

De kwaliteit van grootschalige ‘performance assessments’ gewikt en gewogen

Alexia Deneire, Jetje De Groof, Liesje Coertjens, Vincent Donche, Jan Vanhoof, Peter Van
Petegem en Sven De Maeyer

03/01/2022

Voorwoord

In 2016 leverden we het rapport op voor het onderzoek in opdracht van het departement Onderwijs & Vorming van de Vlaamse overheid (De Maeyer et al. 2016). We rapporteerden enerzijds over een uitgebreide literatuurstudie aangaande cruciale kwaliteitscriteria bij het uitrollen van 'performance assessment' op een grote schaal (bv. met het oog op peilingsonderzoek of centrale toetsing) en anderzijds over een exploratie van buitenlandse voorbeelden (inclusief interviews) waarin dit type van assessment wordt ingezet op grote schaal en de lessen die daaruit af te leiden zijn.

Met dit boek willen we de inzichten verworven doorheen het project en vertaald in het rapport beschikbaar maken voor een breder publiek. Het meten van competenties is geen eenvoudige taak en dit grootschaliger uitrollen brengt een hele reeks van keuzes en uitdagingen met zich mee. Door dit boek open en vrij aan te bieden is het onze hoop dat we verschillende belanghebbenden kunnen bereiken en inspireren.

Inhoud

Voorwoord.....	1
1 Ten Geleide	5
1.1 Context	5
1.2 Leeswijzer	6
2 Probleemstelling en Begrippenkader	9
2.1 Competentieverricht onderwijs, leren en beoordelen	9
2.2 Kwaliteitsmonitoring op een competentieverrichte leerst schoeien	9
2.3 Competenties beoordelen via 'performance assessment'	10
2.4 Doelstellingen van het onderzoek.....	11
2.5 Begrippenkader	11
2.5.1 Competentie	11
2.5.2 'Performance assessment'	12
2.5.3 Kwaliteit bij meten en beoordelen	13
2.5.4 Monitoring op systeemniveau	14
3 Ontwikkeling van de evaluatiematrix	16
3.1 Theoretisch kader	16
3.1.1 Argumentatieve benadering van validiteit	16
3.1.2 Variaties en/of aanvullingen op de argumentatieve benadering van validiteit	19
3.2 Literatuurstudie	24
4 Buitenlandse voorbeelden van grootschalige 'performance assessments'	29
4.1 Australië: National Assessment Program – Literacy and Numeracy (NAPLAN).....	29
4.2 Australië: National Assessment Program (NAP)	31
4.3 Nederland: Periodieke Peiling van het Onderwijsniveau (PPON)	32
4.4 Nieuw Zeeland: National Monitoring Study of Student Achievement (NMSSA)	34
4.5 Schotland: Scottish Survey of Literacy and Numeracy (SSLN)	37
4.6 Verenigde Staten: National Assessment of Educational Progress (NAEP)	38
5 Evaluatiematrix.....	42
5.1 Structuur van de matrix	42
5.2 Matrix: bouwstenen en voorwaarden	46
5.2.1 Bouwsteen 'Doelbepaling'	46
5.2.2 Domeinbeschrijving	48
5.2.3 Opzet en ontwikkeling.....	50
5.2.4 Toetsafname	58

5.2.5 Scoren.....	60
5.2.6 Validiteit.....	66
5.2.7 Niveaubepaling en rapportering.....	69
5.2.8 Haalbaarheid van de bouwstenen	72
5.2.9 Tot slot.....	74
6 Uitdagingen voor grootschalige toetsen die ‘performance assessment’ inschakelen	75
6.1 Uitdaging 1: Voldoende taken voorzien.....	75
6.2 Uitdaging 2: Standaardisering van toetsafname en scores	76
6.3 Uitdaging 3: Vermijden van construct-irrelevante variantie	78
6.4 Uitdaging 4: Het opzetten van taken die recht doen aan de criteriumsituatie	79
6.5 Uitdaging 5: Conform de doelstellingen rapporteren	80
7 Implicaties	83
Woordenlijst	89
Referenties.....	92

1 Ten Geleide

1.1 Context

De tijd dat onderwijs zich louter richtte op (het ontwikkelen van) kennis ligt achter ons. Meer en meer staan ‘competenties’ centraal. Concreet betekent dit dat onderwijsleerprocessen bredere gehelen van kennis, vaardigheden en attitudes bij lerenden erkennen, daarop inspelen en ze verder ontwikkelen. Als de focus van het onderwijs zich verlegt, dan heeft dit ook gevolgen voor de praktijk van het evalueren en beoordelen.

Ook beleidsmakers in Vlaanderen voelden in recente jaren de noodzaak aan om hun toetsprogramma's die zich richten op controle van de kwaliteit van het onderwijs (de zogenaamde peilingsproeven), beter af te stemmen op de evolutie richting competentiegericht onderwijs en competentietoetsing. Dit vormde het vertrekpunt voor het wetenschappelijk onderzoek waaruit deze publicatie voortkomt. In de studie die we in 2016 uitvoerden in opdracht van het departement Onderwijs & Vorming (De Maeyer et al. 2016) gingen we op zoek naar de kwaliteitscriteria van (grootschalige) toetsprogramma's die competenties in kaart willen brengen en daarbij gebruik maken van ‘performance assessment’.

De focus lag ten eerste op ‘performance assessment’, dat in deze publicatie gedefinieerd wordt als de *beoordeling (van competenties) op basis van leerlingprestaties in (levensechte) taken die relevant zijn voor de beoogde competenties*. ‘Performance assessment’ wordt erkend als een krachtige manier om competenties te toetsen, onder andere omwille van het potentieel om leerlingen complexe taken te laten uitvoeren in zo levensecht mogelijke contexten. Het beoordelen van competenties via ‘performance assessment’ is echter niet zonder problemen; zeker niet wanneer dit op grote schaal wordt georganiseerd. Tegen die achtergrond is er nood aan een kader dat houvast biedt aan wie dit soort competentietoetsen wil inzetten.

Ten tweede, was deze studie gericht op grootschalige toetsprogramma's, die kwaliteitszorg op systeemniveau voor ogen hebben. Het zijn met andere woorden toetsprogramma's die eenzelfde doel hebben als het peilingsonderzoek zoals we dat in Vlaanderen kennen: het rapporteren over de prestaties van groepen van leerlingen, met als doel om op systeemniveau een uitspraak te doen over de mate waarin de doelstellingen bereikt worden. Omdat er voor de individuele leerlingen niets van de resultaten van de toetsen afhangt, worden er ook wel eens naar verwezen als ‘low-stakes’ toetsen. Ze verschillen hiermee van grootschalige toetsprogramma's die een uitspraak doen over het competentieniveau van individuele leerlingen, zoals bijvoorbeeld de ‘A-levels’ in het Verenigd Koninkrijk of de Cito-toetsen in Nederland. Deze worden ook wel ‘high-stakes toetsen’ genoemd, omdat de resultaten belangrijk zijn voor (de toekomst van) individuele leerlingen. Daarnaast onderscheiden grootschalige toetsprogramma's zich van toetsen die kleinschaliger zijn en bijvoorbeeld enkel op het niveau van een enkele klas of school worden uitgerold.

In deze publicatie stellen we het pakket onderzoeksresultaten voor dat voortkwam uit boven vernoemd onderzoek (De Maeyer et al. 2016). We maken een stand van zaken op van de kwaliteitseisen die gesteld worden aan 'performance assessments'. Deze kwaliteitseisen worden voorgesteld aan de hand van een evaluatiematrix, die empirisch gefundeerde ondersteuning biedt aan betrokkenen bij (grootschalige) competentiebeoordelingen. We brengen bovendien de meest essentiële uitdagingen in kaart die komen kijken bij competentiebeoordelingen via 'performance assessment' en stellen, waar mogelijk, oplossingen voor. Daarnaast willen we de lezer ook kort meenemen in de ontwikkeling van de matrix door inkijk te geven in de theoretische basis en de onderzoekslijnen die werden uitgezet met het oog op het identificeren van de bouwblokken van de evaluatiematrix. We geven de lezer bovendien informatie over de manier waarop de buitenlandse praktijkvoorbeelden die we analyseerden, te werk gingen bij het opzetten, afnemen en scoren van de toetsen, en hoe ze hierover rapporteerden.

De ontwikkelde evaluatiematrix, stelt stapsgewijs de voorwaarden voor waar grootschalige competentietoetsen die gebruik maken van 'performance assessment' aan moeten voldoen om kwaliteitsvol te zijn. Er wordt stilgestaan bij de afwegingen die daarbij gemaakt moeten worden. We richten ons gezien de oorspronkelijke opdracht in de eerste plaats op grootschalige competentietoetsen, die rapporteren over het prestatieniveau van groepen van studenten ('rapportering op systeemniveau'). Dit neemt niet weg dat de voorgestelde evaluatiematrix en de opgesomde uitdagingen ook relevante inzichten bieden voor wie in bredere zin geïnteresseerd is in het evalueren van competenties. De inzichten kunnen in sommige gevallen bijvoorbeeld ook toegepast worden op 'performance assessments' die in een klas worden ingezet; of op grootschalige competentietoetsen via 'performance assessment' die wel uitspraken doen over de prestaties van individuele leerlingen ('high-stakes' toetsen). Dit maakt dat deze publicatie potentieel interessant lektuur is voor toetsontwikkelaars; beleidsmedewerkers; onderwijsondersteuners; onderzoekers; lerarenopleiders; medewerkers van overheden, inspectie, pedagogische studiediensten, koepels; maar ook leerkrachten. Bij de ontwikkeling van de evaluatiematrix hebben we bovendien veel belang gehecht aan praktische inzetbaarheid. De bouwstenen van de matrix volgen een toetsdesign-insteek, wat betekent dat de logische stappen van het op- en uitzetten van toetsen gevolgd wordt. Dit maakt deze publicatie ook toegankelijk voor lezers die niet psychometrisch of methodologisch onderlegd zijn.

1.2 Leeswijzer

Hieronder geven we een beknopte beschrijving van de verschillende hoofdstukken van deze publicatie.

Hoofdstuk 2 – probleemstelling en begrippenkader schetst de problemen waar we op stoten wanneer het gaat over competentiegericht onderwijs en het beoordelen van 'competenties'. De insteek van kwaliteitsmonitoring op systeemniveau zorgt voor verdere afbakening van de probleemstelling. Met het oog op het beantwoord krijgen van de centrale onderzoeksvraag ('Welke kwaliteitseisen moeten er gesteld worden aan competentiebeoordelingen -- opgezet vanuit het perspectief van kwaliteitsmonitoring -- waarbij men gebruik maakt van performance assessment?'), laten we ook licht schijnen op

de centrale begrippen: 'competentie', 'performance(-based) assessment', 'kwaliteit' en 'monitoring op systeemniveau'.

Hoofdstuk 3 – Ontwikkeling van de evaluatiematrix geeft inzicht in het theoretische kader dat aan de basis ligt van de evaluatiematrix. De keuze voor de 'argumentatieve benadering van validiteit' (Kane 2006) wordt verduidelijkt. Naast Kane wordt ook aansluiting gezocht bij andere auteurs met het oog op de verbetering van de praktische bruikbaarheid van de evaluatiematrix en het hanteren van een brede visie op 'kwaliteit'. Het hoofdstuk geeft ten tweede een beknopt overzicht van de onderzoeksacties die werden uitgezet om verdere invulling aan het kader te geven: de literatuurstudie en de selectie en analyse van buitenlandse praktijkvoorbeelden.

Hoofdstuk 4 – Buitenlandse voorbeelden van grootschalige 'performance assessments' biedt een beknopte beschrijving van elk van de buitenlandse toetssystemen die we onder de loep namen. We geven de lezer inzicht in de belangrijkste elementen van de manier waarop de toets is opgezet en volgen hierbij reeds de bouwstenen zoals die in de evaluatiematrix worden voorgesteld. Dit hoofdstuk geeft inzicht in de verschillende manieren waarop grootschalige toetsen worden opgevat in het buitenland en geeft de nodige basisinformatie voor hoofdstukken 5 en 6.

Hoofdstuk 5 – Evaluatiematrix stelt een raamwerk voor dat gebruikt kan worden voor het opzetten en/of evalueren van grootschalige competentietoetsen op basis van 'performance assessment', die zich richten op kwaliteitsmonitoring op systeemniveau. De focus ligt op de verschillende bouwstenen van een kwaliteitsvolle toets, waarbij grotendeels een toetsdesigninstek wordt gevolgd. Aan elke bouwsteen worden een aantal voorwaarden gekoppeld, waaraan voldaan moet worden om tot een valide uitspraak over het prestatieniveau van een (groep van) leerling(en) te komen. De argumentatieve benadering van validiteit leert ons daarbij dat de niet-kwaliteitsvolle invulling van de ene bouwsteen, gevolgen heeft voor de kwaliteit van de volgende bouwsteen. Waar mogelijk illustreren we de voorwaarden aan de hand van internationale praktijkvoorbeelden. De matrix is enerzijds bedoeld als hulpmiddel voor toetsontwikkelaars. Anderzijds ondersteunt hij beleidsmedewerkers die een uitspraak moeten doen of een beslissing moeten nemen over de kwaliteit van bestaande toetsen en toetsconcepten. Door de stapsgewijze bespreking van de kwaliteitsvoorwaarden is het instrument echter interessant voor alle lezers die geïnteresseerd zijn in praktische handvaten voor het opzetten van competentiebeoordelingen via 'performance assessment'. Hoewel de focus ligt op grootschalige toetsen die een uitspraak doen op systeemniveau, kunnen er ook lessen uit getrokken worden voor kleinschaligere toetsen, of grootschalige toetsen die wel de ambitie hebben op individueel leerlingenniveau een uitspraak te doen.

Hoofdstuk 6 - Essentiële uitdagingen voor grootschalige toetsen die 'performance assessment' inschakelen gaat in op een aantal essentiële uitdagingen met betrekking tot grootschalige toetsen met een 'performance assessment'-component. We identificeerden deze uitdagingen op basis van inzichten uit de literatuur, met name de systematische literatuurstudie en basiswerken over 'performance assessment' enerzijds, en inzichten uit de geanalyseerde praktijkvoorbeelden anderzijds. Het hoofdstuk stipt waar mogelijk ook manieren aan om deze uitdagingen om te buigen en tot alternatieve oplossingen te komen.

Hoofdstuk 7 – Implicaties en uitdagingen voor de praktijk gaat in op de lessen die uit het onderzoek getrokken kunnen worden voor het actuele en toekomstige beleid rond en de implementatie van grootschalige ‘performance assessments’ met het oog op kwaliteitsmonitoring op systeemniveau.

Achteraan de publicatie wordt een ‘Woordenlijst’ voorzien, waar de termen die nadere toelichting behoeven, op een rijtje worden gezet.

2 Probleemstelling en Begrippenkader

De maatschappij verandert aan een hoog tempo en hetzelfde geldt voor het onderwijs. In het voorbije decennium heeft competentiegericht onderwijs meer en meer ingang gevonden in de Verenigde Staten, in Europa en ook in Vlaanderen. Aangezien de manier waarop men beoordeelt een sterke invloed heeft op de wijze waarop men leert, dient ook de omslag gemaakt te worden naar competentiegericht beoordelen en evalueren. Op die manier doet 'performance(-based) assessment' zijn intrede. Brede, complexe constructen zoals competenties, zijn doorgaans immers veel moeilijker te meten op grond van zogenaamde 'klassieke' toetsen (bv. een schriftelijke toets opgesteld uit meerkeuzevragen). Het kwaliteitsvol opzetten van 'performance assessment' is echter geen eenvoudige taak, zeker niet wanneer de toetsing grootschalig is. Dit is zeker zo als het toetsen gericht is op het bewaken van de kwaliteit op het niveau van een onderwijssysteem, wat de aanleiding vormde van de studie die aan de grondslag ligt van deze publicatie.

2.1 Competentiegericht onderwijzen, leren en beoordelen

In reactie op de geschetste evolutie is niet alleen in de Verenigde Staten een beweging vast te stellen richting 'performance-based education' (Darling-Hammond and Adamson 2014); in veel Europese landen is de idee van competentiegericht onderwijs intussen stevig verankerd (Weigel, Mulder, and Collins 2007). Wijzigingen in het onderwijs vereisen ook beoordelingsmethoden die daaraan aangepast zijn, teneinde op een adequate wijze vast te stellen of en in welke mate de vooropgestelde competenties verworven zijn J. Biggs (1996) en J. B. Biggs and Tang (2011) hebben het over 'constructive alignment': de noodzaak om instructie, leren en beoordelen mooi op elkaar af te stemmen. Prodrômou (1995) spreekt in dit verband van het 'backwash effect': wat beoordeeld wordt, bepaalt in sterke mate wat wordt geleerd. Aangezien het onderwijsleerproces in toenemende mate gestuurd wordt vanuit het raamwerk van competenties, duikt het risico op dat de noodzakelijke afstemming tussen de elementen instructie, leren en beoordelen op de helling komt te staan. Indien het onderwijsbeleid en de onderwijspraktijk in Vlaanderen wenst te evolueren naar en/of verder wil inzetten op meer competentiegericht onderwijs, dan mag het beoordelingsproces niet achterblijven en moeten beoordelingen niet enkel gericht zijn op het meten van louter kennis, maar ook van competenties. Zo komen we terecht bij competentietoetsing, die verschillende vormen kan aannemen en uiteenlopende functies kan dienen, bv. toetsen op klasniveau of op systeemniveau, vanuit formatieve en/of summatieve insteek, kleinschalig of grootschalig, met het oog op ontwikkeling of vanuit een verantwoordingsperspectief.

2.2 Kwaliteitsmonitoring op een competentiegericht leerstelsel

Om het peil van het onderwijs te bewaken en te verbeteren organiseert het Vlaams onderwijsbeleid jaarlijks peilingsonderzoek. Peilingsonderzoek vormt dus een van de hoekstenen van kwaliteitszorg op systeemniveau. Dergelijk onderzoek 'peilt' in welke mate bepaalde eindtermen behaald zijn (bv. Hoe is het gesteld met de schrijfvaardigheid in Vlaanderen in een bepaald jaar?). Daarbij gaat men ook na of de prestaties te vergelijken en te verklaren zijn aan de hand van leerling-, klas- en schoolkenmerken. Door systematisch te peilen naar de mate waarin eindtermen behaald worden, zet dit soort onderzoek

desgevallend aan tot bepaalde beleidsinitiatieven. Kenmerkend voor peilingsonderzoek is dat het zich beroept op grote steekproeven van leerlingen om tot statistisch relevante conclusies te komen. Bijgevolg worden hoofdzakelijk deelcomponenten van competenties bevraagd (o.a. specifieke kennis en/of vaardigheden) die met meerkeuzevragen of gesloten vragen in kaart te brengen zijn. Dit resulteert echter in een partieel beeld van de bekwaamheid van leerlingen: de competenties worden immers niet geïntegreerd in beeld gebracht. Waar vroeger het accent lag op het meten van deelaspecten van kennis of specifieke vaardigheden, onderstreept het beleid meer en meer de noodzaak om competentiegericht te evalueren. Dit laatste gebeurt in Vlaanderen ook reeds ten dele bij de praktische proeven, die deel uitmaken van sommige peilingsproeven. Leerlingen moeten bijvoorbeeld een sollicitatiegesprek doen of voeren natuurexperimenten uit, wat hen in staat stelt competenties op integratieniveau te tonen. Omdat de praktische proeven echter bij een beperkte steekproef leerlingen worden afgenomen, kan op basis van deze proeven geen uitspraak gedaan worden over het behaalde niveau van de leerlingen in Vlaanderen, wat het uiteindelijke doel van de peilingstoetsen is. Dit illustreert dat hoewel men meer en meer doordrongen is van de noodzaak om ook toetsprogramma's in het kader van kwaliteitsmonitoring in lijn te brengen met de evolutie richting competentiegericht onderwijs, het geen evidentie is om dit in de praktijk te brengen.

2.3 Competenties beoordelen via 'performance assessment'

Omwille van het complexe samenspel tussen kennis, vaardigheden en attitudes die gepast zijn voor een bepaalde context (Figel 2007), is de beoordeling van competenties niet eenvoudig. Om bijvoorbeeld na te gaan of leerlingen eigen ideeën creatief kunnen vormgeven door gebruik te maken van ICT, is een gesloten kennistoets onvoldoende om tot valide conclusies te komen (Rubin 1996). Leerlingen een schoolaffiche laten ontwerpen met behulp van ICT, heeft daarentegen meer potentieel om een valide beeld te krijgen van de betreffende competentie. In de onderzoeksliteratuur plaatst men dit soort simulatie van realistische taken in realistische situatie(s), waarin de te beoordelen competentie moet worden gebruikt, onder de koepel 'performance assessment' (o.a. Kane, Crooks, and Cohen 1999; Suzanne Lane and Stone 2006). Voor de keuze om competenties bij leerlingen te toetsen via (vormen van) 'performance(-based) assessment' vinden we onder andere steun bij Suzanne Lane (2010). Zij stelt dat:

(w)hen students are given the opportunity to work on meaningful, real world tasks in instruction, students have demonstrated improved performance on performance assessments. Sound educational practice calls for the alignment among curriculum, instruction and assessment, and there is ample evidence to support the use of performance assessments in both instruction and assessment to improve student learning for all students.

Hoewel Suzanne Lane (2010) opmerkt dat 'performance assessment' ook inzetbaar is voor grootschalige beoordelingen (zoals bv. peilingsonderzoek), is het toetsen van competenties via grootschalig 'performance assessment' niet zonder problemen. Bij het opzetten van deze toetsen dienen een aantal keuzes gemaakt te worden, die een effect hebben op de kwaliteit van de toets. Richtinggevende vragen zijn onder meer: Hoeveel taken zijn nodig om de beoogde competentie goed in kaart te brengen? Aan welke elementen dient men aandacht

te besteden bij het uitwerken van de domeinbeschrijving en het daaruit resulterende toetsraamwerk? Welke mate van standaardisering is er vereist op vlak van toetsafname? Hoe ervoor zorgen dat beoordelaars een vergelijkbaar oordeel vellen? Welk cijfer is voldoende om de standaard te behalen?

Om gericht stappen te zetten in de richting van meer competentiegericht peilingsonderzoek, is het nodig om te verhelderen hoe deze keuzes een impact hebben op een kwaliteitsvolle implementatie van 'performance assessment'. Wetenschappelijk onderzoek binnen het domein Onderwijs & Meten ('Educational Measurement') biedt daartoe reeds een aantal inzichten. Een inventarisatie van de meest recente inzichten uit dit domein kan ondersteuning bieden in het uitwerken van een grootschalig, meer competentiegericht toetsprogramma en kan de evaluatie van alternatieve manieren van toetsing van competenties ondersteunen.

2.4 Doelstellingen van het onderzoek

Tegen de geschetste achtergrond beoogde het onderzoeksproject dat aan de basis van deze publicatie ligt (1) een stand van zaken te geven van de inzichten m.b.t. de kwaliteitseisen van 'performance assessment'; en (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. Verder was het doel om (3) op basis van de evaluatiematrix, buitenlandse voorbeelden van grootschalige competentiebeoordelingen te inventariseren en te duiden. Het (grootschalig) meten van competenties is immers ook in andere onderwijssystemen een uitdaging. Inzicht in hoe men hier in realiteit mee omgaat en in welke overwegingen gemaakt kunnen worden, kunnen helpen om de kwaliteitseisen in een realistisch perspectief te zien.

De centrale onderzoeksvraag, 'Welke kwaliteitseisen moeten er gesteld worden aan competentiebeoordelingen — opgezet vanuit het perspectief van kwaliteitsmonitoring – waarbij men gebruik maakt van performance assessment?', werd enerzijds ingegeven door de noodzaak om toetsprogramma's die kwaliteitsmonitoring beogen (meer) toe te spitsen op competenties; en anderzijds door de keuze voor 'performance assessment' om competenties te beoordelen. In de volgende sectie gaan we dieper in op vier begrippen die in de onderzoeksvraag centraal staan.

2.5 Begrippenkader

2.5.1 Competentie

Naast de gangbare omschrijvingen van het begrip 'competentie' die opduiken in Europese en Vlaamse beleidsdocumenten, vinden we ook tal van definities terug in de academische literatuur. Baartman et al. (2007) bijvoorbeeld, stellen in hun analyse van gangbare definities vast dat het begrip 'competentie' op veel verschillende manieren wordt gedefinieerd. De auteurs besluiten dat er, algemeen genomen, twee belangrijke aspecten terug te vinden zijn: (1) de integratie van vaardigheden, kennis en attitudes en (2) een link naar een bepaalde jobcontext. Ook in de definitie die L. Baartman (2008) hanteert, en die teruggrijpt op de omschrijving die ook Lizzio and Wilson (2004) voorstellen, zijn beide

componenten aanwezig. Competentie is voor L. Baartman (2008, 11) : *“(...) the capacity to enact specific combinations of knowledge, skills, and attitudes in appropriate job contexts”*.

Omdat wij ook focussen op leerlingen uit het lager en het secundair onderwijs, verruimen we de context waarbinnen het begrip competentie in Baartmans definitie vorm krijgt. Hiervoor doen we een beroep op de omschrijving die de werkgroep ‘Erkennen van Verworven Competenties’ (EVC) hanteert: *“de reële en individuele capaciteit om kennis (theoretische en praktische kennis), vaardigheden en attitudes in het handelen aan te wenden, en dit in functie van de concrete, dagelijkse en veranderende werksituatie en van persoonlijke en maatschappelijke activiteiten”* Dienst Beroepsopleiding (2008, 6–7) . Met andere woorden, niet enkel de professionele context speelt een rol, maar ook de persoonlijke en maatschappelijke omgeving waarin kinderen, jongeren en adolescenten zich bewegen.

Op grond van de verschillende invullingen van het begrip en gegeven focus van deze publicatie, schuiven we de volgende werkdefinitie voor het begrip ‘competentie’ naar voor.

Competentie:

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.

2.5.2 ‘Performance assessment’

Vanuit het perspectief een breed kwaliteitskader met betrekking tot ‘performance assessment’ van competenties aan te reiken, is het cruciaal dat we ook kiezen voor een brede, open definitie van het begrip ‘performance assessment’ zelf. In de Angelsaksische wereld wordt de term ‘performance assessment’ veelal breed gedefinieerd. Daar omvat ‘performance assessment’ alles wat buiten de categorie meerkeuzevragen valt, wat te verklaren is door de traditie daar om voor bijna alle ‘high stakes’-toetsing, dit is toetsing waarbij er voor de leerling veel op het spel staat, voor meerkeuzevragen te kiezen. De definitie van Basturk (2008, 431–32) illustreert deze invalshoek: *“Performance Assessment refers to a form of evaluation that requires students to perform a task rather than select an answer from a ready-made list.”* Vanuit dit perspectief omvat ‘performance assessment’ een zeer breed gamma aan activiteiten: van het aanvullen van zinnen via enkele woorden, over het schrijven van een grondige analyse, naar het uitvoeren van een onderzoek in een labo en het schrijven van een verslag hierover Stecher (2015). Hoewel deze definitie het beslist mogelijk maakt een breed kwaliteitskader van ‘performance assessment’ van competenties uit te werken, vertrekt ze te veel van wat ‘performance assessment’ niet is (‘het is alles wat niet te bestempelen valt als een toets op basis van meerkeuzevragen’) en gaat ze te weinig in op wat er uniek aan is. De definitie van Fitzpatrick and Morrison (1971, 268) komt hier wel aan tegemoet:

A performance test (performance or product evaluation) has been defined here as a test in which a criterion situation, such as a job, is simulated to a relatively high degree (...) the potential value of the performance test lies in its closer approach to reality – its greater relevance in determining the degree to which the examinee can actually perform the tasks of the criterion job or other situation.

In het verlengde hiervan verwijst Suzanne Lane (2010) en S. Lane (2015) expliciet naar Kane, Crooks, and Cohen (1999) en diens opvatting over de nauwe gelijkheid tussen de 'performance' of prestatie die wordt beoordeeld en de 'performance' of prestatie waarin men is geïnteresseerd, als definiërende eigenschap van 'performance assessment'. 'Performance assessment' verwijst met andere woorden naar simulaties van realistische taken in realistische situatie(s) waarin de te beoordelen competentie moet worden gebruikt.

Cognitieve complexiteit is een ander element dat in sommige definities van 'performance assessment' aanwezig is (o.a. Eisner 1999; Messick 1996). Cognitieve complexiteit verwijst naar de noodzaak om cognitieve strategieën van hogere orde in te schakelen om de taak tot een goed einde te brengen. We kiezen er bewust voor om dit element niet op te nemen in onze definitie. In de begripsomschrijving van 'performance assessment' maken Suzanne Lane and Stone (2006, 388) duidelijk waarom volgens hen niet alle 'performance assessments' 'complexe denkvaardigheden' vereisen: "(...) *the extent to which a performance assessment should require high-level reasoning and problem solving skills is dependent on the performance of interest.*"

Deze elementen samen genomen leidt tot de volgende werkdefinitie voor 'performance assessment'.

Performance assessment:

Een assessment of beoordeling (van competenties) op basis van leerlingprestaties in (levensechte) taken die relevant zijn voor de beoogde competenties.

2.5.3 Kwaliteit bij meten en beoordelen

Hoofddoel is de kwaliteitsvereisten in kaart te brengen die gesteld worden aan grootschalige toetsprogramma's die competenties meten via 'performance assessment'. Hierboven bakenden we reeds de begrippen 'competentie' en 'performance assessment' af; nu stellen we het begrip 'kwaliteit' uit de term 'kwaliteitsvereisten' aan de orde.

Traditioneel komen we bij de conceptualisering van 'kwaliteit' in het kader van meten en beoordelen uit bij begrippen als 'validiteit' en 'betrouwbaarheid'. Over validiteit wordt reeds lang gedebatteerd (Lissitz and Li 2011); talrijk zijn de werken die de betekenis van het begrip, inclusief het meten van validiteit onder de loep namen (cf. AERA APA & NCME 2014; Brennan 2006; L. Cronbach 1971; M. T. Kane 2013; S. Messick 1989). Wij nemen op pragmatische wijze akte van de veelheid aan definities en discussies inzake terreinafbakening en stellen in navolging van Sireci (2009) dat validiteit te maken heeft met de geschiktheid van de interpretatie en het gebruik van toetsscores. Valideringsonderzoek speurt dus naar bewijzen voor een welbepaalde interpretatie en gebruik van scores op beoordelingen of toetsen. Daarbij is het belangrijk te benadrukken dat niet het instrument, een toets, een taak of een toetsscore op zich al of niet valide is, maar wel de interpretatie die men aan de daaruit afgeleide score hecht, alsook de wijze waarop scores gebruikt worden (Lee J. Cronbach and Gleser 1965; Lee J. Cronbach and Meehl 1955; Kane 2006; M. T. Kane 2013; S. Messick 1989). Betrouwbaarheid verwijst onder andere naar de consistentie van scores over replicaties van een toets of beoordelingen heen. De aard en kwaliteit van de respons van een leerling op een toets kunnen variëren van de ene set taken naar de andere,

of van het ene moment van toetsafname naar het andere, zelfs onder gecontroleerde omstandigheden. Verschillende beoordelaars kunnen bovendien andere scores toekennen aan dezelfde prestatie (AERA APA & NCME 2014). Betrouwbaarheidsonderzoek heeft het kwantificeren van de precisie van testcores en het in kaart brengen van de foutenbronnen tot doel (Haertel 2006).

Validiteit en betrouwbaarheid vormen centrale begrippen wanneer het gaat over kwaliteitscriteria voor het opzetten van grootschalige competentiebeoordelingen. Daarnaast zijn er ook zogenaamde ‘alternatieve’ kwaliteitscriteria, zoals authenticiteit, transparantie en eerlijkheid niet uit het oog te verliezen (L. Baartman et al. 2006; P. Newhouse 2013). Authenticiteit verwijst naar de graad van gelijkenis tussen de toetstaken en taken die in het ‘echte leven’ moeten worden uitgevoerd. Eerlijkheid heeft betrekking op het gegeven dat een toets bepaalde groepen niet mag bevoordelen/benadelen en de beoogde kennis, vaardigheden en attitudes (KVA’s) moet weerspiegelen, zonder irrelevante variantie toe te staan. Transparantie impliceert dat een toets bevattelijk is voor alle deelnemers, dat leerlingen de beoordelingscriteria kennen, weten wie de beoordelaars zijn, en wat het doel van de toets is (L. Baartman et al. 2006).

Hoewel we de elementen ‘betrouwbaarheid’, ‘authenticiteit’, ‘transparantie’ en ‘eerlijkheid’ in bovenstaande paragrafen los van de kwestie validiteit aan bod lieten komen, volgen we o.a. AERA APA & NCME (2014) in de erkenning dat dit ten gronde allemaal validiteitskwesties zijn. Als men niet tegemoet komt aan deze kwaliteitscriteria, verkleint de voorspellende waarde van scores ten aanzien van bepaalde criteria, vormen de scores een minder solide vertrekpunt om uitspraken te doen over de leerlingen, en zijn de mogelijkheden voor een degelijke beslissing op basis van de toetsscores beperkt.

We volgen C. P. Newhouse (2011) en Der Vleuten and Schuwirth (2005) bovendien in de vaststelling dat het bepalen van de kwaliteit van een toets steeds een afweging impliceert tussen de onderscheiden kwaliteitscriteria enerzijds en haalbaarheid in termen van tijd en middelen die nodig zijn om deze kwaliteitscriteria te garanderen anderzijds. Daarom moeten tijd en middelen verbonden aan het opzetten en implementeren van een toets ook steeds mee in beschouwing worden genomen.

Samengevat, stellen we deze werkdefinitie voor het begrip ‘kwaliteit’ voor.

Kwaliteit:

...is 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.

2.5.4 Monitoring op systeemniveau

Deze publicatie richt zich op kwaliteitscriteria voor grootschalige competentietoetsen op basis van ‘performance assessment’, vanuit het perspectief van monitoring of kwaliteitsbewaking op systeemniveau. Onze focus is met andere woorden gericht op toetsen op het macroniveau, die iets zeggen over het onderwijssysteem als geheel. Ze

onderscheiden zich van toetsen op meso- en microniveau, die uitspraken doen over respectievelijk scholen en individuele leerlingen. Anderzijds verschillen ze van andere toetsen op macroniveau, die uitspraken doen op individueel leerlingniveau. Grootschalige toetsen vanuit het perspectief van monitoring of kwaliteitsbewaking op systeemniveau worden ontworpen en afgenomen met het oog op het beantwoorden van de vraag wat groepen van leerlingen kunnen en kennen. Mogelijk worden hierbij vergelijkingen gemaakt tussen groepen van leerlingen (bijvoorbeeld naar regio of geslacht) of wordt aangeduid in hoeverre bepaalde groepen de beoogde onderwijsdoelstellingen bereiken (zoals bv. de eindtermen in Vlaanderen) (Mazzeo and Zieky 2006).

Omdat bij toetsen die monitoring op systeemniveau beogen meestal op groepsniveau wordt gerapporteerd, hangt er in principe voor individuele leerlingen en scholen niets van af: er wordt op basis van de resultaten bijvoorbeeld geen beslissing genomen over het al dan niet slagen van individuele leerlingen of over de financiering van scholen. Tegen die achtergrond zijn het 'low-stakes'-toetsen. Dit staat in contrast met 'high-stakes'-toetsen, waar voor de individuele leerling of school wel gevolgen gekoppeld zijn aan de geleverde prestaties

Samengevat stellen we met betrekking tot monitoring op systeemniveau volgende werkdefinitie voor.

Monitoring op systeemniveau:

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.

3 Ontwikkeling van de evaluatiematrix

De ontwikkeling van de evaluatiematrix gebeurde in verschillende stappen en iteraties. De argumentatieve benadering van validiteit (Kane, Crooks, and Cohen 1999; M. T. Kane 2013; Kane 2006) vormde de theoretische insteek. Daarnaast werd een literatuurstudie uitgevoerd. De inzichten verkregen uit de argumentatieve benadering van validiteit, door de literatuurstudie en door lectuur van een aantal basiswerken rond 'performance assessment', stelden ons in staat steeds preciezer de vinger te leggen op de essentiële onderdelen van een kwaliteitsraamwerk voor (grootschalige) competentietoetsen die gebruik maken van 'performance assessment'. Ze werden verwerkt in opeenvolgende versies van de evaluatiematrix. Een stuurgroep en expertengroep gaven feedback op de kwaliteit, volledigheid, inzichtelijkheid en bruikbaarheid van de verschillende concepten van de matrix. Vervolgens werd de evaluatiematrix via de analyse van praktijkvoorbeelden en via interviews grondig aan de praktijk getoetst.

3.1 Theoretisch kader

Het theoretisch kader is geïnspireerd op de argumentatieve benadering van validiteit (Kane, Crooks, and Cohen 1999; M. T. Kane 2013; Kane 2006). In wat volgt verduidelijken we eerst waarom we de argumentatieve benadering van validiteit als uitgangspunt hebben gekozen, om vervolgens de kernelementen van deze benadering aan te stippen. Daarna belichten we bij andere auteurs aanknopingspunten die ons geholpen hebben om (1) de argumentatieve benadering te gebruiken voor de ontwikkeling van de evaluatiematrix, en (2) extra kwaliteitseisen met betrekking tot (grootschalige) 'performance assessment' te identificeren.

3.1.1 Argumentatieve benadering van validiteit

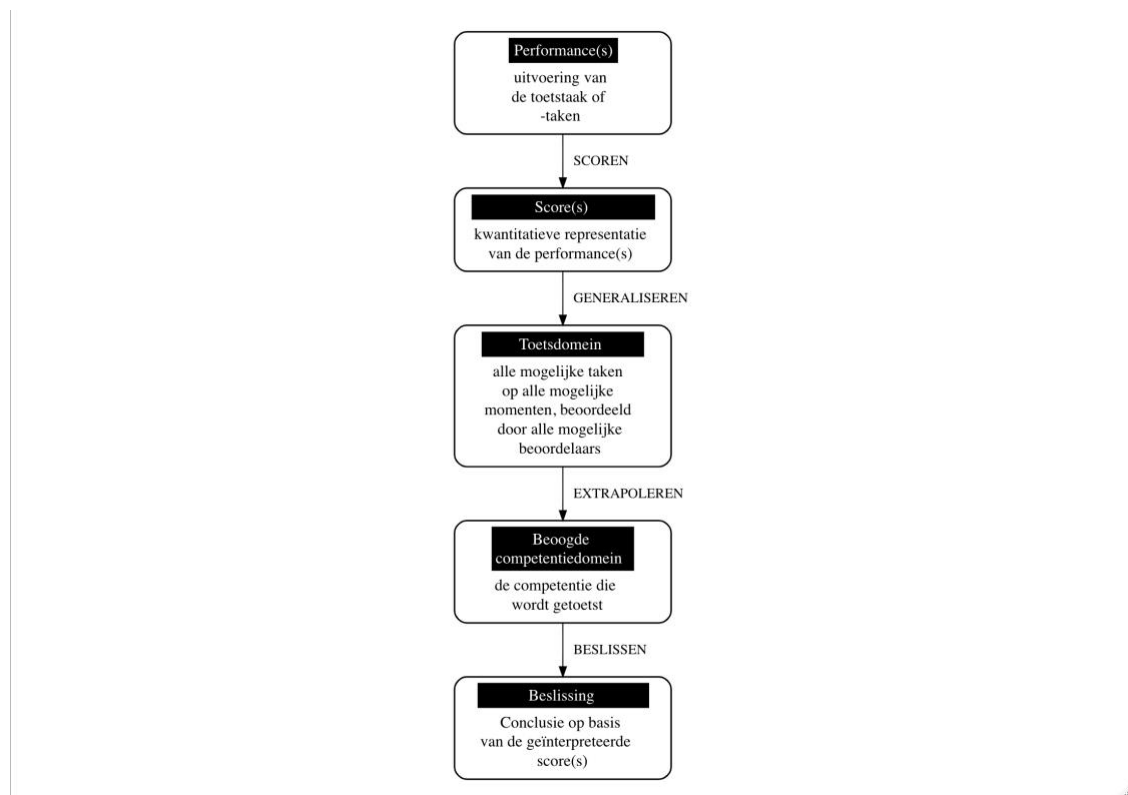
Om de kwaliteit van toetsen en toetsprogramma's te evalueren bestaan er verschillende referentiekaders [L. Baartman (2008); Kane (2006); M. T. Kane (2013); Wools (2015); Wools, Sanders, and Roelofs (2007)]. Vele auteurs (zie o.m. 3.1.2.) schuiven de argumentatieve benadering van validiteit [Kane, Crooks, and Cohen (1999); M. T. Kane (2013); Kane (2006)] 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 van validiteit 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 de aanpak ook handvaten om de argumentatie van validiteit op een methodologisch solide basis te onderbouwen.

In deze benadering staat de volgende vraag centraal: hoe kunnen we op grond van prestaties op één of een beperkte aantal taken uit het toetsdomein (dit zijn alle mogelijke taken, afgenomen op alle mogelijke momenten, beoordeeld door alle mogelijke beoordelaars), tot een valide conclusie komen omtrent verwachte prestaties in het beoogde competentiedomein (dit is de competentie die je wenst te meten)? Kane [-M. T. Kane (2013); -Kane (2006)] stelt voor om dit probleem in kleine deelproblemen op te splitsen en te werken in stappen die in een keten kunnen worden gesitueerd. De vertaalslag tussen

toetsprestaties aan de ene zijde van de keten en een uiteindelijke beslissing op basis van de toetsscore aan de andere zijde van de keten, omvat volgens Kane [-M. T. Kane (2013); -Kane (2006)] 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;
3. **extrapoleren** van deze scores naar scores voor het beoogde competentiedomein;
4. het nemen van **beslissingen**.

Figuur 3.1 visualiseert deze keten, die door Kane zelf het ‘interpretatieve- en gebruiksargument’ (M. T. Kane 2013; Kane 2006) 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 de toets heeft opgezet.



Figuur 3.1: Argumentatief model van validiteit

Het voorbeeld van de leerlingen die een schoolaffiche moeten ontwerpen met behulp van ICT (zie hierboven 1.3.) biedt een goede illustratie. De leerkracht geeft een score voor de uitgevoerde taak (= stap 1: SCOREN). Deze specifieke taak is echter maar een van de vele mogelijke taken die hij had kunnen opstellen. Hetzelfde geldt voor het afnamemoment: de taak had hij evengoed niet op vrijdagmiddag, maar op maandagmorgen of een paar weken later kunnen afnemen. En de taak had net zo goed beoordeeld kunnen worden door een parallelleerkracht of door een extern iemand. Indien aangetoond kan worden dat de

geobserveerde score representatief is voor de (hypothetische) score op alle mogelijke taken, verkregen op alle mogelijke afnamemomenten, vanwege alle mogelijke beoordelaars, kan de score gegeneraliseerd worden naar het toetsdomein of het universum (= stap 2: GENERALISEREN). Om vervolgens de verwachte score voor het toetsdomein te kunnen extrapoleren naar een score voor het ruimere competentiedomein ('eigen ideeën creatief vormgeven door gebruik van ICT'), moet aangetoond worden dat de toetstaken adequate maten zijn 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. Dit betekent dat de leerkracht moet kunnen bewijzen dat het ontwerpen van een schoolaffiche op basis van ICT een goede weerspiegeling is van het creatief vormgeven van ideeën op basis van ICT. De score voor het beoogde competentiedomein reflecteert dus als het ware hoe de leerling daarbuiten, in de echte wereld zou presteren (= stap 3: EXTRAPOLEREN). In de laatste stap ten slotte neemt de leerkracht een beslissing op basis van het behaalde competentieniveau. In functie van het al dan niet behaald zijn van de (prestatie)standaard die hij vooropstelt of had vooropgesteld voor de beoogde competentie, slaagt de leerling of heeft hij/zij een herkansing nodig (= stap 4: BESLISSING NEMEN). Hij verbindt zo verdere consequenties aan de score. Telkens er toetsresultaten gebruikt worden om conclusies te trekken of beslissingen te nemen, wordt deze logische argumentatie van stappen toegepast (Kane 2006).

Met betrekking tot de stappen 'generaliseren' en 'extrapoleren' stoten we, zeker in het geval van 'performance assessments', op een onvermijdelijke paradox (Kane, Crooks, and Cohen 1999; Kane 2006). 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 op een specifieke manier vastlegt ((in zeker mate verengt); tegelijkertijd ondermijnt ze door deze vastlegging, ook de mogelijkheid tot extrapoleren. Standaardisering kan er soms toe leiden dat het toetsdomein ten opzichte van het beoogde competentiedomein te veel beperkt wordt en geen recht doet aan de diversiteit van het geheel. Nemen we opnieuw het voorbeeld van de schoolaffiche erbij, dan kan de leerkracht de taak bijvoorbeeld standaardiseren door de veelheid aan mogelijke software-toepassingen die leerlingen zouden kunnen gebruiken om hun affiche te ontwerpen, te beperken tot één programma. Een gevolg van deze standaardisering is dat gemakkelijker kan worden aangetoond dat de score die aan deze (verengde) taak wordt toegekend, representatief is voor het (ingeperkte) toetsdomein. Een negatief gevolg van deze standaardisering echter, is dat de leerkracht niet of moeilijker kan argumenteren dat het ontwerpen van een schoolaffiche waarbij leerlingen maar één programma mogen gebruiken, een goede weerspiegeling is van de volledige competentie 'creatief vormgeven van ideeën op basis van ICT'.

Eigen aan de opeenvolging binnen de keten is dat we de vier stappen en de onderliggende veronderstellingen in het argument expliciteren en de bewijzen aan een serie kritische tests onderwerpen. Dit kunnen bijvoorbeeld logische analyses en/of empirische studies zijn. De argumentatieve benadering van validiteit onderscheidt daarom twee types argumenten [M. T. Kane (2013); Kane (2006)]. Het interpretatieve- en gebruiksargument specificeert de beoogde interpretatie en het beoogde gebruik van scores, door de volledige keten van gevolgtrekkingen en onderliggende assumpties tussen geobserveerde prestaties en

conclusies en beslissingen op basis van deze interpretaties, uiteen te leggen en te expliciteren. Het validiteitsargument of de beoordelingsfase omvat de evaluatie van het interpretatieve- en gebruiksargument [L. Cronbach (1971); Kane (2006); M. T. Kane (2013); Toulmin (2003)]. Na het verzamelen en structureren van de noodzakelijke evidentie in het interpretatieve- en gebruiksargument, worden de verschillende logische en/of empirische bewijzen voor de gemaakte deducties en onderliggende assumpties naar boven gehaald en aan kritische testen onderworpen [L. Cronbach (1971); Kane (2006); M. T. Kane (2013); Kane, Crooks, and Cohen (1999); S. Messick (1989)]. Analytische bewijzen zijn bijvoorbeeld verslagen over de rationale van de item- en taakconstructie. Empirische bewijzen zijn bewijzen gebaseerd op (in)directe waarnemingen en worden op grond van (statistische) analyses op de verzamelde gegevens verzameld. Op grond van verschillende soorten bewijzen afkomstig uit diverse bronnen kan er worden aangetoond dat er sprake is van een voldoende valide interpretatie en/of voldoende valide gebruik van de toetsscore (of niet). De notie 'voldoende' wijst er overigens op dat het bepalen van de validiteit geen kwestie is van alles of niets. Over validiteit spreken we in termen van minder of meer.

3.1.2 Variaties en/of aanvullingen op de argumentatieve benadering van validiteit

Bij de ontwikkeling van de evaluatiematrix consulteerden we naast Kane ook andere auteurs (Crooks, Kane, and Cohen 1996; Chapelle, Enright, and Jamieson 2010; Chapelle 2012; Shaw, Crisp, and Johnson 2012; Wools 2015). Deze auteurs vertrekken evenzeer vanuit een argumentatieve benadering van validiteit: de keten van scoren, generaliseren, extrapoleren en beslissen vormt ook de kern van hun betoog. De specifieke invulling van het valideringskader van deze auteurs vertoont echter interessante nuanceverschillen. Deze verschillen zijn het gevolg van een andere toetsinstek (bv. taaltoetsen, traditionele toetsen of meer competentiegerichte assessments), of van de nadruk die de auteurs willen leggen op specifieke aspecten van het beoordelingsproces. Deze auteurs hebben ons verder op weg gezet om het gedachtengoed van Kane te vertalen en om de matrix ook vorm te gaan geven vanuit de stappen die toetsontwikkelaars traditioneel volgen. Deze beide elementen hebben een positieve impact gehad op de concrete bruikbaarheid van de matrix.

In de volgende paragrafen schetsen we kort om welke verschilpunten het gaat en welke lessen wij daaruit trokken met het oog op de verdere uitwerking van de evaluatiematrix.

Crooks en collega's (1996)

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 aaneengesloten fasen: toetsafname, scoren, aggregeren, generaliseren, extrapoleren, beslissing en impact. Interessant voor ons is het feit dat toetsafname en scoren uit elkaar gehaald zijn. Dit onderscheid biedt ook voor onze evaluatiematrix een belangrijke toegevoegde waarde. In de fase van de toetsafname schuilen immers belangrijke valkuilen die, indien ze niet gemedend worden, de mogelijkheid tot generaliseren en extrapoleren van toetsscores hypothekeren.

Chapelle (2012) en collega's (2010)

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, Enright, and Jamieson 2010; Chapelle 2012).

Het opnemen van de stap 'domeinbeschrijving' vloeit rechtstreeks voort uit volgende observatie van Kane (2006, 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, Enright, and Jamieson (2010, 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). Het expliciteren van een stap 'domeinbeschrijving' zien wij als een duidelijke toegevoegde waarde bij het uitwerken van de evaluatiematrix.

Shaw en collega's (2011)

Shaw, Crisp, and Johnson (2012) 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 concrete valideringsvraag gelinkt. Dergelijke duidelijke handvaten zijn niet geëxpliciteerd in de argumentatieve benadering van Kane. Net daarom vinden we deze aanpak - met het oog op de uitwerking van onze evaluatiematrix - een verbetering. De gerichtheid op praktische bruikbaarheid maakt 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 overigens 'constructrepresentatie' als extra (eerste) stap naar voor. Hiermee onderschrijven ook zij het belang van een grondige domeinbeschrijving.

Wools (2015)

Het valideringskader van Wools (2015) 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, scores, 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 extrapoleren van het toetsdomein naar het competentiedomein. De volgende stap trekt de mogelijkheid tot extrapoleren door van dat competentiedomein naar het praktijkdomein. Het praktijkdomein omvat dan situaties uit het dagdagelijkse leven die mensen kunnen tegenkomen in hun toekomstige beroepsleven. De mogelijkheid tot het extrapoleren van scores van het toetsdomein naar het competentiedomein komt neer op de operationalisering van de competentie die gemeten wordt. Concreet wil dit zeggen dat het opstellen van een goede domeinbeschrijving in de ontwikkelingsfase van de toets - waarbij ook experts en vertegenwoordigers uit het werkveld betrokken worden - de extrapoleerbaarheid van toetsscores vergroot. Het gaat dan zowel over de mogelijkheid om scores op een toets te extrapoleren naar het competentiedomein, als over de mogelijkheid om deze vervolgens te extrapoleren naar het praktijkdomein. Deze stap is overigens vergelijkbaar met wat Chapelle (2012) en Chapelle, Enright, and Jamieson (2010) 'domeinbeschrijving' noemt.

Niet alleen de extrapoleerbaarheid, maar ook de generaliseerbaarheid van scores hangt overigens af van beslissingen en stappen die eerder in het proces genomen worden: in eerste instantie bij de voorbereiding en de opzet en ontwikkeling van de toets, maar ook bij de toetsafname en het scoren. Dit inzicht bracht ons ertoe om in ons kwaliteitskader de stappen die te maken hebben met het toetsdesign te scheiden van de elementen die te maken hebben met de representativiteit van scores ten aanzien van het toets- en competentiedomein.

Wat we meenemen

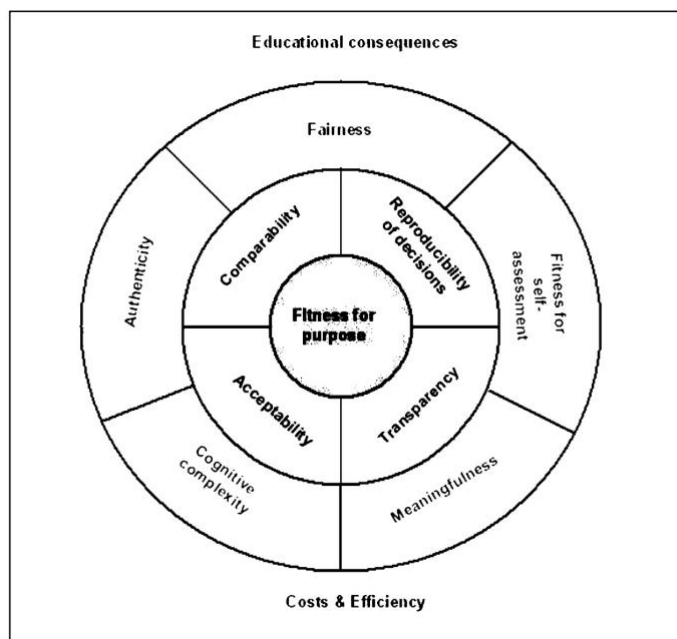
- We benadrukken in onze evaluatiematrix het belang van een bouwsteen 'domeinbeschrijving'.
- Ook 'toetsafname' nemen we – afzonderlijk van scores - expliciet als bouwsteen op.
- Om de interpretatieve benadering van Kane ook bruikbaar te maken voor een breder publiek dan alleen psychometrici (zie ook 1.2), werken we niet met de volledige argumentatiestructuur (gevolgtrekkingen én onderliggende assumpties), maar formuleren we eenvoudigweg 'voorwaarden' waaraan competentiebeoordelingen op grond van 'performance assessment'-technieken dienen te voldoen.

3.1.3 Op zoek naar een breder kwaliteitskader

Toetsen en toetsresultaten staan niet los van de context waarbinnen ze georganiseerd en gebruikt worden. De argumentatieve benadering van validiteit besteedt hier aandacht aan door te benadrukken 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 gevolgtrekkingen, namelijk het inschatten van de implicaties van de toetsresultaten, verwijst naar de ruimere context waarbinnen toetsen plaatsgrijpen. Toch besteedt de argumentatieve benadering van validiteit in de eerste plaats aandacht 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.

L. Baartman (2008) pleit in het kader van competentiebeoordelingen voor een kwaliteitskader dat verder reikt dan de traditionele, psychometrische noties van betrouwbaarheid en validiteit. Dit past binnen de edumetrische benadering, die een alternatief vormt om de specifieke karakteristieken van de beoordelingscultuur 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. L. Baartman (2008) en L. Baartman et al. (2006) vullen dit ruimere kwaliteitskader in op grond van het zogenaamde wiel van competentiebeoordeling (zie [Figuur 3.2](#)). Hiermee bouwen ze voort op het werk van onder andere Linn, Baker, and Dunbar (1991).



Figuur 3.2: Wiel van competentieassessment (bron: Baartman et al., 2006, p. 166)

Centraal in het wiel staat 'fitness for purpose' (geschiktheid voor onderwijsdoelen), wat impliceert dat een competentiebeoordeling maar kwaliteitsvol kan zijn indien zij geschikt is voor het doel waarvoor zij wordt ontwikkeld. Het betreft een principe dat ook in onze evaluatiematrix verder uitgewerkt wordt. Het wiel omvat verder nog elf andere kwaliteitscriteria:

- acceptatie of de mate waarin alle betrokken partijen de beoordeling accepteren
- authenticiteit of de mate van overeenkomst tussen de beoordeling en de eigenlijke praktijk
- betekenisvolheid of de mate waarin de beoordeling waarde heeft voor de persoonlijke ontwikkeling en de beroepsontwikkeling
- cognitieve complexiteit of de overeenkomst in (denk)processen met de eigenlijke praktijk
- eerlijkheid of de mate waarin de beoordeling de gelegenheid biedt om alle bedoelde competenties te tonen en te beoordelen
- onderwijsgevolgen of de mate waarin de beoordeling een positieve invloed heeft op het leerproces en de motivatie
- herhaalbaarheid van beslissingen of de mate van accuraatheid of betrouwbaarheid
- geschiktheid voor zelfbeoordeling of de mate waarin de beoordeling zelfsturend leren stimuleert
- efficiëntie en kosten of de mate waarin de beoordeling efficiënt is en organiseerbaar binnen de beschikbare tijd en op basis van de beschikbare financiële middelen
- transparantie of de mate waarin alle betrokkenen het beoordelingsproces (goed) begrepen hebben
- vergelijkbaarheid of de mate waarin de beoordeling (taken, criteria en context) consistent is opgezet

Het belang van criteria als authenticiteit, eerlijkheid en transparantie in deze opsomming, erkennen we reeds bij de probleemstelling en begripsafbakening (zie 2.5.3.). Omwille van dat belang nemen we deze drie alternatieve kwaliteitscriteria – samen met fitness for purpose- mee bij de uitbouw van de matrix.

Een interessant raamwerk, naast dat van Baartman, wordt ons aangereikt door C. P. Newhouse (2011). Teruggrijpend op het werk van Kimbell et al. (2007) onderscheidt hij een aantal dimensies die er samen voor zorgen dat een bepaalde toets in een bepaalde context haalbaar is. De eerste dimensie ‘beheersbaarheid’ duidt op de handelbaarheid van de toetsafname, terwijl de tweede, ‘technische dimensie’, heel specifiek verwijst naar technische uitdagingen die voortvloeien uit het inzetten van ICT voor ‘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 gaan in op betrouwbaarheid enerzijds en validiteit anderzijds. Een uitgebreide analyse naar de haalbaarheid gebeurt uiteindelijk op basis van deze verschillende elementen, en omvat steeds een afweging tussen de verschillende dimensies, rekening houdend met het doel van de toets.

Het afwegen van aspecten komt overigens ook heel duidelijk naar voren in de formule die Der Vleuten and Schuwirth (2005) voorstellen ten aanzien van de kwaliteit en bruikbaarheid van toetsen. Deze auteurs stellen namelijk dat de bruikbaarheid van een beoordelingstool afhankelijk is van vijf verschillende factoren. De formule die zij in dit verband naar voor schuiven is de volgende:

bruikbaarheid van een beoordelingstool = validiteit x betrouwbaarheid x aanvaardbaarheid x impact op het onderwijs(leerproces) x kosten-effectiviteit.

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 kwaliteitsvol men een toets(programma) kan opzetten Suzanne Lane and Stone (2006). In het geval leerlingen bijvoorbeeld elk veertig taken moeten uitvoeren die vervolgens elk door twintig beoordelaars worden gescoord, zal dit tot zeer betrouwbare scores leiden. 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

- Bij het bepalen van de kwaliteit van een toets dient rekening te worden gehouden met het doel waarvoor die werd ontwikkeld ('fitness for purpose').
- De argumentatieve benadering van validiteit is ruimer en omvat een breder kwaliteitskader, waarin er meer expliciete aandacht is voor 'edumetrische' aspecten zoals bijvoorbeeld authenticiteit, transparantie en eerlijkheid van de toets(taken).
- In het geval van grootschalige 'performance assessments' is het cruciaal dat een ontwikkelde toets ook haalbaar en bruikbaar is. Kosten en efficiëntie zijn een belangrijk criterium om de kwaliteit van competentiebeoordelingen te bepalen.
- De kwaliteit van een toets resulteert uit het onderling afwegen van verschillende criteria.

Het resultaat van deze theoretische verkenning was de identificatie van noodzakelijke bouwstenen om, in het kader van grootschalige competentietoetsing met het oog op monitoring op systeemniveau, kwaliteitsvolle 'performance assessments' op te kunnen zetten (zie hoofdstuk 5).

3.2 Literatuurstudie

Na lectuur van een aantal basiswerken over 'performance assessment' en over de argumentatieve benadering van validiteit (zie 3.1.) kwamen we tot de conclusie dat (1) er in de voorbije jaren veel onderzoek is uitgevoerd naar 'performance assessment'; en dat (2) 'kwaliteit' vaak in enge, psychometrische zin wordt opgevat, zonder veel oog voor de context waarin de toets plaatsvindt. Tegen die achtergrond voerden we een systematische literatuurstudie uit naar de kwaliteitseisen gesteld aan 'performance assessments' van competenties, om zo verder invulling te geven aan de ontwikkeling van de evaluatiematrix. Daarnaast was het doel van de literatuurstudie om mogelijke uitdagingen en problemen die zich aandienen bij het opzetten van grootschalige competentietoetsen via 'performance assessment', te identificeren, en te zoeken naar relevante input vanuit potentieel inspirerende buitenlandse praktijkvoorbeelden.

De overkoepelende vragen van de literatuurstudie waren:

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?

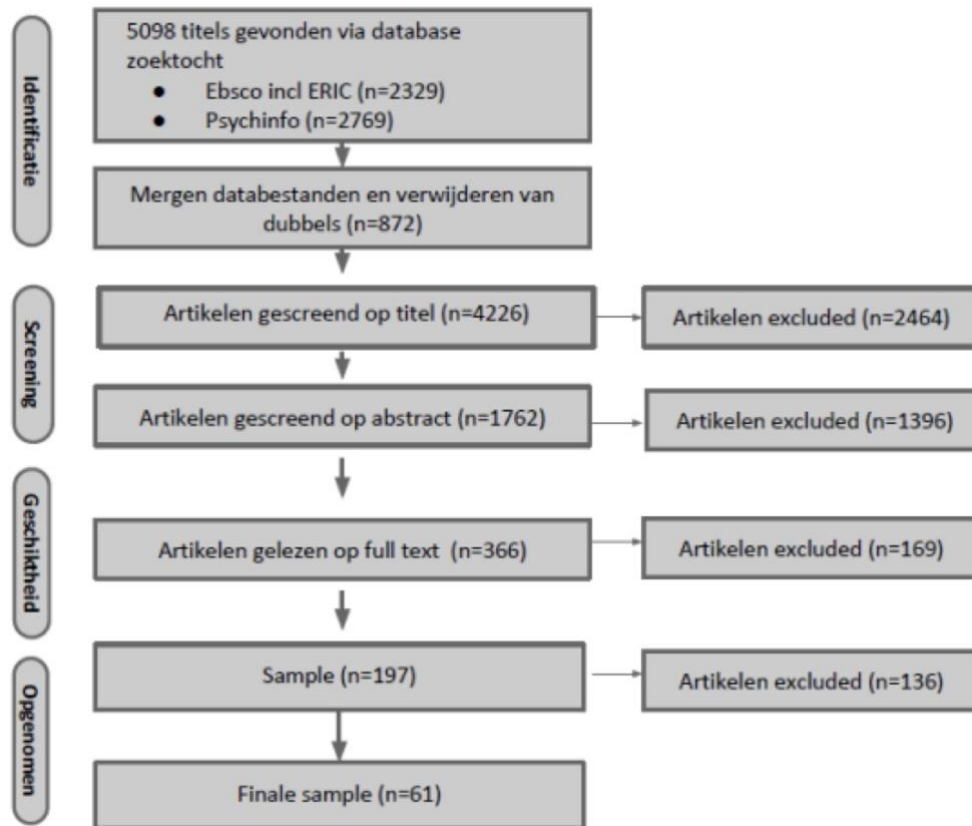
Welke zijn de implicaties voor het beoordelingsbeleid en voor de beoordelingen zelf?

We formuleerden de vragen voor de literatuurstudie bewust breed. Ten eerste beperkten we ons niet tot het lager en secundair onderwijs, maar namen we ook het hoger onderwijs mee. Een eerste verkenning van de literatuur leerde immers dat een belangrijk aandeel van de literatuur rond 'performance assessment' zich binnen de context van het hoger (i.c. medisch) onderwijs situeert. Hieruit kunnen echter ook lessen getrokken worden voor beide andere onderwijsniveaus. Ten tweede werd bepaald om de review niet te beperken tot louter grootschalige toetsen met het oog op systeemmonitoring. Beweegreden hiervoor was dat ook uit meer kleinschalig opgezette toetsen of toetsen die andere doelstellingen hebben dan monitoring op systeemniveau, zinvolle suggesties meegenomen kunnen worden over kwaliteitsvolle grootschalige toetsen met een 'performance assessment'-component.

Gelet op de gegeven definitie van het begrip 'competentie' (zie [Hfdst. 2.5.1](#)), namen we in de literatuurreview alleen studies mee waarin het beoogde construct meer omvat dan zuiver kennis, louter vaardigheden of enkel attitudes. Gezien de gekozen definitie van 'performance(-based) assessment' kwamen alleen studies in aanmerking die rapporteren over toetsvormen die taken aanbieden met een zekere graad van authenticiteit.

Qua databanken raadpleegden we enerzijds Ebsco (ERIC, Business Source Premier en e-book collection) en anderzijds PsychInfo. Deze databanken bevatten internationale, voornamelijk Engelstalige, tijdschriften met collegiaal getoetste artikelen. De zoektocht in deze databanken leverde aanvankelijk 5.092 artikelen op, die op basis van verschillende selectierondes herleid werden tot een uiteindelijke sample van 61 artikelen. [Figuur 3.3](#) geeft een overzicht van deze verschillende selectiefasen. Voor meer details over de methodologie op het vlak van selectie en codering verwijzen we naar het onderzoeksrapport De Maeyer et al. (2016).

We analyseerden en codeerden deze 61 artikelen, met als bedoeling: problemen die zich voordoen met betrekking tot de kwaliteit van 'performance assessments' van competenties, te identificeren en de essentie ervan te rapporteren; deze problemen vervolgens te situeren in één of meerdere van de bouwstenen van de evaluatiematrix; en de oplossing(en) voor de gestelde problematiek die in het artikel aan bod komen, helder te stellen.



Figuur 3.3: Overzicht selectiestappen en weerhouden sample

De inzichten die we verkregen uit de literatuurstudie, werden meegenomen in de ontwikkeling van de evaluatiematrix en de uitwerking van essentiële uitdagingen (zie ook Hoofdstukken 5 en 6).

3.3 Verzamelen en analyseren van buitenlandse praktijkvoorbeelden

Een derde onderzoekslijn, die de evaluatiematrix mee voedde, was de beschrijving en analyse van praktijkvoorbeelden in buitenlandse onderwijscontexten die (deelcomponenten van) competenties toetsen via 'performance assessment'. In tegenstelling tot de fase van de literatuurreview, ging aandacht uit naar grootschalige initiatieven gericht op het bewaken van leerlingenprestaties op systeemniveau. De bedoeling van de analyse van deze praktijkvoorbeelden was dubbel: enerzijds het verder verfijnen en toetsen van de evaluatiematrix op praktische hanteerbaarheid, anderzijds het zoeken naar evidentie om typische probleemgebieden en oplossingen in verband met grootschalige competentietoetsen op grond van 'performance assessment', te illustreren.

De screening en de selectie van de potentiële praktijkvoorbeelden gebeurden in drie stappen. Met het oog op een inventarisatie van bestaande systemen voerden we in een eerste stap een systematische screening uit van activiteiten en producten van buitenlandse overheidsinstanties en niet-overheidsinstanties, onderzoeksinstituten, universiteiten en agentschappen, die zich toelagen op de ontwikkeling en/of evaluatie van grootschalige

competentietoetssystemen. Daarnaast vroegen we ook aan verschillende experts input over mogelijke interessante praktijkvoorbeelden.

Met de inzichten gewonnen uit de systematisch opgezette literatuurstudie, konden we het uitgebreide overzicht van potentieel interessante praktijkvoorbeelden op een meer gefundeerde wijze screenen. Bij deze oefening plaatsten we drie expliciete criteria voorop, waaraan praktijkvoorbeelden moesten voldoen om geselecteerd te worden. De praktijkvoorbeelden dienden beoordelingssystemen te zijn

- met als doel: het bewaken van leerlingenprestaties op systeemniveau;
 - m.a.w. duidelijke link met een nationaal curriculum/nationale standaarden
 - m.a.w. grootschalig
- die zich richten op het beoordelen van competenties (waarbij competenties geïntegreerd worden opgevat);
- en die gebruik maken van ‘performance assessment’.

De initiële selectie van twaalf potentiële praktijkvoorbeelden werd uiteindelijk herleid tot zeven praktijkvoorbeelden (zie [Tabel 3.1](#)). Selectie-argumenten waren ofwel inhoudelijk van aard (bv. dat de focus van het beoordelingssysteem te veel afwijkt of dat de mate waarin het systeem ‘performance assessment’-technieken aanwendt te gering of nihil is), ofwel pragmatisch (bv. dat we de taal waarin de informatiebronnen zijn opgesteld, niet beheersen, of het ontbreken van reacties van verantwoordelijken van de toetsprogramma’s). Voor meer duiding bij de verschillende selectierondes verwijzen we naar het onderzoeksrapport (De Maeyer et al. 2016).

Tabel 3.1: Finale pool van praktijkvoorbeelden van grootschalige toetsprogramma’s waarbij gebruik wordt gemaakt van ‘performance assessment’.

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

Met het oog op de verdere verfijning van de evaluatiematrix en de inventarisatie van uitdagingen en oplossingen bij grootschalige ‘performance assessments’, analyseerden we

algemene en technische rapporten en andere relevante documenten. Daarnaast organiseerden we met betrekking tot elk van de praktijkvoorbeelden, twee diepte-interviews. Terwijl in ronde 1 de verschillende bouwblokken van de evaluatiematrix systematisch werden doorlopen, gingen we in de tweede ronde dieper in op een aantal cruciale aspecten. Met het oog op de interviews werkten we, volgens het stramien van de evaluatiematrix, een vragenpool uit. Uit deze pool selecteerden we voor elke toets afzonderlijk en naargelang de hiaten en vraagtekens, een individuele set vragen.

De inzichten die we verkregen uit deze analyse, werden meegenomen in de ontwikkeling van de evaluatiematrix en de uitwerking van essentiële uitdagingen. Ze werden met andere woorden mee verwerkt in hoofdstukken 5 en 6 van deze publicatie.

4 Buitenlandse voorbeelden van grootschalige ‘performance assessments’

De analyse van de buitenlandse praktijkvoorbeelden, waarover we in hoofdstuk 3 rapporteerden, stelde ons in staat de evaluatiematrix verder te verfijnen en waar nodig aan te vullen. De interviews gaven ons verdere voeling met de afwegingen die toetsontwikkelaars moeten maken in de zoektocht naar kwaliteitsvolle oplossingen voor de problemen waarmee ze zich geconfronteerd zien. Net omdat de praktijkvoorbeelden een belangrijke bron voor de bevindingen van het onderzoek waren, dat aan de basis van deze publicatie ligt, geven we in dit hoofdstuk een beknopte beschrijving van deze toetsystemen. Op die manier biedt dit hoofdstuk ook de nodige achtergrond voor hoofdstukken 5 en 6, waar regelmatig naar de buitenlandse toetsystemen wordt verwezen.

Voor elk buitenlandse praktijkvoorbeeld beschrijven we eerst de context. Vervolgens zoomen we in op de specifieke toets die we analyseerden en geven we inzicht in verschillende van de bouwstenen van de toets, zoals doel, opzet, ontwikkeling, toetsafname, scoren en rapportering.

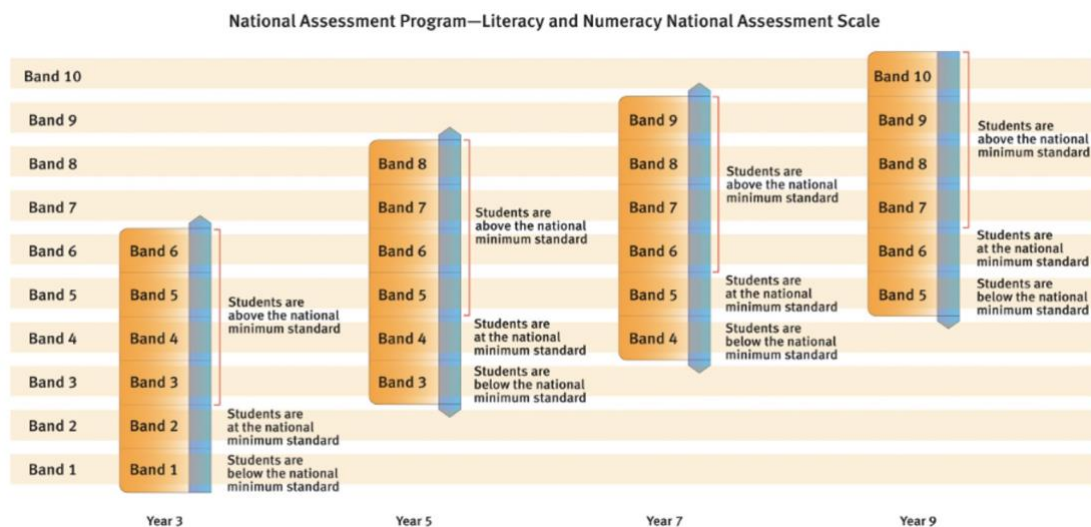
4.1 Australië: National Assessment Program – Literacy and Numeracy (NAPLAN)

In Australië wordt sinds 2008 jaarlijks het National Assessment Program – Literacy and Numeracy (NAPLAN) afgenomen. NAPLAN is een toetsprogramma dat als doel heeft de vooruitgang van de kennis en vaardigheden van alle leerlingen in het derde leerjaar (8-9 jaar oud), vijfde leerjaar (10-11 jaar oud), zevende leerjaar (12-13 jaar oud) en negende leerjaar (14-15 jaar oud) met betrekking tot lezen (‘Reading’), schrijven (‘Writing’), taalconventies (‘Language Conventions’) en wiskundige geletterdheid (‘Numeracy’) na te gaan. De toetsen geven een breed beeld van de taal- en wiskunde-aspecten die vervat zitten in de curricula van alle deelstaten en regio’s van Australië. Zo wil men te weten komen in welke mate de nationale standaarden voor deze vier gebieden behaald worden.

ACARA (Australian Curriculum, Assessment and Reporting Authority) ontwikkelt de NAPLAN-toetsen in samenwerking met lokale overheden, private schoolsectoren en de federale overheid. Elke staat/territorium in Australië heeft een eigen instantie die bevoegd is voor het afnemen van de toetsen, het verzamelen van gegevens en het bezorgen van rapporten. De focus van de kwaliteitsbewaking ligt op verschillende niveaus: primair op het niveau van de deelstaten en regio’s en de scholen, maar daarnaast worden er ook resultaten teruggekoppeld aan individuele leerlingen en zijn er ook data beschikbaar op het federale niveau. De gegevens worden beschikbaar gesteld in de vorm van openbare rapporten (resultaten deelstaten & regio’s en federale niveau), de My School-website (resultaten op schoolniveau) en individuele leerlingrapporten voor elke deelnemende leerling.

In ACARA - Australian Curriculum (2014), de toets die we voor deze publicatie bestudeerden, moesten de leerlingen voor de schrijftoets een overtuigende tekst schrijven (‘Persuasive Writing’). Een overtuigende tekst wordt omschreven als een tekst die als belangrijkste doel heeft om een mening te presenteren en probeert om de lezer te overtuigen. Het meten van schrijfvaardigheid werd met andere woorden geoperationaliseerd door één schrijfggenre te toetsen, wat ook wel geleid heeft tot kritiek

(zie hoofdstuk 5, 5.2.7.). De toets was voor alle deelnemende leerlingen, m.a.w. voor de leerlingen uit het derde, vijfde, zevende en negende leerjaar, dezelfde. Iedereen kreeg hetzelfde stimulusmateriaal (de zogenaamde ‘writing prompt’). Het thema was: *“It is cruel to keep animals in cages. What do you think? Do you (dis)agree? Perhaps you can think of ideas for both sides of this topic”*. De afname en het scoren van de schrijftoets gebeurde door getrainde toetsassistenten en beoordelaars, die door de lokale toetsautoriteit op het niveau van de deelstaat of ‘territory’ waren aangesteld. Hoewel elk(e) deelstaat en regio een eigen instantie heeft die bevoegd is voor de toetsafname, garandeert NAPLAN een hoge mate van standaardisering dankzij een nationaal overeengekomen, systematisch ontwikkeld ‘National Protocol for Test Administration’. Alle beoordelaars kregen dezelfde, intensieve training, waarbij ingegaan werd op zowel de procedure als de te gebruiken materialen. Elk schrijfstuk werd (slechts) door één beoordelaar beoordeeld. Alle beoordelaars maakten voor het beoordelen van de schrijfstukken gebruik van dezelfde rubric. Alle NAPLAN-schrijftaken werden online gescoord. Er werden controletaken ingezet om de nauwkeurigheid van beoordelaars te monitoren. Onervaren beoordelaars werden opgevolgd en kregen hertraining indien nodig. De resultaten van NAPLAN ‘Persuasive Writing’ werden gerapporteerd op een gestandaardiseerde, nationale prestatieschaal (zie [Figuur 4.1](#)). De schaal was opgedeeld in tien velden of ‘proficiency bands’; deze vormen een weerspiegeling van de toenemende complexiteit op vlak van kennis en vaardigheden van het derde tot het negende leerjaar.



Figuur 4.1: De NAPLAN schaal (bron: ACARA, 2014, p.v.)

Sinds de NAPLAN van 2015 werd, onder andere na consultatie met curriculumexperts, een andere taak voorzien voor de leerlingen van het derde en het vijfde leerjaar enerzijds en de leerlingen van het zevende en negende leerjaar anderzijds. De taak behoort wel steeds tot hetzelfde genre. Sinds 2015 wordt bovendien niet uitsluitend voor het genre overtuigend schrijven gekozen. De schrijftaak van NAPLAN omvat ofwel een narratieve, ofwel een overtuigende schrijfoefening. Er wordt op voorhand niet aangekondigd voor welk genre werd geopteerd.

4.2 Australië: National Assessment Program (NAP)

Het doel van de NAP 'sample assessments' is te kunnen rapporteren over de vooruitgang die leerlingen maken in het behalen van de nationale onderwijsdoelstellingen. Dit gebeurt op nationale basis, via toetsen die afgenomen worden bij steekproeven leerlingen uit steekproeven van scholen uit het lager onderwijs. De verantwoordelijkheid voor de NAP 'sample assessments' ligt bij ACARA, dat ook instaat voor NAPLAN. De NAP 'sample assessments' toetsen vaardigheden en begrip van leerlingen uit het zesde en het tiende leerjaar, met betrekking tot volgende drie domeinen: wetenschappelijke geletterdheid, maatschappij en burgerschap, en informatie- en communicatietechnologie (ICT). De NAP 'sample assessments' zijn gestart in 2003 en elk domein wordt driejaarlijks in kaart gebracht: wetenschappelijke geletterdheid in 2003, 2006, 2009 en 2012; maatschappij en burgerschap in 2004, 2007, 2010 en 2013; en ICT-geletterdheid in 2005, 2008, 2011 en 2014.

ACARA - Australian Curriculum (2014) hanteerde de volgende definitie van 'ict-literacy': *"The ability of individuals to use ICT appropriately to access, manage and evaluate information, develop new understandings, and communicate with others in order to participate effectively in society"*. NAP-ICTL was zodanig opgesteld dat het de typische dagdagelijkse toepassing van ICT weerspiegelde. De toets bestond uit negen scenario-gebaseerde modules, die elk een lineaire narratieve sequentie volgden. Elke leerling kreeg, op toevalsbasis, vier modules toegewezen. Aan NAP – ICTL 2014 nam een representatieve steekproef van 649 scholen deel met in totaal 10.562 leerlingen, wat neerkwam op 87% van de getrokken leerlingen uit het zesde leerjaar en 77% van de getrokken leerlingen uit het tiende leerjaar. Centraal getrainde toetsassistenten namen de toets in de geselecteerde scholen af. Een steekproef van 5% van de deelnemende scholen werd bezocht door getrainde kwaliteitsmonitoren. Via observatie van de toetsassistent, gingen de kwaliteitsmonitoren de uniformiteit en de consistentie van de afnameprocedures in de deelnemende scholen na. Een pool van beoordelaars stond in voor het scoren van de toetsen. Tijdens het scoringsproces zelf werden de antwoorden ofwel automatisch gescoord, ofwel bewaard en later centraal gescoord. Voor elk verschillend item en taaktype werd een afzonderlijke scoringsprocedure en -tool gebruikt. 10 % van de antwoorden werd dubbel gescoord door de aangestelde coördinator. In het geval van inconsistente scores werden de beoordelaars hertraind met betrekking tot dat specifieke item en werden de antwoorden opnieuw gescoord. In totaal werden er 133 items gebruikt om via IRT (1-parameter model) een unidimensionele schaal te bekomen (de 'NAP-ICTL proficiency scale'). De items en taken in de 'trend modules' konden gebruikt worden als link items. Via 'common item equating' werd de schaal van 2014 geëquivalet met die uit 2011. Over de resultaten werd op drie verschillende manieren gerapporteerd. In eerste instantie op basis van de NAP-ICTL-schaal (gemiddelde schaalscores voor ICT-geletterdheid). Ten tweede als percentage leerlingen binnen de zes onderscheiden bekwaamheidsniveaus, vergezeld met een beschrijving van de ICT-bekwaamheden geassocieerd met dat bepaalde niveau. En ten derde, onder de vorm van het percentage leerlingen dat de prestatiestandaard ('Proficient Standard') haalt.

4.3 Nederland: Periodieke Peiling van het Onderwijsniveau (PPON)

In 1986 startte de Nederlandse Minister van Onderwijs, Cultuur en Wetenschappen (OCW) het project Periodieke Peiling van het Onderwijsniveau (PPON), met als belangrijkste doel om een evaluatiekader te bieden voor de kerndoelen van het basisonderwijs. Het peilingsonderzoek had als doel om uitspraken te doen over het bereikte niveau op systeemniveau; rapportering op school- of leerlingniveau werd niet beoogd. Tot 2014 voerde het Centraal Instituut voor Toetsontwikkeling (Cito) het peilingsonderzoek in opdracht van het Ministerie OCW uit. Cito nam de peilingsonderzoeken in hoofdzaak af bij steekproeven leerlingen einde basisonderwijs (jaargroep 8, leeftijd 11-12 jaar), maar voor andere leerdomeinen vond ook onderzoek plaats in jaargroep 5 (8-9 jaar) en op scholen voor speciaal basisonderwijs. Sinds 2014 ligt de regie van het peilingsonderzoek bij de Onderwijsinspectie in het project 'Peil.onderwijs', dat in brede zin de kennis, vaardigheden en houding van leerlingen aan het einde van het primair onderwijs in kaart brengt.

In 2009 vormde schrijfvaardigheid van leerlingen één van de foci van de periodieke peiling in Nederland. Het was één van de aspecten die getoetst werden binnen van het leergebied 'schrijven'. Naast de schrijfopdracht(en) kregen leerlingen ook toetsen spelling, interpunctie, tekstrevisie, grammatica, zinsontleding en woordbenoeming. Binnen schrijfvaardigheid onderscheidde men een inhoudelijke, structurele, stilistische en communicatieve component. Verschillende genres kwamen aan bod; de schrijfopdrachten resulteerden in informatieve, instructieve, verhalende of overtuigende teksten. De schrijfopdrachten streefden naar een hoge graad van authenticiteit; men vroeg leerlingen bijvoorbeeld om een telefonische boodschap door te geven of een briefje op te stellen aan buurtbewoners over de verdwijning van een kat. De opdrachten bestonden telkens uit drie onderdelen: het uitgangsmateriaal (meestal een 'stel-je-voor-situatiebeschrijving' aangevuld met bronnenmateriaal, tekeningen of foto's), een taakinstructie met een expliciet geformuleerde schrijfopdracht, en de schrijfopdracht. Bij het ontwerpen van de schrijfopdrachten werd ernaar gestreefd typische en levensechte taken te ontwikkelen, relevant en betekenisvol voor leerlingen uit groep 5 en groep 8. De totale pool taken in de toets van 2009 omvatte twaalf schrijfopdrachten: leerlingen in groep 5 kregen één of twee opdrachten (45 min); leerlingen in groep 8 meestal vier (90 min). Cito maakte voor de peiling schrijfvaardigheid gebruik van taken uit de peiling van 1999, onder andere met het oog op vergelijking van de resultaten.

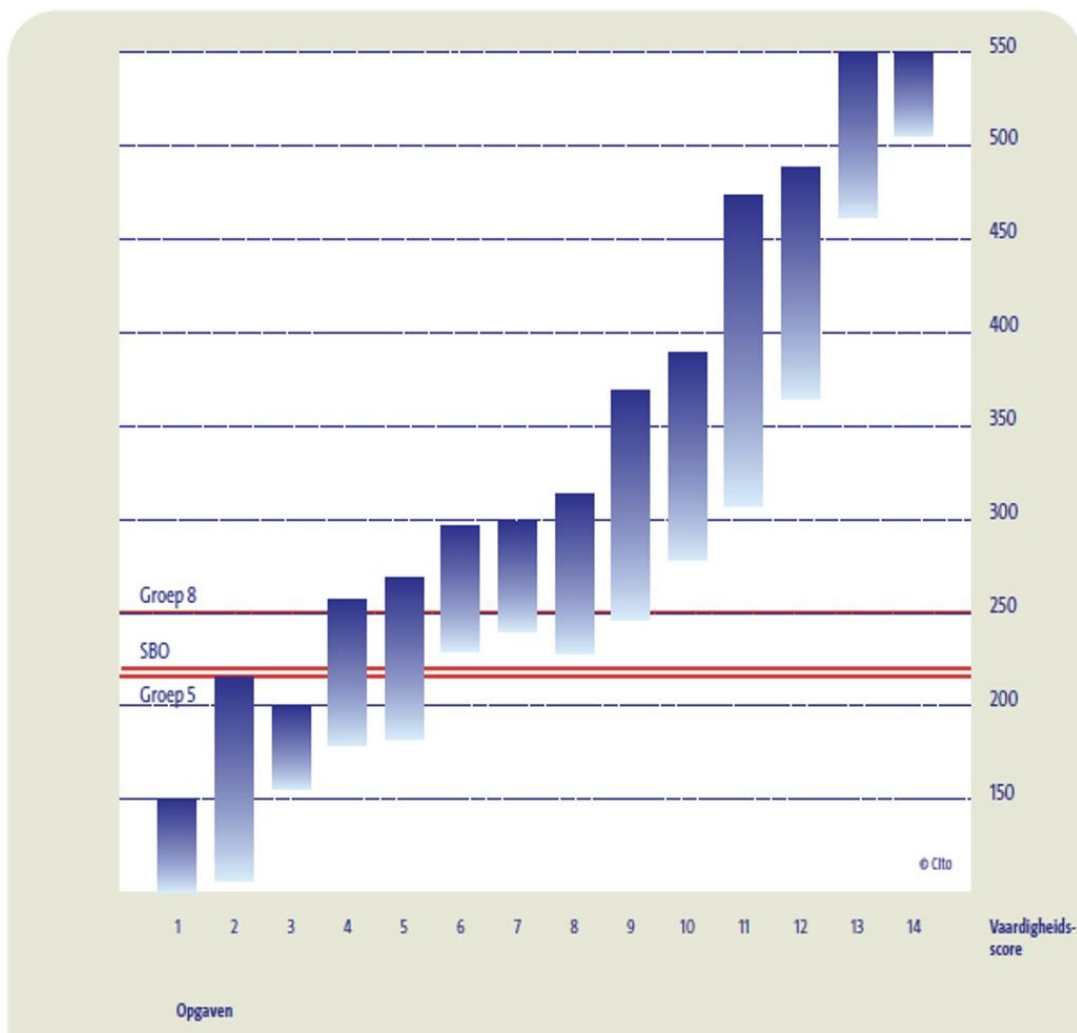
Voorbeeld van enkele taken voor PPON Schrijfvaardigheid 2009

Afspraak (informatief). De leerling schrijft een briefje aan zijn of haar moeder. Het doel is een telefonische boodschap die via een cd-rom ten gehore wordt gebracht door te geven.

Beste boek (betogend). De leerling schrijft een brief aan de juffrouw van de bibliotheek en vertelt haar van welk boek of welke film hij of zij het meeste houdt. De leerling moet daarin vertellen waar het over gaat, wat er zo goed aan is en waarom.

Brievenbus (instructief). De leerling leest eerst drie brieven van kinderen die een relationeel probleem hebben en om advies vragen. Vervolgens maakt de leerling een keuze voor één van de drie problemen en schrijft een briefje met advies en tips om te helpen.

De peiling vond plaats op twee momenten in 2009: in het voorjaar voor groep 8 en de eindgroep van het speciaal basisonderwijs en in het najaar voor groep 5. De leerlingen uit deze jaargroepen werden geselecteerd via een tweetrapssteekproeftrekking, waarbij eerst scholen werden getrokken en vervolgens, binnen die scholen, leerlingen. De afname van de schrijfopdrachten gebeurde in groep 5 onder leiding van een centrale toetsassistent, in groep 8 onder leiding van de eigen leerkracht (voor twee van de opdrachten) en de toetsassistent (voor de overige twee opdrachten). Vuistregel hierbij was dat, met het oog op standaardisering, enkel taken die eenvoudig uit te voeren waren, aan de leerkrachten werden overgedragen. Taken die bijvoorbeeld audiofragmenten omvatten, werden afgenomen door toetsassistenten. Leerkrachten dienden een afnameprotocol in te vullen, dat een indicatie kon geven van wanneer er iets mis was gegaan met de toetsafname. Het beoordelen van de schrijfstukken was de verantwoordelijkheid van centraal ingehuurde, getrainde leraren ($n = 50$); elke tekst werd door twee van hen beoordeeld. Er waren meerdaagse trainingsprogramma's voorzien, waarin bijvoorbeeld aandacht was voor oefeningen waarbij schaalpunten geïllustreerd worden aan de hand van goede en slechte schrijffragmenten van leerlingen. Om alle opgaven op één gemeenschappelijke schaal uit te drukken, maakte men gebruik van het één-parameter logistisch model (OPLM; IRT). PPON rapporteerde op landelijk niveau en niet over individuele leerlingen of afzonderlijke scholen. Men beeldde de schrijfvaardigheid van de leerlingen en de moeilijkheidsgraad van de opgaven telkens in één figuur af (de vaardigheidsschaal; zie [Figuur 4.2](#) hieronder) (Kuhlemeier et al. 2013). Aan de hand van een set voorbeeldopgaven rapporteerde men zo voor ieder aspect van schrijfvaardigheid afzonderlijk (inhoud, organisatie en structuur, stijl en communicatie) over welke vaardigheid (zeer) zwakke (P10 en P25), gemiddelde (P50) en (zeer) goede schrijvers (P75 en P90) beschikken in termen van aantal opgaven dat ze onvoldoende, matig en goed beheersen. Per jaargroep werd ook een algeheel niveau van schrijfvaardigheid berekend op basis van alle 193 opgaven overheen de inhoudelijke, organisatorische, stilistische en communicatieve kwaliteit.



Figuur 4.2: Vaardigheidsschaal 'Organisatorische en Structurele kwaliteit' (bron: Kuhlemeier, van Til et al., 2013, p. 83).

4.4 Nieuw Zeeland: National Monitoring Study of Student Achievement (NMSSA)

De National Monitoring Study of Student Achievement (NMSSA) is een peilingstoets in opdracht van het Ministerie van Onderwijs van Nieuw-Zeeland. Hij wordt uitgevoerd door de 'Educational Assessment Research Unit' (EARU) van de Universiteit van Otago en de 'New Zealand Council for Educational Research' (NZCER). De NMSSA startte in 2012 en heeft als doel een momentopname te maken van leerlingenprestaties in relatie tot het Nieuw-Zeelandse curriculum. NMSSA wordt in een vijfjarige cyclus herhaald met het oog op het identificeren van trends. Rapportering over de resultaten gebeurt jaarlijks; enkel op nationaal niveau, niet op het niveau van individuele leerlingen of scholen. Bij de rapportering focust men op de vooruitgang van sleutelpopulatiegroepen van Māori- and Pasifika-leerlingen, en leerlingen met speciale onderwijsnoden.

NMSSA focust elk jaar op twee leergebieden uit het Nieuw-Zeelandse curriculum. In 2016 zoomde NMSSA in op technologie en stak men ook tijd in de heranalyse van reeds

verzamelde gegevens. Kunst en luisteren en kijken vormden de kern van het beoordelingsprogramma in 2015. In 2014 lag de focus op sociale wetenschappen en lezen. Het jaar daarvoor bracht de NMSSA de leerlingenprestaties op vlak van gezondheid en lichamelijke opvoeding enerzijds en wiskunde en statistiek anderzijds in kaart. In het eerste afnamejaar 2012 stonden schrijven (Engels) en wetenschappen in de kijker.

We beschrijven in de volgende paragraaf het leergebied 'Gezondheid en Lichamelijke Opvoeding' (Educational Assessment Research Unit & NZCER - New Zealand Council for Educational Research 2014), aangezien daar onze focus initieel op lag. Tijdens de interviews kregen we ook interessante inzichten over het leergebied 'Kunst', dat in 2015 aan de beurt was. De relevante inzichten over dit laatste gebied worden aangestipt in hoofdstukken 5 en 6. De doelstellingen van het leergebied 'Gezondheid en Lichamelijke Opvoeding' omvatten het welbevinden van leerlingen, van anderen en van de maatschappij door het leren in en over contexten die gerelateerd zijn aan gezondheid en beweging. Gelet op de focus van 'performance assessment', ging onze aandacht in de eerste plaats uit naar dat deel van de peilingstoets waar leerlingen een assortiment van bewegingsvaardigheden moesten tonen. Het 'performance assessment'-deel van de toets bestond uit drie taken: twee taken waarmee de bewegings- en strategische actievaardigheden van de leerlingen werden getoetst tijdens een authentiek spel ('Rippa Tag' en 'Rua Tapahwa'); en een derde taak die een opeenvolging van bewegingen beoordeelde. De taken werden afgenomen en gescoord door getrainde toetsassistenten en beoordelaars.

Toelichting bij de taken van de NMSSA (2014)

In het eerste spel, 'Rippa Tag', was het de bedoeling dat twee leerlingen in een beperkte ruimte het velcro lint van de gordel van de tegenstander trokken. Tijdens deze activiteit ging de aandacht van de beoordelaars onder andere uit naar wendbaarheid en beweeglijkheid, evenwicht en strategische actievaardigheden. In het tweede spel, 'Rua Tapahwa' moesten leerlingen de bal in het veld van de tegenstander (een getrainde beoordelaar) werpen zodat deze er niet in zou slagen de bal op te vangen na één keer botsen. Leerlingen werden geëvalueerd op werpen, vangen, 'defensive tracking' en strategische actievaardigheden. In de derde activiteit moesten de leerlingen een bewegingssequentie creëren, waarbij ze gebruik moesten maken van verschillende hulpmiddelen. Nadat ze de bewegingssequentie hadden uitgevoerd werd aan de leerlingen gevraagd om een nieuwe sequentie uit te voeren, met name de vorige, inclusief een nieuwe beweging. Leerlingen werden o.a. geëvalueerd op hun controle van het materiaal, het gebruik van hun lichaam, de variaties in bewegingen etc.

De selectie van de toevalssteekproef gebeurde in NMSSA getrapt: eerst werden scholen getrokken, daarna - in de getrokken scholen - de leerlingen. Het projectteam trok zo een toevalssteekproef van respectievelijk 100 scholen voor het vierde leerjaar (8 – 9 jaar oud) en 100 scholen voor het achtste leerjaar (12 – 13 jaar oud). Uit elk van de scholen trokken ze in een tweede fase, opnieuw op toevalsbasis, 25 tot 28 leerlingen (3 reserve) voor deelname aan het schriftelijk gedeelte van de toets. Deze toets had overigens betrekking op een ander leergebied, m.n. dat van wiskunde en statistiek. De peiling 'Health and Physical Education' werd uitgevoerd bij een subset van 8 leerlingen uit de steekproef van 25 leerlingen, wat neerkomt op een totale steekproefgrootte van circa 800 leerlingen per

leerjaar. De toets werd afgenomen door centraal opgeleide toetsassistenten, die nauwgezette instructies meekregen. De antwoorden van de leerlingen op de toets worden op video en schriftelijk vastgelegd en elektronisch bijgehouden met het oog op het scoren. Alle beoordelaars waren ervaren leerkrachten, die daarenboven een specifieke training kregen. Elke taak werd door één beoordelaar beoordeeld. Wel werd gewerkt met een pool van een twintigtal beoordelaars, die tegelijkertijd scoorden. Er waren verschillende kwaliteitszorgprocedures voorzien om de variantie tussen deze beoordelaars binnen de perken te houden. Ook werd voor elk van de drie fysieke activiteiten waarmee de bewegingsvaardigheden van de leerlingen werden getoetst, een aparte rubric opgesteld. Deze rubric definieerde een set vaardigheden die een lage mate van competentie vertegenwoordigden; vervolgens een set die een middelmatige mate vertegenwoordigen; en ten slotte een reeks die een sterke mate vertegenwoordigen (zie [Figuur 4.3](#) voor een voorbeeld van een rubric van één van de drie taken).

Low Range	Mid Range	High Range
<p>Shows:</p> <ul style="list-style-type: none"> • Uses one strategy- does not try something new when one way doesn't work • Heavy on feet, flat on feet • Stumbles forward to grab and when repositioning body after snatch • Unable or slow to change direction effectively • Unbalanced- perhaps feet too far apart/base of support too wide • Tires, gives up • Running not dodging 	<p>Able to competently use:</p> <ul style="list-style-type: none"> • Quick, light feet • Dodge by pushing off outside foot • Defend the space/tag • Balanced so able to transfer weight fluidly • Re-position themselves to gain advantage e.g. moves towards to attack/moves away from opposition • Checking opposition 	<p>Able to competently use:</p> <ul style="list-style-type: none"> • Pivot • Rotate body to snatch and avoid opposition • Lower centre of gravity so able to change direction quickly e.g. crouched position when on attack • On balls of feet - readiness for movement • Quick decision making e.g. change direction or speed, anticipation of opposition moves/tactics • Uses both hands • Consistency over defence/attack • Competitive/shows commitment

Figuur 4.3: 'Scoring guide' van de Rippa Tag (bron: EARU & NZCER, 2014, p. 44).

Over de resultaten van de peiling van 'Health and Physical Education' werd op twee manieren gerapporteerd: via een schaal rond 'Critical Thinking in Health and Physical Education', die afgestemd was op de niveaus van het 'New Zealand Curriculum'; en via aparte descriptieve rapportering van het inzicht van leerlingen in welzijn en hun prestaties inzake bewegingsvaardigheden. Over het 'performance assessment'-gedeelte van de peilingstoets werd apart en louter beschrijvend gerapporteerd.

Descriptieve rapportering bewegingsvaardigheden NMSSA (2014)

Rippa Tag: proportion of students at both year levels scoring at:

- Level 'Student displays all/almost all aspects from high range movement list'
- Level 'Student displays a variety of aspects – mainly mid range with some high range movements'
- Level 'Student displays a few aspects from mid range with some low range movements'
- Level 'Student displays low range movements'

Rua Tapawhā: proportion of students at both year levels scoring at:

- Level 'Student displays all/almost all aspects from high range movement list'
- Level 'Student displays a variety of aspects – mainly mid range with some high range movements'
- Level 'Student displays a few aspects from mid range with some low range movements'
- Level 'Student displays low range movements'

Movement sequence: Students' aggregated scores for the three aspects of the movement sequences task are reported > proportion of students at both year levels scoring at

- low range (No response/don't know, does not complete 3 movements, no evidence of consistency, and no evidence of cooperative work)
- low-mid range (Includes at least 3 movements/ 1 element, 1 aspect of consistency, and 1 aspect of cooperative work)
- mid-high range (Includes 2-3 elements, 2 aspects of consistency, and 2 aspects of cooperative work)
- high range (Includes 4 or more elements, 3 aspects of consistency, and 3 aspects of cooperative work)

4.5 Schotland: Scottish Survey of Literacy and Numeracy (SSLN)

De Scottish Survey of Literacy and Numeracy (SSLN) werd in 2009 ontwikkeld om het Schotse curriculum ('Curriculum for Excellence' - CfE) te toetsen. SSLN is volledig afgestemd op het CfE; de toetsen zijn zo opgezet dat ze een weerspiegeling vormen van de curriculumdoelstellingen. SSLN wordt jaarlijks, alternerend voor geletterdheid (2012, 2014, 2016) en wiskundige geletterdheid (2011, 2013, 2015), afgenomen bij leerlingen uit het vierde jaar primair onderwijs (8-9 jaar), het zevende jaar primair onderwijs (11-12 jaar) en het tweede jaar secundair onderwijs (13-14 jaar), met het oog op het bewaken van de prestaties van leerlingen. Het betreft een momentopname van de verworvenheden op een welbepaald moment in de tijd. De focus ligt ook op het maken van vergelijkingen tussen afnamejaren. Er wordt enkel op systeemniveau gerapporteerd. De resultaten van de survey zetten ook aan tot verdere ontwikkeling op vlak van leren, lesgeven en toetsen doorheen de ontwikkeling van 'Professional Learning Resources'.

SSLN is een partnerschap tussen de Schotse regering, 'Education Scotland', het Schotse agentschap voor kwalificaties ('Scottish Qualifications Authority': SQA), de 'Association of Directors of Education in Scotland' en lokale autoriteiten. Alle Schotse scholen nemen deel aan SSLN; leerlingen worden op basis van toeval geselecteerd. Om de belasting voor scholen zo klein mogelijk te houden, selecteert men slechts drie leerlingen uit het vierde en zevende jaar primair onderwijs en 14 leerlingen uit het tweede jaar secundair onderwijs. Jaarlijks nemen per leerjaar ongeveer 4.000 leerlingen deel.

SSLN-2014 focuste op drie aspecten van geletterdheid: lezen, schrijven en luisteren & spreken. Schrijven werd getoetst op grond van twee schrijfproducten (per leerling) in 50% van de scholen; luisteren & spreken op basis van een groepsdiscussie onder leerlingen, in

40% van de scholen. De steekproef leerlingen voor SSLN-2014 omvatte per school 2 leerlingen in het vierde jaar primair onderwijs, 2 leerlingen in het zevende jaar primair onderwijs en 12 leerlingen in het tweede jaar secundair onderwijs. De schrijfstukken werden geselecteerd door de leerkracht zelf op grond van centrale richtlijnen. Met betrekking tot de component 'luisteren & spreken' ontwikkelde men telkens twaalf groepsdiscussietaken. Er werden 120 personen aangesteld om de groepsdiscussietaken af te nemen en terzelfdertijd te beoordelen. Zij volgen vooraf een uitgebreid, geaccrediteerd opleidingsprogramma. Het beoordelen van de stukken gebeurde centraal, door getrainde beoordelaars. De groepsdiscussietaak (15 min), die op gang getrokken werd rond een bepaald topic, werd afgenomen en terzelfdertijd beoordeeld door externe, getrainde beoordelaars. Elk schrijfstuk werd beoordeeld door drie onafhankelijke beoordelaars (onderwijsprofessionals, doorgaans leerkrachten in functie) aan de hand van de criteria uit de scoringstool. Het in 'real time' scoren en beoordelen van de prestaties van de leerlingen op de groepsdiscussietaken gebeurde door externe assessoren teneinde potentiële bias vanwege de eigen leerkracht, die de leerlingen kent, te voorkomen. Dit gebeurde ook aan de hand van een scoringstool. De prestatiecategorieën ('not yet working within the level', 'working within the level', 'performing well at the level', 'performing very well at the level', 'performing beyond the level') waren vastgelegd vanuit het CfE en werden vervolgens toegepast op SSLN. Deze prestatiecategorieën fungeerden ook als rapporteringscategorieën: voor elk van de componenten van geletterdheid rapporteerde men resultaten onder de vorm van het aandeel leerlingen in de verschillende leerken in elk van de vijf prestatiecategorieën.

4.6 Verenigde Staten: National Assessment of Educational Progress (NAEP)

De National Assessment of Educational Progress (NAEP) heeft als doel om informatie te leveren over de prestaties van groepen van leerlingen (bv. alle leerlingen uit het vierde leerjaar) en subgroepen (bv. meisjes) in verschillende disciplines. Er worden geen resultaten opgeleverd op het niveau van individuele leerlingen of scholen. De resultaten zijn gebaseerd op representatieve steekproeven van leerlingen uit het vierde leerjaar (9-10 jaar), het achtste leerjaar (13-14 jaar) en het twaalfde leerjaar (17-18 jaar). De betrokken inhoudsdomeinen zijn kunst, maatschappijleer, vreemde talen, wiskunde, lezen, wetenschappen, technologie en technische geletterdheid, geschiedenis van de Verenigde Staten, wereldgeschiedenis en schrijven. Twee belangrijke spelers betrokken bij de uitwerking van het NAEP zijn de 'National Assessment Governing Board' (NAGB) en het 'National Center for Education Statistics' (NCES). Het NAGB is verantwoordelijk voor de domeinbeschrijving, het toetsraamwerk en het vastleggen van de prestatiestandaarden. Het NCES draagt de eindverantwoordelijkheid over de ontwikkeling van de toetsen. De eigenlijke ontwikkeling is vaak in handen van toetsontwikkelaars die in onderaanneming werken. Voor de studie die aanleiding gaf tot deze publicatie, bestudeerden we NAEP-Science (2009) en NAEP-Technology & Engineering Literacy (2013).

NAEP-Science van 2009 peilde naar de kennis en vaardigheden van leerlingen uit het vierde, achtste en twaalfde jaar met betrekking tot natuurwetenschappen, levenswetenschappen en aarde- en ruimtewetenschappen. De hoofdstudie bestond uit een klassieke toets met meerkeuzevragen en werd afgenomen bij een steekproef van 156.500 vierdejaars, 151.000 achtstejaars en 11.100 twaalfdejaars. Daarnaast waren er twee aparte

toetsen, 'Interactive Computer Tasks' (ICTs) and 'Hands-On Tasks' (HOTs), die proefgedraaid werden, elk bij een aparte nationaal representatieve steekproef van 2000 leerlingen voor elk van de drie leerjaren. Dit laatste deel van de toets had als doel de vaardigheden van de leerlingen op het vlak van 'scientific inquiry' (wetenschappelijk onderzoek) na te gaan. De 'hands-on' taken duurden veertig minuten en gaven leerlingen de mogelijkheid aan te tonen in hoeverre ze in staat zijn om wetenschappelijk onderzoek te plannen en uit te voeren, te redeneren doorheen complexe problemen en hun wetenschappelijke kennis toe te passen in reële contexten. Hiervoor konden ze gebruik maken van verschillende soorten materiaal en laboratoriumuitrusting om echte experimenten uit te voeren. De interactieve taken duurden twintig tot veertig minuten en vereisten dat leerlingen een wetenschappelijk probleem oplosten in een gesimuleerde (natuurlijke of laboratorium-)omgeving. NAEP maakte voor alle toetsen gebruik van uitvoerig getrainde toetsassistenten. Bovendien werd voor alle toetsafnames ook het 'eigen' materiaal meegenomen, zoals laptops, kits met het materiaal voor de hands-on taken, om op die manier standaardisering van afname doorheen het land te verzekeren. Er werd gezorgd voor speciale regelingen qua afname voor leerlingen met een functiebeperking. Voor toetsen die via de computer werden afgenomen, startte de toets met een 'tutorial' die ervoor moet zorgen dat leerlingen succesvol doorheen de taken kunnen navigeren. Voor elk item en elke taak wordt tegelijkertijd en door dezelfde mensen een rubric ontwikkeld. NAEP maakte voor alle toetsen gebruik van een uitgebreid protocol voor het scoren van de toetsen, waarbij naast het uitwerken en testen van de scoringstools ook procedures werden uitgewerkt die consistent, valide en objectief scoren in de hand werken. Items en taken werden gescoord door getrainde beoordelaars. Beoordelaars werden ingedeeld in teams van acht tot twaalf leden en werkten in dat team voor de gehele duur van het scoringsproces, dat ongeveer een maand duurde, en gebald voor één inhoudsdomrein op één locatie gebeurde. Alle scores voor een welbepaalde taak of item gebeurden binnen één team, om een consistente toepassing van de rubric te verzekeren. Daarbij werd er binnen het team taak per taak gescoord. Aan het hoofd van het team stond een trainer, die de training en eventuele hertraining op zich nam. Daarnaast was er een scoresupervisor, die de data monitorde terwijl er werd gescoord. De vereiste was dat in het geval van drie- of vierpuntsschalen een beoordelaarsovereenstemming van 75% werd gehaald. Afhankelijk van de omvang van de steekproef was een aandeel van de taken dubbel gescoord, tussen 5 en 25%. De resultaten voor de ICTs en de HOTs werden niet geschaald. De data werden geanalyseerd om een samenvatting te bieden van de prestaties van leerlingen. Percent-correctstatistieken werden per taak en over de verschillende taken heen berekend en gerapporteerd (zie [Figuur 4.4](#) en [Figuur 4.5](#)). Dit stond in contrast met de resultaten van de hoofdstudie, die wel werd geschaald en waar wel een niveaubepaling op werd toegepast in de vorm van 'basic'-, 'proficient'- of 'advanced'-prestatieniveaus.

Average percent correct score for all hands-on tasks in NAEP science, by selected student characteristics and grade: 2009									
	Gender		Race/ethnicity				Eligibility for NSLP ¹		
	All students	Male	Female	White	Black	Hispanic	Asian/ Pacific Islander	Eligible	Not eligible
Grade 4	47	45	49	51	37	42	53	41	52
Grade 8	44	43	45	48	35	37	45	38	48
Grade 12	40	39	41	45	29	35	43	—	—

— Not available.

¹ National School Lunch Program.

NOTE: Black includes African American, Hispanic includes Latino, and Pacific Islander includes Native Hawaiian. Race categories exclude Hispanic origin.

Figuur 4.4: Percent-correct-scores voor HOTs (bron: NCES, 2012, p. 5).

Percentage of twelfth-grade students who successfully completed each step of the Maintaining Water Systems hands-on task in NAEP science, by selected student characteristics: 2009

		Gender		Race/ethnicity			
		Male	Female	White	Black	Hispanic	Asian/ Pacific Islander
Predict	Step 1	63	64	68	39	58	76
	Step 2	74	76	84	50	59	66
Observe	Step 3	9	13	13	3	6	17
Explain	Step 4	17	11	17	1	6	13
	Step 5	30	27	31	13	19	39

NOTE: Black includes African American, Hispanic includes Latino, and Pacific Islander includes Native Hawaiian. Race categories exclude Hispanic origin.

Figuur 4.5: Percentage leerlingen dat elk van de stappen van een taak correct uitvoert (bron: NCES, 2012, p. 13).

NAEP-Technology & Engineering Literacy (2013) werd in 2013 voor de eerste keer afgenomen bij leerlingen uit het achtste leerjaar. Het uitwerken van de domeinbeschrijving vormde een uitdaging in de zin dat er geen curricula rond 'technology and engineering' in

de VS aanwezig waren, waarvan kon worden vertrokken. Omdat TEL crosscurriculair was, vertrok het van verschillende inhoudsdomeinen uit het curriculum. Tegen die achtergrond lag een brede bevraging aan de basis van de domeinbeschrijving. De toets was volledig ICT-gebaseerd; er werd vertrokken van realistische scenario's om leerlingen verschillende taken te laten uitvoeren. De toets duurde vijftig minuten. Hij bestond uit korte scenario's, lange scenario's en 'discrete items', die via een matrix-sampling design in verschillende samenstellingen werden aangeboden aan de leerlingen. De langere scenario's waren complexer en bevatten meer ingebedde taken. De toetsmatrijs achter deze toetsen is weergegeven in [Figuur 4.6](#).

	Technology and Society	Design and Systems	Information and Communication Technology
Understanding Technological Principles	Analyze advantages and disadvantages of an existing technology Explain costs and benefits Compare effects of two technologies on individuals Propose solutions and alternatives Predict consequences of a technology Select among alternatives	Describe features of a system or process Identify examples of a system or process Explain the properties of different materials that determine which is suitable to use for a given application or product Analyze a need Classify the elements of a system	Describe features and functions of ICT tools Explain how parts of a whole interact Analyze and compare relevant features Critique a process or outcome Evaluate examples of effective resolution of opposing points of view Justify tool choice for a given purpose
Developing Solutions and Achieving Goals	Select appropriate technology to solve a societal problem Develop a plan to investigate an issue Gather and Organize data and information Analyze and Compare advantages and disadvantages of a proposed solution Investigate environmental and economic impacts of a proposed solution Evaluate trade-offs and impacts of a proposed solution	Design and Build a product using appropriate processes and materials Develop forecasting techniques Construct and Test a model or prototype Produce an alternative design or product Evaluate trade-offs Determine how to meet a need by choosing resources required to meet or satisfy that need Plan for durability Troubleshoot malfunctions	Select and Use appropriate tools to achieve a goal Search media and digital resources Evaluate credibility and solutions Propose and Implement strategies Predict outcomes of a proposed approach Plan research and presentations Organize data and information Transform from one representational form to another Conduct experiments using digital tools and simulations
Communicating and Collaborating	Present innovative, sustainable solutions Represent alternative analyses and solutions Display positive and negative consequences using data and media Compose a multimedia presentation Produce an accurate timeline of a technological development Delegate team assignments Exchange data and information with virtual peers and experts	Display design ideas using models and blueprints Use a variety of media and formats to communicate data, information, and ideas Exhibit design of a prototype Represent data in graphs, tables, and models Organize, Monitor, and Evaluate the effectiveness of design teams Request input from virtual experts and peers Provide and Integrate feedback	Plan delegation of tasks among team members Provide and Integrate feedback from virtual peers and experts to make changes in a presentation Critique presentations Express historical issues in a multimedia presentation Argue from an opposing point of view Explain to a specified audience how something works Address multiple audiences Synthesize data and points of view

Figuur 4.6: Classificatie van soorten 'assessment targets', ingedeeld volgens domeinen en praktijken van TEL (bron: National Assessment Governing Board, 2014, p. 3-4).

De steekproef bestond uit 20.000 leerlingen (steekproef representatief op het nationale niveau). Voor het afnemen en scoren van deze toets werden dezelfde richtlijnen gevolgd, die ook voor NAEP-Science golden (zie 4.6.2.). De resultaten van de toets werden op één TEL-schaal geplaatst. Daarnaast werden er ook schalen ontworpen voor de subdimensies die werden onderscheiden in het toetsraamwerk. Er werden ook prestatiecriteria gezet, al gebeurde dit alleen voor de geïntegreerde TEL-schaal. Deze standaardzetting was een uitdagende oefening, enerzijds door de mix van meerkeuzevragen en 'constructed-response'-vragen, en anderzijds door de vele interactieve scenariotaken.

5 Evaluatiematrix

Het theoretisch kader, de literatuurstudie en de praktijkvoorbeelden gaven invulling aan de ontwikkeling van de evaluatiematrix. Op basis van deze matrix kunnen toetsprogramma's die competenties toetsen op grond van 'performance assessment'-technieken, op kwalitatieve wijze opgezet worden en/of op hun kwaliteit getoetst worden. In dit hoofdstuk stellen we de verschillende bouwstenen van de matrix voor. We bespreken de voorwaarden waar grootschalige competentietoetsen aan dienen te voldoen om kwaliteitsvol te zijn. We geven hierbij ook enkele aanwijzingen voor verdere lectuur. Om de voorwaarden te illustreren en bouwstenen te concreetiseren, verwijzen we bovendien naar de praktijkvoorbeelden die we in hoofdstuk 4 voorstelden.

5.1 Structuur van de matrix

De evaluatiematrix (zie [Figuur 5.1](#), [Figuur 5.2](#) en [Figuur 5.3](#)) omvat zeven bouwstenen, met daaraan telkens één of meerdere voorwaarden gekoppeld. De matrix geeft globaal en per bouwsteen aan, aan welke voorwaarden of kwaliteitseisen een toets 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 over het competentiepeil van groepen (leerlingen) op systeemniveau.

We volgen de argumentatieve benadering van validiteit (zie hoofdstuk 3) 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 gemaakt moet worden tussen wat de meest kwaliteitsvolle oplossing is, respectievelijk in termen van betrouwbaarheid (generaliseerbaarheid), validiteit (extrapoleerbaarheid) en 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 ook duidelijk een toetsdesign-insteek; de matrix volgt de logische stappen van het op- en uitzetten van toetsen (al of niet in het kader van een ruimer toetsprogramma). De oranje pijl aan de rechterzijde van de matrix in bijlage geeft deze standaardafwikkeling weer. Het startpunt vormt de toets 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.

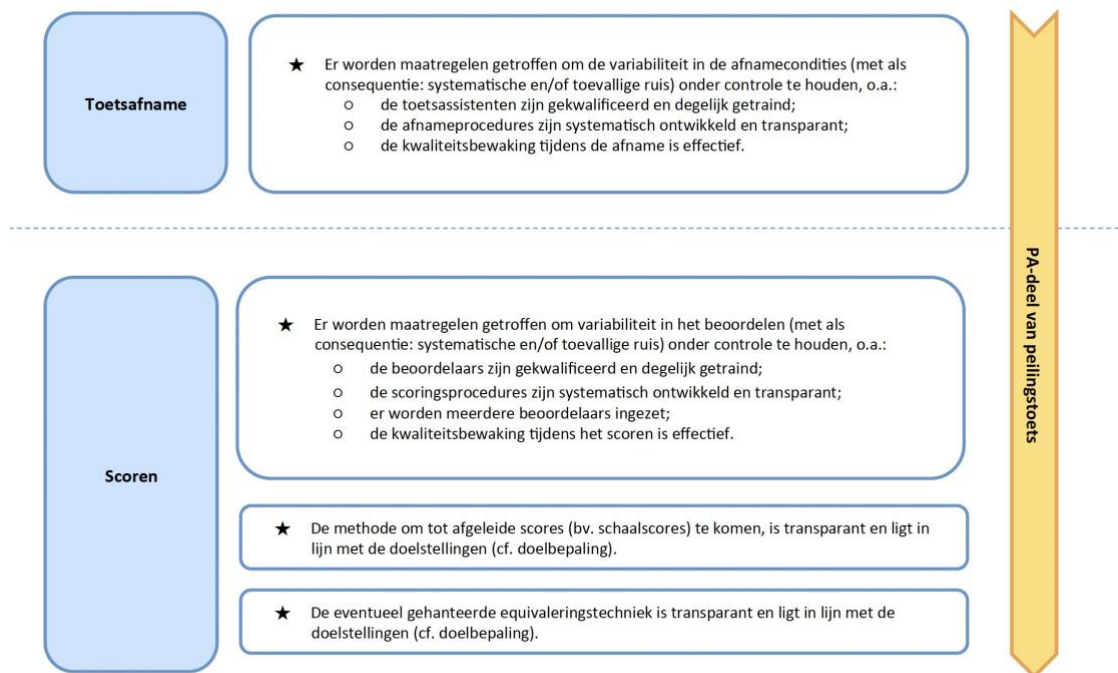
De zeven bouwstenen zijn als volgt benoemd:

- doelbepaling (1 voorwaarde)
- domeinbeschrijving (2 voorwaarden)
- opzet en ontwikkeling (7 voorwaarden)
- toetsafname (1 voorwaarde)
- scoren (3 voorwaarden)

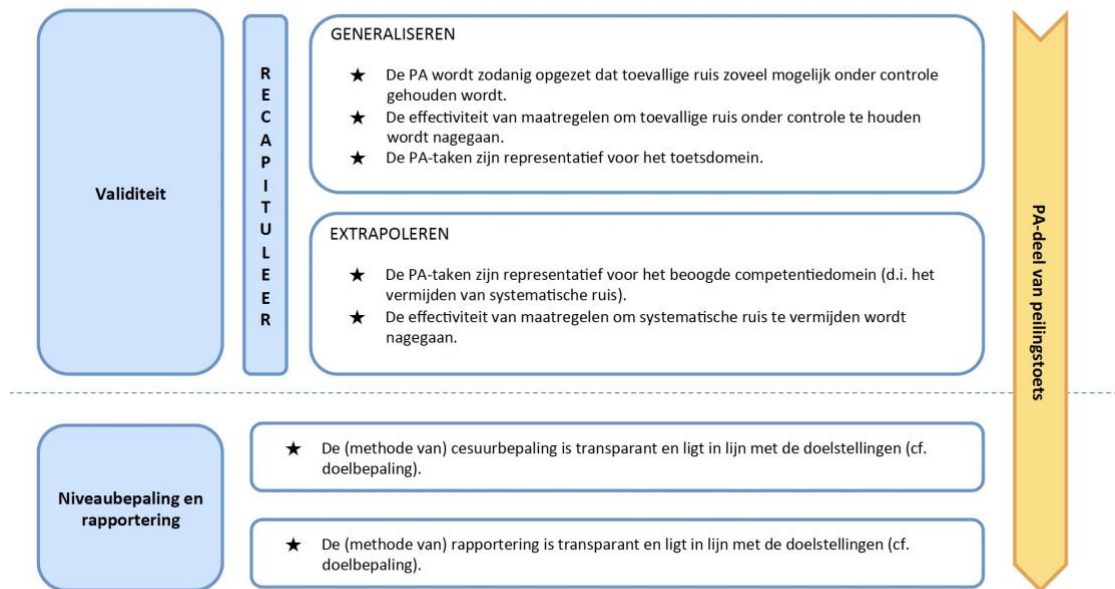
- validiteit (recapitulatie van 5 voorwaarden)
- niveaubepaling en rapportering (2 voorwaarden)

BOUWSTENEN	VOORWAARDEN	
Doelbepaling	<ul style="list-style-type: none"> ★ De bedoeling van de peilingstoets en het peilingsonderzoek wordt duidelijk geëxpliciteerd (o.a. waarom, wat, wie, soort conclusies). 	Peilingstoets als geheel
Domeinbeschrijving	<ul style="list-style-type: none"> ★ De competentie die men wil meten (het beoogde competentiedomein) is gepreciseerd. ★ Vanuit het beoogde competentiedomein wordt het toetsdomein zodanig afgebakend, dat het een duidelijk vertrekpunt vormt voor de ontwikkeling van de peilingstoets en de toetsaken. 	
Opzet en ontwikkeling	<ul style="list-style-type: none"> ★ Alle dimensies van de beoogde competentie, gepreciseerd in het toetsdomein (cf. domeinbeschrijving), zijn weerspiegeld in de peilingstoets (in de vorm van de toetsitems en/of taken en de scoringstools). 	Focus op PA-deel
	<ul style="list-style-type: none"> ★ Voor elke (combinatie van) dimensie(s) selecteert men de meest geschikte toetsvorm. 	
	<ul style="list-style-type: none"> ★ Er zijn voldoende taken om taakvariabiliteit tegen te gaan. 	PA-deel van peilingstoets
	<ul style="list-style-type: none"> ★ De steekproefopzet van leerlingen ligt in lijn met de doelstellingen (cf. doelbepaling). 	
	<ul style="list-style-type: none"> ★ De taken doen op vlak van authenticiteit en complexiteit recht aan de beoogde competentie (cf. domeinbeschrijving). 	
	<ul style="list-style-type: none"> ★ Bij de constructie van taken waakt men erover dat systematische ruis vermeden wordt. 	

Figuur 5.1: De evaluatiematrix (Deel 1)



Figuur 5.2: De evaluatiematrix (Deel 2)




Figuur 5.3: De evaluatiematrix (Deel 3)

De bouwsteen ‘doelbepaling’ staat voor het expliciteren van onder meer het waarom, het wat, het wie en het type conclusies en resultaten van het toetsstelsel (of –programma) en de toets. 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 beoordelingsproces. In deze bouwsteen gaat het voorts ook om het bepalen van afgeleide scores en eventueel ook het equivaleren van de scores. Dit houdt ook in dat gekeken wordt naar de 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. Hierbij worden ook gekeken naar 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 vastleggen van de prestatiecriteria waartegen toetsscores worden afgezet en het rapporteren van de toetsscores zelf. Even belangrijk is dat elk van deze bouwstenen geflankeerd wordt door de 'haalbaarheidsvoorwaarde': kwaliteitsvolle oplossingen kunnen enkel worden geïmplementeerd indien ze ook haalbaar zijn in termen van tijd en middelen.

5.2 Matrix: bouwstenen en voorwaarden

Binnen elke bouwsteen worden in onderstaande tekst een aantal voorwaarden geformuleerd waaraan moet worden voldaan om grootschalige competentietoetsen op basis van 'performance assessment' kwaliteitsvol uit te bouwen. Tekstvakken in de alinea verwijzen naar verdere lectuur in verband met bepaalde voorwaarden. Daarnaast zijn er ook kaders die bedoeld zijn als illustratie van bepaalde voorwaarden, uitdagingen en/of alternatieve oplossingen, aan de hand van buitenlandse praktijkvoorbeelden. Omwille van de vertrouwelijke aard van sommige van de gegevens die we via interviews verkregen, blijven we bij sommige van deze praktijkillustraties op de vlakte over de concrete casus waarover het gaat. We namen in onderstaande tekst de volgende symbolen op om de leesbaarheid te verbeteren:

 Verwijzing naar literatuur

Telkens we expliciet verwijzen naar de literatuur om keuzes te onderbouwen, schrijven we dit weg in dit type van tekstvak.

 Verwijzing naar een illustratie

Inzichten en illustraties uit de praktijkvoorbeelden worden weergegeven in dit type tekstvak.

5.2.1 Bouwsteen 'Doelbepaling'

De eerste bouwsteen focust op het expliciet maken waarom men een toets wil opzetten, welke competentie men - bij wie - beoogt te meten en welke soort conclusies men wil trekken in de fase van de rapportering.

Voorwaarde 1: De bedoeling van het toetsprogramma en/of de toets wordt duidelijk geëxpliciteerd (o.a. waarom, wat, wie, beoogde conclusies).

Een kwaliteitsvolle toets ontwikkelen begint met een gestructureerde doelbepaling. Verschillende deelcomponenten komen hier aan bod: 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 met betrekking tot elk van deze deelcomponenten beïnvloeden elkaar wederzijds. Bovendien hebben ze ook gevolgen voor wat betreft de verdere ontwikkeling van de toets. Een kwaliteitsvolle doelbepaling is met andere woorden cruciaal, omdat ze sturend is voor elk van de bouwstenen die erop volgen. Beslissingen met betrekking tot elk van de volgende bouwstenen moeten dus in lijn liggen met de doelstellingen die in de doelbepaling geëxpliciteerd werden.

We onderscheiden vier centrale antwoorden die cruciaal zijn opdat een doelbepaling een adequaat vertrekpunt vormt voor de rest van de toets, met name een antwoord op:

- de waarom-vraag
- de wat-vraag
- de wie-vraag
- de vraag naar het type conclusies/de vorm van de resultaten

De waarom-vraag staat centraal in de doelbepaling ('waarom willen we toetsen?') en schetst welke de vragen zijn waarop men door middel van de toets een antwoord wil krijgen. Wat betreft de toetsen die wij analyseerden, is het doel of de functie evident: men beoogt kwaliteitsbewaking, en dit op het niveau van het onderwijssysteem. Andere denkbare richtvragen zijn bijvoorbeeld of de toets formatieve of summatieve doeleinden heeft; of de toets opgezet wordt vanuit een typisch ontwikkelings- of een verantwoordingsperspectief; of men op basis van de toets individuele leerlingen beoogt te selecteren, dan wel diploma's of getuigschriften dient uit te reiken.


Toetsprogramma's en toetsen kunnen overigens multiple doelen dienen. 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. Bij toetsen die een hybride doelbepaling hebben, is wel waakzaamheid geboden, net omdat aan deze uiteenlopende doelstellingen andere kwaliteitsvereisten voor het opstellen van de toets gekoppeld kunnen zijn, dan wel sterker of minder sterk kunnen doorwegen.

De wat-vraag ('wat willen we toetsen?') heeft als doel om in algemene termen te duiden welk(e) construct, competentie, inhoudsdomein, leerdoel men in kaart wil brengen. De eigenlijke precisering en concretisering van de competentie die men wil meten hoort thuis in de fase van de domeinbeschrijving (zie bouwsteen 'Domeinbeschrijving').

Gekoppeld aan de wat-vraag stelt zich meteen ook de wie-vraag ('bij wie willen we dit toetsen?'): wie maakt deel uit van de doelpopulatie? Zijn dit bv. leerlingen basisonderwijs of secundair onderwijs? Of specifieker: leerlingen aan de start, in het midden of aan het einde van hun schoolloopbaan in het basisonderwijs of secundair onderwijs?

Bovenstaande vragen zijn op hun beurt gelinkt aan de vraag welk type conclusies men wil trekken uit de resultaten en scores en in welke vorm er gerapporteerd dient te worden. In

dit verband duikt bijvoorbeeld de vraag op of alle gemeten dimensies van het construct/de competentie op één schaal moeten worden gezet, dan wel of er, voor een aantal subdimensies, via aparte schalen zal worden gerapporteerd. Waar ook over nagedacht dient te worden, is de eventuele wens om resultaten te vergelijken. Mogelijk moeten de scores van een welbepaald afnamejaar worden vergeleken met een ander afnamejaar, of dienen prestaties van leerlingen van een bepaald leerjaar gelinkt te worden aan die van een ander leerjaar. In vele toetssystemen of - programma's is er bovendien de vereiste dat prestaties van deelgroepen met elkaar worden vergeleken (bv. op basis van gender, sociaal-economische status, deelstaat, enz.). Ook kan het een expliciete vereiste zijn dat de toetsscores na afloop teruggekoppeld worden (op school- of zelfs op leerlingenniveau) aan de scholen die hebben deelgenomen. In verband met het type conclusies dat men wil trekken en de vorm van de resultaten, dient men zich ten slotte ook te buigen over de vraag hoe scores 'betekenisvol' kunnen worden gemaakt voor het lezerspubliek. In dit verband duiken keuzes op zoals bv. een focus op criteriumgerefeerd beoordelen, wat doorgaans het gebruik van prestatiestandaarden impliceert, of andere manieren om scores betekenisvol te maken, zoals via voorbeeldtaken die representatief zijn voor verschillende punten of locaties op de gehanteerde schaal (zie ook bouwsteen 'Niveaubepaling en rapportering').


 Zie National Research Council (2014, 134) voor een illustratie van welke vragen typisch gesteld worden in het kader van grootschalige toetsen gericht op kwaliteitsbewaking.

5.2.2 Domeinbeschrijving

In deze bouwsteen wordt het beoogde competentiedomein nader uitgewerkt. Vanuit het competentiedomein worden vervolgens de te toetsen dimensies (het 'toetsdomein') afgebakend.

Voorwaarde 2: De competentie die men wil meten (het beoogde competentiedomein) is gepreciseerd.

Zeker bij het in kaart brengen van competenties geldt dat het beoogde competentiedomein vaak breed en moeilijk scherp af te lijnen of precies te specificeren is (bv. 'geletterdheid', 'creativiteit', 'zelfsturing'). Toch dient men in de domeinbeschrijving duidelijk te beschrijven wat er precies met het beoogde construct of de beoogde competentie wordt bedoeld en welke verschillende dimensies er deel van uitmaken. Hiervoor kan men een beroep doen op inhoudsexperten en gebruik maken van wetenschappelijke inzichten.


 *In het kader van Periodieke Peiling van het Onderwijsniveau (PPON) – Schrijfvaardigheid (zie 4.3.) vormden de wettelijke vereisten (Kerndoelen enerzijds en Tussendoelen en Leerlijnen anderzijds) het vertrekpunt voor het conceptualiseren van het begrip 'schrijfvaardigheid'. De kerndoelen bieden vaak echter onvoldoende handvaten voor het construeren van peilingsonderzoek. Tegen die achtergrond maakt Cito voor elke peiling een domeinbeschrijving, die als basis dient voor het peilingsontwerp en de bijhorende instrumentontwikkeling. Een dergelijke domeinbeschrijving legt bijvoorbeeld de didactisch*

betekenisvolle eenheden vast en beschrijft die vervolgens ook. In de publicatie Balans van de schrijfvaardigheid in het basis- en speciaal basisonderwijs is voor PPON-‘Schrijfvaardigheid’ (2009) zo’n domeinbeschrijving opgenomen. De domeinbeschrijving is onder meer gebaseerd op wetenschappelijke inzichten en analyse van veel gebruikte methoden en bijgesteld op basis van het commentaar van vakinhoudelijke deskundigen, leerkrachten en geïnteresseerde leken.


Voorwaarde 3: Vanuit het beoogde competentiedomein wordt het toetsdomein zodanig afgebakend, dat het een duidelijk vertrekpunt vormt voor de ontwikkeling van de toets en de toetstaken.

Bij het afbakenen van het toetsdomein is het belangrijk erover te waken dat het toetsdomein het beoogde competentiedomein weerspiegelt, onder meer in termen van inhoud en cognitieve complexiteit. Men dient op transparante wijze duidelijk te maken welke dimensies wel en welke dimensies niet worden meegenomen in het toetsdomein en op basis van welke argumenten dit gebeurt.

Vanuit het beoogde competentiedomein bakent men dus het toetsdomein af. De systematische omschrijving van dit toetsdomein gebeurt in het toetsraamwerk, ook wel toetsspecificaties of blauwdruk genoemd. Deze systematische omschrijving is noodzakelijk om ervoor te zorgen dat de taken die ontwikkeld zullen worden, het beoogde construct of de beoogde competentie goed vertegenwoordigen.

 *Een systematische afbakening van het toetsdomein vormt een noodzakelijk vertrekpunt voor de toetsconstructie. Wanneer het beoogde competentiedomein daarbij ‘vernauwt’, heeft dit vanzelfsprekend gevolgen voor de (validiteit van de) interpretatie van de scores. Dat dit een evenwichtsoefening blijft die niet te ver kan worden doorgedreven, wordt geïllustreerd door het voorbeeld van National Assessment Program – Literacy and Numeracy (NAPLAN) - ‘Persuasive Writing’ (zie 4.1.). Om schrijfvaardigheid te toetsen, werd een toets opgezet rond één genre (‘overtuigend schrijven’); het brede domein ‘schrijfvaardigheid’ werd in het toetsraamwerk van NAPLAN met andere woorden ingeperkt. Dit had als gevolg dat de prestaties van de leerlingen op de taak minder representatief waren voor het beoogde competentiedomein, nl. schrijfvaardigheid in het algemeen.*

Het is cruciaal dat de domeinbeschrijving en het toetsraamwerk gedetailleerd en transparant zijn, zeker bij toetsen die competenties beogen te meten via ‘performance assessment’. Het toetsraamwerk omvat steeds de inhoud en de aard van de cognitieve processen die beoordeeld moeten worden. Daarnaast vinden we in het toetsraamwerk ook de psychometrische eigenschappen van de taken en relevante informatie voor de afname (bv. instructies, voorziene tijd, al of niet keuzemogelijkheid wat de opdracht betreft, gebruik van materialen, ...) terug.

 Zie Suzanne Lane and Stone (2006) voor meer informatie over de inhoud van toetsraamwerken.

Documenten die als uitgangspunt gebruikt worden voor de domeinbeschrijving en het toetsraamwerk (bv. nationale curricula en standaarden) bieden vaak onvoldoende houvast. Zulke documenten limiteren of prioriteren niet, wat bij een toetsraamwerk wel het geval hoort te zijn. Daarom dienen andere bronnen te worden geconsulteerd, zoals bijvoorbeeld wetenschappelijke inzichten met betrekking tot de beoogde competentie of toetsraamwerken uit het buitenland. Het is ook interessant experts te betrekken in dit proces; zij kunnen uit deze documenten de meest relevante elementen selecteren, wat bijdraagt tot een degelijke domeinbeschrijving en een uitgewerkt toetsraamwerk.

💡 *Het domein 'Technology and Engineering Literacy' (TEL) van National Assessment of Educational Progress (NAEP) is crosscurriculair (zie 4.6.3.). Bovendien is het een nieuw te toetsen domein, wat betekent dat NAEP zich niet kon baseren op bestaande toetsraamwerken. Ze maakten gebruik van heel uiteenlopende documenten, zoals bestaande eindtermen inzake technologie uit de deelstaten, invloedrijke technologiestandaarden uit andere landen, het NAEP-toetsraamwerk voor Wetenschappen, documenten van belangenorganisaties en onderzoeksrapporten. Met het oog op de uitwerking van de domeinbeschrijving en het toetsraamwerk, betrokken zij experts vanuit heel diverse inhoudsdomeinen en sectoren: (1) vertegenwoordigers uit scholen, het bedrijfsleven en de ingenieurswetenschappen; (2) internet-, toets- en onderwijsexperts, en experts op vlak van wetenschapsonderwijs en '21st century skills'.*

5.2.3 Opzet en ontwikkeling

Na de bepaling van het toetsdomein staat in de volgende bouwsteen de ontwikkeling van de toets centraal. Cruciale elementen zijn onder andere de taakconstructie (met aandacht voor authenticiteit), de uitwerking van de scoringstools, de toetscompilatie en de opzet van de steekproef (van zowel leerlingen als taken). Deze bouwsteen omvat in totaal zeven voorwaarden.

Voorwaarde 4: Alle dimensies van de beoogde competentie, gepreciseerd in het toetsdomein, zijn weerspiegeld in de toets (in de vorm van toetsitems en/of taken en scoringstools).

De domeinbeschrijving mondt uit in de eigenlijke opzet en ontwikkeling van de toets, waarin alle dimensies van de beoogde competentie, afgebakend in het toetsdomein, weerspiegeld dienen te zijn. Indien dit niet het geval is, spreken we van onderrepresentatie van het construct. Dit impliceert dat de toets belangrijke aspecten (inhoud en/of processen) van het beoogde construct of de beoogde competentie niet vat. Het gevolg is dat de betekenis die aan de toetsscores gehecht kan worden, verengd wordt. Het tegengaan van deze ondervertegenwoordiging van het construct begint eigenlijk al in de fase van de domeinbeschrijving. Zeker bij competentietoetsen, die vaak een breed en complex domein dienen te bestrijken, is deze ondervertegenwoordiging een uitdaging.

💡 Voor de ontwikkeling van de toets, de taken en de scoringstools hanteert men in NAEP een 'evidence centered design'. Een dergelijk design impliceert dat erg uitgebreid en in verschillende fasen getest wordt of met de taken wel gemeten wordt wat men beoogt te meten. Hierbij heeft men zowel oog voor onderrepresentatie van het construct als voor construct-irrelevante variantie (zie voorwaarde 10). Educational Testing Service (ETS) ontwikkelde eind jaren negentig het 'evidence-centered design' (ECD)-raamwerk. Van dit raamwerk kan men gebruik maken bij het opstellen van coherente toetsen en het in praktijk brengen van deze bewijsketen. Het ECD-raamwerk moet ervoor zorgen dat de manier waarop de toets wordt opgezet in alle fasen in lijn is met wat men beoogt te meten (zie 'doelbepaling'). De gemeenschappelijke designarchitectuur zorgt er bovendien voor dat de manier waarop informatie verzameld en geïnterpreteerd wordt, gecoördineerd verloopt voor de verschillende partijen die bij het proces betrokken zijn (zoals taakontwikkelaars, statistici, interface-designers, enz.). Centraal in het ECD-raamwerk staan het 'Conceptual Assessment Framework' (CAF), dat zich vooral richt op de processen die nodig zijn om de toets systematisch te ontwikkelen enerzijds, en de 'four-process delivery architecture', dat zich in de eerste plaats focust op de oplevering ('delivery') van de toets anderzijds.

Grootschalige toetsen die 'performance assessment' inzetten, omvatten vaak minder taken dan klassieke toetsen items omvatten (zoals bijvoorbeeld klassieke, schriftelijke toetsen op basis van meerkeuzevragen). Daarom is het belangrijk dat voor elke taak wordt nagedacht welke dimensie(s) van het construct ze vertegenwoordigt. Wat bijdraagt aan een volledige(re) dekking van het beoogde construct, is het systematisch ontwikkelen van toetsen, taken en scoringstools. Verschillende componenten van het ontwikkelproces dienen hierbij in beschouwing genomen te worden, zoals:


- het design van het constructieproces;
- de selectie en training van de ontwikkelaars;
- de toetsconstructie zelf;
- de review door experts en de daaruit voortvloeiende noodzakelijke aanpassingen; en
- het piloot-testen van de taken en scoringstools.

📖 Suzanne Lane and Stone (2006) en Schmeiser and Welch (2006) geven een overzicht van wat een systematische ontwikkeling inhoudt.

Parallel aan de taken ontwikkelt men de scoringstools (beoordelingscriteria en -schaal). Deze tools ondersteunen een consistente toepassing van de scoringscriteria. Net als de ontwikkeling van de taken en opdrachten, is de ontwikkeling van scoringstools een iteratief proces.

Wat de beoordelingscriteria en -schaal betreft, moeten er keuzes gemaakt worden. Deze keuzes hebben betrekking op de te hanteren scoringsprocedure (analytisch of holistisch), de beoordelingscriteria en de wijze waarop die geoperationaliseerd worden en ten slotte

ook het aantal punten op de beoordelingsschaal, inclusief de omschrijving van de schaalpunten of bekwaamheidsniveaus.

 Zie Kuhlemeier et al. (2013) voor toelichting inzake drie types schalen: normatieve schalen, descriptieve schalen of rubrics, en productschalen.

Met betrekking tot de scoringsprocedure maken Suzanne Lane and Stone (2006) het onderscheid tussen analytische en holistisch scoringstools duidelijk aan de hand van het voorbeeld van het beoordelen van een schrijfstuk. Bij een holistische beoordeling geven de beoordelaars een enkel, holistisch oordeel over de kwaliteit van het schrijfstuk. Ze kennen één score toe, waarbij ze gebruik maken van een rubric met criteria en meestal ook van papers die een illustratie vormen van het vereiste prestatieniveau voor elke score. Bij analytisch scoren beoordeelt de beoordelaar het schrijfstuk aan de hand van een aantal kenmerken, zoals bijvoorbeeld inhoud, structuur en grammatica. Er wordt een score gegeven om het niveau van elk van deze componenten weer te geven. Soms worden de verschillende criteria gewogen, waarbij criteria waarvan verondersteld wordt dat ze belangrijker zijn voor het construct dat gemeten wordt, meer bijdragen in de totaalscore dan andere. Onderzoek geeft geen eenduidig antwoord op de vraag of holistische dan wel analytische scoringstools de voorkeur genieten (Johnson, Penny, and Gordon 2009; S. Lane 2015). Er gaan stemmen op om beide aanpakken te combineren. Zo is er de optie om holistisch te scoren om een totaalscore te bekomen én analytisch te scoren om feedback te kunnen verschaffen. Deze methode brengt wel hogere kosten met zich mee. Johnson, Penny, and Gordon (2009) raden aan om de keuze te laten afhangen van de manier waarop de informatie gebruikt zal worden en de beschikbare middelen. Het advies is om holistisch te scoren bij eenvoudige prestaties, die niet meer dan één belangrijk kenmerk in kaart brengen. Analytisch scoren kan dan eerder gebruikt worden bij meer complexe prestaties, met verschillende belangrijke componenten en in het geval achteraf feedback gegeven dient te worden op eventueel deeldimensies.

Voorwaarde 5: Voor elke (combinatie van) dimensie(s) selecteert men de meest geschikte toetsvorm.

Deze volgende voorwaarde in het kader van de opzet en de ontwikkeling houdt verband met de keuze voor een geschikte toetsvorm. Het inzetten van 'performance assessment' bij het grootschalig toetsen van competenties dient weloverwogen te gebeuren. De evaluatiematrix 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 beargumenteerde keuze gemaakt worden met betrekking tot de te gebruiken toetsvormen.

Hoewel 'performance assessments' door de band genomen beter in staat zijn om complexe cognitieve processen van leerlingen in kaart te brengen en toetsen samengesteld uit meerkeuzevragen vaak beschouwd worden als vooral nuttig voor het evalueren van lagere-orde cognitieve processen, hoeft dit niet steeds het geval te zijn. Meerkeuzevragen kunnen,

indien ze zorgvuldig zijn opgesteld, ook bepaalde complexe cognitieve processen vatten en niet elke 'performance assessment' meet sowieso meer complexe processen (Suzanne Lane and Stone 2006; National Research Council 2014). Daarenboven brengen dit soort toetsen een logistieke en financiële meerkost met zich mee. Veel tijd en (financiële) middelen zijn nodig voor o.a. de noodzakelijke hulpmiddelen voor de toets (bv. computers of labokits), de doorgedreven training van beoordelaars en het inzetten van meerdere beoordelaars. Tegen die achtergrond is het raadzaam een overwogen keuze te maken voor een welbepaalde toetsvorm en 'performance assessment' alleen dan in te zetten wanneer het echt een meerwaarde oplevert. Hierbij dient expliciet te worden gemaakt waarom uiteindelijk voor deze of gene toetsvorm werd geopteerd.

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.

Voorwaarde 6: Er zijn voldoende taken om de implicaties van tussen-taken-variabiliteit te minimaliseren.

Deze cruciale voorwaarde houdt verband met de noodzaak om de variabiliteit, die het gevolg is van het werken met taken, te minimaliseren. De taken in 'performance assessments' vormen immers een belangrijke foutenbron, die de betrouwbaarheid van de toetsscores in negatieve zin kan beïnvloeden. De aard en kwaliteit van de respons van een leerling op een toets kunnen namelijk erg variëren van de ene steekproef van taken naar de andere. Deze 'taakgerelateerde error' heeft te maken met de unieke kenmerken van de taak en de interactie van deze kenmerken met de kennis en ervaring van de leerling (National Research Council 2014). Het heeft als gevolg dat de scores van leerlingen overheen verschillende 'performance assessment'-taken die in principe hetzelfde zouden moeten meten, vaak niet erg consistent zijn. Dit fenomeen duiden we aan met de term 'task sampling variability' of tussen-takenvariabiliteit.

Bij 'performance assessments' is deze taakgerelateerde error typisch groter dan bij toetsen samengesteld uit meerkeuzevragen. Bovendien worden bij dit laatste type toetsen toevallige fouten uitgemiddeld op basis van de vele meerkeuzevragen die leerlingen beantwoorden. Dit is niet het geval bij 'performance assessment', waar het aantal taken dat bij een enkele leerling wordt afgenomen doorgaans (erg) klein is. Een mogelijke oplossing bestaat erin voldoende taken (per leerling) af te nemen en zo over deze verschillende taken heen een uitspraak te doen over hun prestatieniveau.

Grootschalige toetsen met het oog op systeemmonitoring die gebruik maken van 'performance assessment' dienen uit verschillende taken te bestaan, om valide en betrouwbaar te zijn. Het voorzien van voldoende taken is echter praktisch vaak niet haalbaar. Op basis van een analyse van internationale praktijkvoorbeelden en screening van

empirische literatuur, identificeerden we volgende mogelijke aanpakken voor de vastgestelde problematiek:

- Het inperken van het competentiedomein naar het toetsdomein via een kwaliteitsvolle domeinbeschrijving (zie bouwsteen 'Domeinbeschrijving').
- Het gebruik van matrix sampling (zie verder, voorwaarde 8).
- Het inzetten van verschillende types items en taken in één toets. Door verschillende types van items en taken (bijvoorbeeld, open vragen waarbij geantwoord moet worden met een woord, meerkeuzevragen, essay-vragen, ...) te combineren, laat men betrouwbaarheidskwesties die verbonden zijn aan één bepaald vraagtype (bv. de variabiliteit die wordt geïntroduceerd door het beoordelingsproces bij 'performance assessment' of variabiliteit die wordt geïntroduceerd doordat leerlingen het antwoord gokken bij toetsen die bestaan uit meerkeuzevragen) minder doorwegen.


💡 *Bij het 'National Assessment Program' (NAP) – ICT Literacy (zie 4.2.) zijn alle items en taken binnen een module inhoudelijk gelinkt aan een verhaallijn. In een module bevinden zich verschillende formats (bv. meerkeuzevragen, korte antwoordvragen, taken die de inzet van een enge vaardigheid vereisen, taken die een meer complexe prestatie vereisen). Een module eindigt meestal met een grote taak, waarbij aan leerlingen wordt gevraagd om een product te creëren. Daarvoor komen er een aantal voorbereidende items en taken (zgn. 'lead-up items'), die naar de betreffende grote taak toe leiden (ook inhoudelijk). De voorbereidende items en taken kunnen verschillende vormen aannemen: vaardigheidstaken (meestal software simulation items), 'information-based' items en 'evaluation' items, waarvoor meerkeuzevragen, vragen waarbij leerlingen elementen moeten slepen en (elders) neerzetten en korte antwoordvragen worden gebruikt.*

- Het inzetten van toetsen met een grotere inbedding in de eigen (klas)context.


💡 *In Nieuw-Zeeland werkte men met betrekking tot 'National Monitoring Study of Student Achievement' (NMSSA) - Arts (2016) voor het eerst met lokale klasleerkrachten. Hun taak was het reële competentieniveau van een steekproef leerlingen te beoordelen. Als richtlijn gold dat ze zich moesten baseren op de typische prestaties van de leerlingen doorheen het jaar. De beslissing om leerkrachten in te zetten kaderde in de overtuiging dat de eigen klasleerkracht zich in de beste positie bevindt om een oordeel te vellen over het reële prestatieniveau. Het alternatief zou zijn om de toets via een gestandaardiseerde taak af te laten nemen door een centrale toetsassistent. Gezien de verschillende te toetsen subdomeinen (dans, drama, muziek en visuele kunsten) zou dit echter praktisch niet haalbaar zijn. De werkwijze paste bovendien in een bredere strategie in Nieuw-Zeeland om de professionele kennis van leerkrachten op vlak van prestaties en vooruitgang van leerlingen, te optimaliseren en benutten. Grootschalige professionaliseringstrajecten voor leerkrachten werden opgezet met het oog op kwaliteitsvolle (formatieve) evaluaties. Merk op dat de hier beschreven werkwijze met lokale leerkrachten enkel voor het 'performance assessment'-gedeelte van de NMSSA-Arts (2016) werd gehanteerd, en overigens ook voor het eerst in het kader van het nationale toetsprogramma.*

Voorwaarde 7: De steekproefopzet van leerlingen ligt in lijn met de doelstellingen (zie doelbepaling).

Voor de steekproeftrekking van leerlingen in het kader van grootschalige toetsen met het oog op kwaliteitsmonitoring, ligt de focus op technieken voor en consequenties van 'probability sampling' en 'multistage cluster sampling'.

 Voor meer informatie over steekproeftrekking verwijzen we naar Kish (2005) en Mazzeo and Zieky (2006).

Hierbij is het steeds belangrijk voor ogen te houden dat de steekproefopzet in lijn ligt met de doelstellingen die men eerder in de doelbepaling preciseerde.

 *In een bepaalde casus formuleerde de opdrachtgever vrij laat in het proces de verwachting dat deelnemende scholen achteraf een schoolfeedbackrapport zouden krijgen. De projectleiding kon hier echter geen rekening meer mee houden: de steekproeven waren reeds getrokken. Het gevolg was dat de resultaten niet voldoende betrouwbaar waren voor rapportering op schoolniveau. Er zijn dus risico's verbonden aan het gebruik van de rapporten, met name wat betreft (de validiteit van) de interpretatie van de resultaten. De bijkomende doelstelling om te rapporteren op schoolniveau lag immers niet in lijn met de initiële doelstellingen.*

Voorwaarde 8: De eventuele matrix-sampling ligt in lijn met de doelstellingen (zie doelbepaling).

Wanneer we voorwaarden 6 en 7 gecombineerd in beschouwing nemen, komen we uit op de volgende voorwaarde, die betrekking heeft op de techniek van matrix-sampling. Een van de uitdagingen om toetsscores te generaliseren naar het toetsdomein, heeft te maken met het voorzien van voldoende taken. De constructen die via grootschalige competentietoetsen op basis van 'performance assessment' gemeten worden, bestrijken vaak een breed domein. Gecombineerd met de problematiek van de tussen-takenvariabiliteit zorgt dit ervoor dat deze toetsen 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 veel gebruikt wordt, is de techniek van matrix-sampling.

Bij matrix-sampling worden steekproeven van taken en items uit een ruimere pool afgenomen bij steekproeven leerlingen. Verschillende groepen van leerlingen krijgen met andere woorden verschillende combinaties van taken en/of items voorgelegd. Samen zeggen de prestaties van alle leerlingen op alle taken iets over de prestaties van de groep van leerlingen met betrekking tot het construct of de competentie dat/die men beoogt te meten. Om de prestaties van de leerlingen op één schaal te kunnen zetten, en dus

gezaamenlijk over de prestaties te kunnen rapporteren, worden 'link items' of 'link taken' voorzien. Dit zijn taken die door alle, of ten minste door een deel van de leerlingen worden uitgevoerd, en de prestaties dus aan elkaar kunnen linken. Via deze werkwijze zijn er minder taken per leerling nodig om tot betrouwbare scores op groepsniveau te komen. Matrix-sampling is met andere woorden een kostenefficiënte en technisch degelijke methode om alle dimensies van de beoogde competentie te kunnen meten.

💡 *NAPLAN koos ervoor om de brede competentie 'schrijfvaardigheid' te verengen tot één schrijffgenre: overtuigend schrijven (zie illustratie bij tweede voorwaarde m.b.t. bouwsteen 'domeinbeschrijving' en ook 4.1.). Bij CITO ging men voor de PPON-schrijfvaardigheid anders te werk om hetzelfde probleem het hoofd te bieden (zoe ook 4.3.). Daar ontwikkelde men een takenpool die 12 taken omvatte en die de verscheidenheid van het domein vertegenwoordigde door verschillende teksttypes en -genres op te nemen. Vervolgens werd er via matrix-sampling voor gezorgd dat de toets qua tijdsinvestering voor de leerlingen beheersbaar bleef.*

Hoewel matrix sampling effectief en efficiënt is, moet over de opzet grondig nagedacht worden. Deze techniek is immers, zeker voor toetsen die 'performance assessment' inzetten, niet zonder problemen. Het juiste design vinden is niet vanzelfsprekend en elke oplossing brengt steeds een aantal nieuwe uitdagingen met zich mee.

📖 Hambleton (2006) geven inzicht in een aantal van de criteria die in overweging dienen te worden genomen bij het uitwerken van het ontwerp van matrix sampling, om het hoofd te kunnen bieden aan de uitdagingen verbonden aan het inzetten van matrix sampling voor performance assessment.

Voorwaarde 9: De taken doen op vlak van authenticiteit en complexiteit recht aan de beoogde competentie (zie domeinbeschrijving).


Authenticiteit staat centraal bij 'performance assessment' en verwijst naar de mate waarin de toets erin slaagt de kennis en vaardigheden te reflecteren die belangrijk zijn in de alledaagse context van de leerlingen. Er wordt in dit kader ook wel van 'fidelity' gesproken.


Taken die plaatsvinden in authentieke, levensechte contexten zijn betekenisvol en motiverend voor leerlingen. Naarmate taken of opdrachten sterker lijken op die in de criteriumsituatie, kunnen er ook meer valide voorspellingen worden gedaan over het functioneren in de criteriumsituatie.

💡 *NAP-ICTL is zodanig opgesteld dat het de typische dagdagelijkse toepassing van ICT weerspiegelt (zie 4.2.). De leerlingen voeren taken uit op computers en gebruiken software die zowel gesimuleerde als live-applicaties omvat. De toets omvat naast de software-toepassingen zowel meerkeuzevragen als open-antwoordvragen. Deze zijn gegroepeerd in negen modules, elk met een eigen uniek thema, dat de authentieke basis vormt voor het uitvoeren van de taken. Deze scenario-gebaseerde modules volgen elk een lineaire narratieve sequentie.*

Met betrekking tot de criteriumsituatie stelt zich overigens de vraag of men naar het product wenst te kijken, naar het proces dat ertoe leidt of naar beide. Om representatief te zijn voor het beoogde competentiedomein of de criteriumsituatie is het belangrijk dat de omstandigheden van de observatie representatief zijn voor deze in de criteriumsituatie. Bij de opzet van een toets schrijfvaardigheden bijvoorbeeld, betekent dit dat er in principe ook ruimte moet zijn voor voorafgaande studie van de literatuur en planning en revisie achteraf.

Het is echter niet omdat 'performance assessments' het potentieel hebben om authentiek te zijn, dat ze dat in werkelijkheid ook zijn. Zaak is zodanig te werk te gaan, dat de taken en scoringstools die worden ontwikkeld de beoogde competentie effectief weerspiegelen en daar bewijs voor te leveren. Daarom is het belangrijk de cognitieve processen die bij leerlingen in gang gezet worden door de 'performance assessment'-taak expliciet te maken. Verschillende methodieken kunnen hiervoor aangewend worden: protocolanalyse (ook: cognitieve interviews of 'cognitive labs' genoemd), 'analysis of reasons' en 'analysis-of-errors'.

 Zie Darling-Hammond and Adamson (2014) voor toelichting bij deze methodieken die cognitieve processen expliciteren.

 Bij NAEP (zie 4.6.) zien we dat ze, in het kader van hun 'evidence-centered design', gebruik maken van cognitieve interviews of 'cognitive labs' waarbij leerlingen ertoe worden aangezet om luidop te denken bij het oplossen of uitvoeren van een taak ('think aloud') en vervolgens ook nog geïnterviewd worden door de onderzoeker na uitvoering van de taak (bv. 'stimulated recall').

Voorwaarde 10: Bij de constructie van taken waakt men erover dat systematische ruis vermeden wordt.

Naast 'onderrepresentatie van het construct' (zie voorwaarde 4) leidt ook 'construct-irrelevante variantie' tot systematische ruis. Construct-irrelevante variantie (CIV) verwijst naar variantie in een score die resulteert uit iets anders dan het construct dat men beoogde te meten. De beoordeling is met andere woorden 'te breed'; ze omvat informatie die niet relevant is voor de competentie die men beoogt te meten. Illustratief is een toets die wil peilen naar probleemoplossend vermogen, en die uitgevoerd wordt op een computer. Het onvermogen van een leerling om een bepaalde taak in dat verband af te werken, kan het resultaat zijn van het ontbreken van (voldoende) probleemoplossend vermogen, maar kan even goed het gevolg zijn van beperkte ICT-vaardigheden.

Ook buitenlandse praktijkvoorbeelden van grootschalige toetsen die gebruik maken van 'performance assessment' worden geconfronteerd met construct-irrelevante variantie. Wat

in dit verband met name opvalt is dat de reductie van CIV een spanning kan opleveren met de keuze voor authenticiteit.

💡 *In een van de praktijkvoorbeelden wou men net taken ontwikkelen die geen CIV introduceren. De organisatie was van mening dat controversiële thema's, die interessant en uitdagend zijn voor leerlingen, wel thuishoren in de dagdagelijkse klaspraktijk, maar niet in een nationale toets. Ook wou men de potentiële invloed van voorkennis op het competentieniveau uitschakelen. Zo ontwikkelde men bijvoorbeeld geen schrijfoopdrachten rond wetenschappelijk topics omdat de (wetenschappelijke) voorkennis bij leerlingen mogelijks kan verschillen. Critici vroegen zich met betrekking tot deze keuzes af hoe authentiek en engagerend een toetstaak kan zijn voor de leerlingen, indien het gekozen thema niet over iets gaat dat hen aanbelangt. Tegenover dit voorbeeld staan andere voorbeelden die levensechtheid en authenticiteit van de taken laten primeren. Zo klonk bij een van de praktijkvoorbeelden de waarschuwing dat CIV er sowieso altijd is, zeker bij scenario-gebaseerde toetsen. Werken met scenario's en levensechte contexten brengt dan weer verschillende andere uitdagingen met zich mee. Ten eerste kan deze context voor bepaalde leerlingen net boeiend zijn; andere leerlingen haken er op af. Ten tweede zorgen scenario-gebaseerde toetsformats ervoor dat opeenvolgende items of taken van elkaar afhankelijk zijn: een leerling scoort mogelijk slecht op stap 2 in een scenario, omdat zij stap 1 slecht heeft afgewerkt. Ten derde kunnen ook te uitgebreide tekstpassages aan het begin van sommige taken voor sommige leerlingen belemmerend werken, bijvoorbeeld omdat ze de tekst niet begrijpen. Men erkende dus dat CIV bij 'performance assessment' een van de uitdagingen is en blijft. Zaak is de CIV in de pilootfase zo goed mogelijk in kaart te brengen, en de taken vervolgens verder te ontwikkelen, zodat ze engagerend zijn voor de leerlingen, rekening houdend met mogelijke bronnen van CIV.*

5.2.4 Toetsafname

De focus van deze bouwsteen ligt op het minimaliseren van mogelijke variabiliteit in de afnamecondities. Of anders gesteld, op het belang de afname voldoende gestandaardiseerd - en dus vergelijkbaar en eerlijk - te laten verlopen.


Voorwaarde 11: Er worden maatregelen getroffen om de variabiliteit in de afnamecondities (met als consequentie: systematische en/of toevallige ruis) onder controle te houden, o.a.:

- **de toetsassistenten zijn gekwalificeerd en degelijk getraind;**
- **de afnameprocedures zijn systematisch ontwikkeld en transparant;**
- **de kwaliteitsbewaking tijdens de afname is effectief.**

Scores van toetsen zijn onderhevig aan meetfouten die veroorzaakt worden door variabiliteit, onder meer in de afnamecondities. We kunnen die variabiliteit onder meer vermijden en/of onder controle houden door de meetprocedure te standaardiseren. Standaardisering van de toetsafname houdt in dat aspecten zoals tijd, taken, materiaal,


locatie bij verschillende toetsafnames zo veel mogelijk op dezelfde manier vorm worden gegeven. Dit draagt bij tot de vergelijkbaarheid van toetsscores overheen contexten en tot de eerlijkheid van een toets. Alle (grootschalige) toetsen besteden aandacht aan standaardisering, ook toetsen die geen gebruik maken van 'performance assessment'. Eigen aan 'performance assessment' is echter dat het grote(re) risico's op variabiliteit in zich draagt. Dit heeft bijvoorbeeld te maken met de grotere complexiteit van 'performance assessment'-taken en het risico dat leerlingen op de ene locatie meer begeleiding krijgen bij het oplossen van de taak dan elders.

Om een voldoende graad van standaardisering te bereiken, is het noodzakelijk om richtlijnen uit te werken voor de toetsassistenten die de toets en de 'performance assessment'-taken zullen afnemen, onder de vorm van een toetshandleiding. Ook de selectie en training van toetsassistenten en beoordelaars (zie verder bij 'scoren') zijn kritische elementen in het kader van een gestandaardiseerde afname.

 We verwijzen naar Johnson, Penny, and Gordon (2009) voor een volledig overzicht van potentiële topics die in de toetshandleiding, resp. de training van toetsassistenten, aan bod kunnen komen.

Een andere belangrijke maatregel om variabiliteit in de afnamecondities onder controle te houden is het bewaken van de kwaliteit van de toetsafname zelf. Dit kan gebeuren door de toetsassistenten afwijkende gebeurtenissen tijdens de afname te laten noteren, zodat op basis van deze gegevens beslissingen kunnen worden genomen over het al of niet meenemen van de gegevens. Een andere optie is kwaliteitsmonitoren in een steekproef van de scholen te laten controleren of de toetsafname conform de richtlijnen gebeurde en of de steekproef voldoende integer bleef.

Voor de toetsafname wordt veelal teruggevallen op centraal getrainde toetsassistenten. Dit is echter een dure en logistiek soms omslachtige werkwijze. Alternatieve maatregelen zijn bijvoorbeeld het inzetten van lokale leerkrachten, in combinatie met centrale aansturing, en het gebruik van digitale systemen om het evenwicht tussen standaardisering en authenticiteit te helpen vormgeven.

 *Bij SSLN deed men een beroep op lokale leerkrachten (zie 4.5.). De leerkrachten werden ingezet om per leerling twee schrijfstukken te selecteren die representatief waren voor het prestatieniveau van de leerlingen. Het genre van deze schrijfstukken moest verschillend zijn en de schrijfstukken moesten afkomstig zijn uit twee verschillende vakgebieden. Leerkrachten konden ervoor kiezen om een bestaand schrijfproduct te selecteren of ze konden leerlingen in de klas een tekst laten. Leerkrachten kregen bij de selectie uitgebreide aanwijzingen en er werd gemonitord of deze wel werden opgevolgd. Bovendien werd elk schrijfproduct beoordeeld door drie onafhankelijke, externe beoordelaars. Men koos voor deze werkwijze omdat men leerkrachten het best in staat acht om het reële prestatieniveau van de leerling in te schatten. Leerkrachten hebben bovendien jarenlange ervaring met het werken met de prestatiestandaarden en zijn daarom in staat de selectie op adequate wijze door te voeren.*

NAP-ICTL (4.2.) werkte in het kader van haar toetsprogramma met scenario's. In deze scenario's werden verschillende taken in een welbepaalde volgorde geplaatst om zo een verhaal neer te zetten dat betekenisvol was voor de leerlingen. Scenario's zorgen ervoor dat de toets authentiek opgezet kan worden dan voordien mogelijk was. Tegelijkertijd zorgt de digitale omgeving er ook voor dat de afname sterk gecontroleerd wordt. Elke toetsmodule omvatte 'point in time items', waarbij leerlingen op dat moment zelf een 'enge' vaardigheid moesten tonen. De module eindigde steeds met een uitgebreidere taak. Deze werd in een meer 'open' omgeving uitgewerkt, in tegenstelling tot de sterk gecontroleerde 'point-in-time' items. Leerlingen konden bv. op hun stappen terugkeren, bronnen op het web bekijken, deze bronnen gebruiken. De rationale voor de 'open taken' was dat deze werkwijze overeenstemt met de omgeving in de echte wereld, waar men dit soort handelingen ook kan uitvoeren. Tegelijkertijd was de context nog steeds sterk gecontroleerd: alle leerlingen hadden toegang tot exact dezelfde hulpmiddelen, er waren grenzen aan de beschikbare tijd om de taak af te werken en enkel die software-features die leerlingen echt nodig hadden, werd ter beschikking gesteld. Omdat de taken plaatsvonden in een ICT-omgeving, was het relatief eenvoudig deze mate van controle uit te voeren.

Standaardisering van de toetsafname is belangrijk om meetfouten onder controle te houden. Een ver doorgedreven standaardisering houdt echter ook risico's in. Het standaardiseren van aspecten van de toetsprocedure die niet vastliggen in de criteriumsituatie, heeft mogelijk tot gevolg dat toetsscores niet representatief zijn voor het beoogde competentiedomein (zie verder bij de bouwsteen 'validiteit').

5.2.5 Scoren

Bij grootschalige competentietoetsen is het ook cruciaal om mogelijke variabiliteit met betrekking tot het scoren te minimaliseren. Het is belangrijk dat het beoordelingsproces zoveel mogelijk wordt gestandaardiseerd - en dus vergelijkbaar en eerlijk - verloopt. In deze bouwsteen gaat het ook om het belang van transparantie en doelgerichtheid wat betreft de methodes die gebruikt worden voor het omzetten van de ruwe toetsscores en eventueel ook voor het equivaleren van de afgeleide scores.

Voorwaarde 12: Er worden maatregelen getroffen om de variabiliteit in het beoordelen (met als consequentie: systematische en/of toevallige ruis) onder controle te houden, o.a.:

- **de beoordelaars zijn gekwalificeerd en degelijk getraind;**
- **de scoringsprocedures zijn systematisch ontwikkeld en transparant;**
- **er worden meerdere beoordelaars ingezet;**
- **de kwaliteitsbewaking tijdens het scoren is effectief.**

Net als bij de toetsafname het geval was, komt het er ook bij deze bouwsteen op aan de nodige maatregelen te treffen om variabiliteit onder controle te houden. Deze maatregelen

concentreren zich rond beoordelaars, scoringsprocedures en de kwaliteitsbewaking tijdens het scoringsproces.

Het veronderstelde gebrek aan betrouwbaarheid bij het scoren is een van de vaakst gehoorde kritieken op 'performance assessment'. Bij 'performance assessment' is het consistent en ondubbelzinnig beoordelen van de kwaliteit van de 'performance' inderdaad veel minder eenvoudig dan bij klassiek opgezette toetsen. Een 'performance assessment'-taak kan immers meestal niet eenvoudig in termen van 'juist' of 'fout' beoordeeld worden. Het vereist een inschatting van de beoordelaar, waardoor de beoordeling steeds een bepaalde mate van subjectiviteit inhoudt. Indien echter voldoende aandacht besteed wordt aan het selecteren en trainen van de beoordelaars en aan (het monitoren van) het scoringsproces zelf, kan er zeker voldoende betrouwbaarheid gegarandeerd worden.


💡 *In het kader van NAP-ICTL ontwikkelde men een uitgebreid protocol voor het scoren dat o.m. ook procedures omvat om consistent, valide en objectief te beoordelen (4.2.). In de fase van het scoren werden 18 beoordelaars en 2 coördinatoren aangesteld die ook reeds betrokken waren in het verkennend onderzoek en/of vorige cycli van NAP-ICTL. Deze werkwijze droeg bij tot de consistentie in het gebruik van de scoringstools voor de trenditems, en maakte het trainingsproces zelf ook efficiënter en meer betrouwbaar. De training werd opgebouwd rond elk item en elke taak afzonderlijk. Meteen na elke training werden alle antwoorden voor het betreffende item of de betreffende taak gescoord. Zo bleven de beoordelaars meer op het betreffende item gefocust, was het eenvoudiger om de scoringscriteria te onthouden en werden de beoordelaars in staat gesteld om snel een grote set gegevens te beoordelen. Wat de trainingssessie meer specifiek betreft, selecteerde men voor elk item en elke taak 5 tot 20 antwoorden. De coördinator kende meteen scores toe aan deze selectie. Naarmate de beoordelaars vervolgens bij wijze van training doorheen de items en taken gingen en scores toekenden, gaf de software aan wanneer er sprake was van inconsistentie. In dergelijke gevallen, werden de scoringscriteria nogmaals verduidelijkt. Voor elk verschillend item en taaktype werd een afzonderlijke scoringsprocedure en -tool gebruikt. 10 % van de antwoorden werd dubbel gescoord door de aangestelde coördinator. In het geval van inconsistente scores werden de beoordelaars opnieuw getraind met betrekking tot dat specifieke item en werden de antwoorden opnieuw gescoord.*

Het trainen van beoordelaars is essentieel om betrouwbare resultaten te bekomen. Dergelijke trainingen kunnen zowel centraal als lokaal (al of niet via de computer) georganiseerd worden. Via training streeft men ernaar de beoordelaars de beoordelingscriteria op dezelfde wijze te laten toepassen. Heldsinger and Humphry (2010) zochten naar een oplossing om deze cruciale afstemming minder tijdrovend te maken. In hun studie gingen ze na of een alternatieve aanpak, die uit twee stappen bestaat, ook tot een aanvaardbaar niveau van consistentie in de beoordelingen kan leiden. In een eerste stap evalueren leerkrachten schrijfoopdrachten van leerlingen uit de lagere school via een systeem van paarsgewijze vergelijking, om zo tot een set van gekalibreerde 'exemplars' te komen: voorbeelden die als ankerpunt dienen om andere opdrachten mee te vergelijken. Deze voorbeelden worden vervolgens in een tweede stap gebruikt door de leerkrachten om andere schrijfoopdrachten te beoordelen. Deze werkwijze bleek een betrouwbaar, valide en


bovendien tijd- en kostenefficiënt alternatief voor groepen van leerkrachten die samenkomen om opdrachten te bediscussiëren in relatie tot de beoordelingscriteria en hun beoordelingen te modereren.

Een ander belangrijk aspect dat aan bod dient te komen in de training heeft betrekking op het onder de aandacht brengen van mogelijke bias vanwege de beoordelaars.

Vertekeningen kunnen bijvoorbeeld te maken hebben met: herhaling (toekennen van een lagere score omdat de beoordelaar reeds vertrouwd is met het topic of het antwoord eerder al gelezen heeft); lengte (toekennen van hogere scores aan langere antwoorden); en strengheid/mildheid (neiging om consistent te streng/te mild te zijn in de scores).

 Johnson, Penny, and Gordon (2009, 211) en Engelhard (2002) geven meer informatie over verschillende types vertekingen.

Het scoren zelf, inclusief het bewaken van dit scoringsproces, is een volgend belangrijk aspect in het streven naar betrouwbare scores (zie ook bovenstaande praktijkillustratie van NAP-ICTL). Vertekeningen dragen bij tot meetfouten die mogelijks resulteren in onbetrouwbare toetsscores. Daarom is het belangrijk om tijdens het scoringsproces blijvend toezicht te houden op de kwaliteit van de beoordelingen. Gericht feedback kan ertoe leiden dat beoordelaars hun scoringsgedrag aanpassen en discrepanties oplossen. Daartoe kunnen verschillende methodes aangewend worden, zoals bijvoorbeeld het nakijken van bepaalde beoordelaarsstatistieken, het beoordelen van de beoordelaarsovereenstemming en/of het gebruik van recalibratiesets. Tijdens de fase van het toezicht is het ook belangrijk aandacht te besteden aan de manier waarop men omgaat met afwijkende scores.


 Er zijn een aantal manieren om kwesties met betrekking tot tegenstrijdige scores opgelost te krijgen. Zie Johnson, Penny, and Gordon (2009, 241) voor een overzicht.

Ook het inzetten van verschillende beoordelaars bij het beoordelen van elke taak leidt tot meer betrouwbare scores (zie ook bouwsteen 'validiteit'). Indien het niet haalbaar is om voor elke taak verschillende beoordelaars in te zetten, kan men ook een subset taken dubbel laten scoren.

Al deze maatregelen om betrouwbare scores te garanderen, worden beschouwd als kenmerken van een goede toetspraktijk. Er gaan echter ook stemmen op om subjectiviteit bij de beoordelaars niet steeds als een bedreiging op te vatten. Zo kan het net verrijkend zijn om gebruik te maken van meerdere perspectieven bij het beoordelen van complexe performances. Een zekere graad van overeenstemming onder de beoordelaars is weliswaar een vereiste, maar kleine verschillen tussen de beoordelaars hoeven niet problematisch te zijn.

De maatregelen hierboven om variabiliteit in het scoren aan te pakken, hebben evenwel consequenties in termen van middelen en tijd. Het terugvallen op menselijke beoordelaars maakt het beoordelen sowieso duur: het uitgebreid trainen van beoordelaars kost tijd en

geld, alsook de inzet van meerdere beoordelaars. Tegen deze achtergrond bieden zich een drietal alternatieve oplossingen aan. Een eerste alternatief is de comparatieve beoordeling of paarsgewijze vergelijking, die aan beoordelaars vraagt om prestaties van leerlingen paarsgewijs te vergelijken en steeds aan te geven welke prestatie zij als beste van de twee identificeren. Meerdere beoordelaars beoordelen de prestaties meerdere keren. Op basis van al deze vergelijkingen wordt een rangorde van de prestaties opgesteld. De methode maakt er overigens aanspraak op om meer valide te zijn, omdat de beoordelaars hun keuze baseren op basis van een holistische evaluatie.

 Heldsinger and Humphry (2010), S. Heldsinger and Humphry (2013), Lesterhuis et al. (2015), Lesterhuis et al. (2017), van Daal et al. (2019) en Steedle and Ferrara (2016) hebben het potentieel van paarsgewijze vergelijking wat betreft haalbaarheid aangetoond.


Een andere mogelijke oplossing wordt aangereikt door recente software-ontwikkelingen, die het mogelijk maken om producten zoals bijvoorbeeld essays automatisch te scoren. Voldoende omzichtigheid bij het inzetten van geautomatiseerd scoren is echter op zijn plaats. Zo wijst S. Lane (2015) er bijvoorbeeld op dat het geloof in geautomatiseerd scoren vooral steunt op studies die de uitwisselbaarheid van automatisch toegekende scores en scores vanwege ‘menselijke’ beoordelaars, onderzoeken. In dergelijke studies wordt meestal aangetoond dat de relatie tussen automatische scores en menselijke scores vergelijkbaar is met die verkregen tussen twee menselijke beoordelaars. Er zijn echter minder studies die ingaan op het scoringsproces zelf bij automatisch scoren, terwijl dit net belangrijke informatie oplevert over de validiteit van de scoringsmethode. Dit type onderzoek (bijvoorbeeld Ben-Simon and Bennett (2007)) toont meer specifiek aan dat de dimensies die experts belangrijk vinden bij het evalueren van schrijven, niet noodzakelijk dezelfde zijn als deze die gehanteerd worden in de automatische scoringsprocedures. Ook Lu (2012) is van mening dat er te weinig verschillende types bewijsmateriaal worden verzameld bij het valideren van automatisch scoren van essays. Haar onderzoek toonde problemen aan bij de validiteit van het automatisch scoren: terwijl het theoretische onderscheid tussen hogere-orde- en taaleigenschappen werd bevestigd, was dit niet het geval bij het automatisch scoren. Bovendien zijn programma’s voor automatisch scoren er niet op gericht individualiteit of bijvoorbeeld poëtische inspiratie te appreciëren en focussen ze daarom meer op conformiteit.


Nog een oplossing bestaat erin om lokale leerkrachten in te zetten om de eigen leerlingen te beoordelen (zie bijvoorbeeld NMSSA-Arts, 5.2.2., Voorwaarde 6). Dit vereist echter extra waakzaamheid, met name in verband met het opduiken van meetfouten. Kuhlemeier, Hemker, and Bergh (2013) geven bijvoorbeeld aan dat leerkrachten milder blijken te zijn in hun beoordelingen en de neiging hebben om minder accuraat onderscheid te maken tussen kandidaten dan gerechtvaardigd is op grond van de eigenlijke performance. Onderzoek in de VS (National Research Council, 2014) toonde dan weer aan dat de interbeoordelaarsbetrouwbaarheid tussen leerkrachten, die de neiging hebben milder te zijn in hun oordeel, en externe beoordelaars, na enkele jaren van werken met het systeem sterk verbeterde.

Voorwaarde 13: De methode om tot afgeleide scores (bv. schaalscores) te komen is transparant en ligt in lijn met de doelstellingen (cf. doelbepaling).

Met het oog op het rapporteren van de prestaties van leerlingen hebben de ruwe scores van leerlingen weinig betekenis. Om vergelijkbaarheid tussen scores te bekomen, worden ruwe scores veelal getransformeerd naar een schaal. Doorgaans gebruikt men modellen uit de item response theorie (IRT) om toetsen die louter uit 'performance assessment'-taken bestaan of toetsen die uit 'performance assessment'-taken en meerkeuzevragen bestaan, te schalen. IRT-modellen laten toe de resultaten van alle leerlingen op een gemeenschappelijke schaal te plaatsen, niettegenstaande het feit dat deze individuen andere steekproeven van taken en items hebben ontvangen. Bovendien kan men op basis van IRT in de fase van de rapportering, eenzelfde schaaleenheid naar voor schuiven, los van de verschillende testvormen en -taken die in de toets zijn opgenomen (Hambleton 2006).

Aan het gebruik van IRT bij 'performance assessment' zijn echter ook een aantal uitdagingen verbonden. Deze uitdagingen hebben te maken met de assumptie van lokale onafhankelijkheid van items en/of taken, de assumptie van unidimensionaliteit en het inzetten van link- of ankertaken.

 Meer informatie hieromtrent is terug te vinden bij o.a. Davey et al. (2015), Kolen and Brennan (2014) en National Research Council (2014).


 *Typierend voor 'performance assessment' is dat de toetsvragen/activiteiten verband houden met elkaar, om op die manier voor de leerling een betekenisvol geheel te kunnen vormen. De assumptie van lokale onafhankelijkheid van items en/of taken vereist echter dat toetsvragen/activiteiten idealiter niet met elkaar in verband mogen staan. In het kader van een van de casussen, omvatten de toetsmodules sequenties van items en taken die ontwikkeld werden rond een specifieke verhaallijn ('narrative') en zodoende bijeen horen. Hoewel men erkent dat het vanuit meetoogpunt ideaal zou zijn om leerlingen ad random items en taken voor te leggen, houdt dit een duidelijk verlies aan authenticiteit in en limiteert men zich in wat men kan meten. In hun visie is dit een frictie waar ze mee moeten omgaan: met het oog op de authenticiteit van de toets, wil men levensechte, scenariogebaseerde toetsen aanbieden, niettegenstaande het feit dat dit leidt tot het schenden van de assumptie van lokale onafhankelijkheid, waardoor ongewenste variantie geïntroduceerd wordt. Tegelijkertijd wordt de lokale afhankelijkheid reeds vanaf de pilots van de toets gemonitord. In het geval bepaalde scoringscriteria (items) een hoge graad van lokale afhankelijkheid hebben, passen ze de criteria bijvoorbeeld aan door ze samen te voegen (voorbeeld van 2 criteria -'inhoud titel' en 'lay-out titel'- die conceptueel verschillend zijn, maar die bij de toepassing van de criteria 1 criterium blijken te zijn: 'kwaliteit titel').*

Voorwaarde 14: De eventueel gehanteerde equivaleringstechniek is transparant en ligt in lijn met de doelstellingen (zie doelbepaling).

Afhankelijk van de doelstellingen die in de fase van doelbepaling werden vastgelegd, dient men toetsscores al of niet te equivaleren. Equivalering is een methode om scores op twee toetsen zodanig rechtstreeks te linken dat de scores inwisselbaar worden, alsof ze afkomstig zouden zijn van één enkele toets. Een toetssysteem of -programma kan bijvoorbeeld de ambitie hebben om de scores van het vierde jaar secundair onderwijs te vergelijken met scores van het tweede jaar secundair onderwijs, of om scores voor probleemoplossend denken uit een recente peiling te vergelijken met die uit een of meerdere voorgaande jaren. Equivalering helpt in deze gevallen om de vergelijkbaarheid van toetsscores te garanderen.

Voorwaarde om te kunnen equivaleren is dat er een set gemeenschappelijke taken, zogenaamde link- of ankertaken, afgenomen werden in de betreffende leerjaren/kalenderjaren. Op dat punt duikt voor 'performance assessment' een moeilijkheid op. Het feit dat de meeste equivaleringsdesigns steunen op het hergebruik van minstens een aantal van de gebruikte taken ('ankertaken') is problematisch voor toetsen die een 'performance assessment'-component bevatten. 'performance assessment'-taken zijn immers vaak makkelijk te memoriseren door leerlingen en kunnen daarom moeilijker gebruikt worden als link tussen verschillende afnames van een toets. Gevolg is immers dat (andere) leerlingen op voorhand kunnen oefenen wat een zekere invloed kan hebben op de mate waarin de taak het beoogde construct meet.

Equivaleringsmethoden kunnen gebaseerd zijn op IRT of niet. National Research Council (2014) en Kolen and Brennan (2014) gaan dieper in op de problematiek rond equivalering op basis van IRT en reiken een aantal manieren aan om hiermee om te gaan. Naast IRT bestaan er ook alternatieve technieken voor equivalering, die ook resulteren in betrouwbare en valide scores, zoals bijvoorbeeld de comparatieve beoordeling of paarsgewijze vergelijking [Heldsinger and Humphry (2010), S. Heldsinger and Humphry (2013), Lesterhuis et al. (2015), Lesterhuis et al. (2017), van Daal et al. (2019) en Steedle and Ferrara (2016)].

 *In NAPLAN-Persuasive Writing (zie 4.1.) maakte men voor de equivalering gebruik van een combinatie van paarsgewijze vergelijking, IRT en regressie-analyses. In het kader daarvan voorziet men voor elk kalenderjaar een andere, weliswaar vergelijkbare schrijfofdracht. Deze taak wordt elk jaar gescoord aan de hand van eenzelfde rubric met tien criteria. De scores op de rubric worden opgevat als 'items' en worden vervolgens in een IRT analyse gehanteerd (Partial Credit Model). Op grond van die modellen leidt men schaalscores af die vervolgens gerapporteerd kunnen worden.*

De grote uitdaging in dit proces is het garanderen van vergelijkbaarheid overheen verschillende afnamejaren (bv. 2011 versus 2014). Binnen NAPLAN kiest men expliciet om elk afnamejaar een andere taak te voorzien. Er zijn dus geen gemeenschappelijke taken. Bovendien is het ook niet zo dat bij een deel van de leerlingen uit 2014 twee taken afgenomen worden (zowel schrijftaken uit 2011 als 2014). De consequentie hiervan is dat men voor het equivaleren niet kan terugvallen op een 'common item'- of een 'common person'-design.

De alternatieve strategie die ze bij NAPLAN-Persuasive Writing aanwenden om toch te kunnen equivaleren berust op de volgende redenering: gesteld dat de twee verschillende schrijftaken in de respectievelijke afnamejaren even moeilijk zijn en gesteld dat de beoordelaars op beide momenten even streng zijn, dan zijn de scores die toegekend worden aan de hand van de gebruikte rubric perfect vergelijkbaar overheen beide afnamejaren. Net om deze assumptie te toetsen, gebruikt men de methode van paarsgewijze vergelijking. Voor de paarsgewijze vergelijkingsstudie selecteerde men een set van schrijfproducten uit 2014 en een set van schrijfproducten uit 2011. Beide sets werden vervolgens tesamen paarsgewijs beoordeeld door een deel van de beoordelaars. Een beoordelaar kreeg telkens twee schrijfproducten te zien en diende aan te geven welke van beide de beste was. Deze paren konden overigens enkel uit teksten uit 2011 bestaan, of enkel uit teksten uit 2014, maar net zo goed kon het om een tekst uit 2011 en een tekst uit 2014 gaan. Elke tekst in de subset werd ongeveer 30 keer vergeleken met een andere tekst. Op basis van een het 'Bradley Terry Luce model' (een variant op het Rasch model, voor paarsgewijze data) werd een gemeenschappelijke schaal (en bijhorende scores) van schrijfproducten uit 2011 en 2014 gevormd. Vervolgens werd voor de subset van schrijfproducten uit 2014 de schaalscore resulterend uit de paarsgewijze vergelijking statistisch gerelateerd aan de scores die deze schrijfproducten behaalden op basis van de rubrics. Daartoe voerde men een regressieanalyse uit. Voor de subset van schrijfproducten uit 2011 die paarsgewijs vergeleken werden, deed men net hetzelfde. De resultaten van beide regressieanalyses geven inzicht in de vergelijkbaarheid van scores. Indien er geen significant verschil werd vastgesteld tussen intercepten in beide regressieanalyses, concludeerde men eenvoudig dat er geen aanpassingen nodig waren om de scores vergelijkbaar te maken. Indien dit verschil in intercepten wel significant was, dan werd gebruik gemaakt van de informatie uit beide regressieanalyses om de scores van alle teksten uit 2014 (en dus niet enkel deze subset die deel uitmaakte van het equivaleringsonderzoek) bij te stellen en vergelijkbaar te maken met de scores uit 2011.

Los van de techniek of methode die gehanteerd wordt om toetsscores vergelijkbaar te maken, is het van belang dat voldoende transparant gemaakt wordt welke de gehanteerde technieken zijn en dat de gemaakte keuzes verantwoord worden.

5.2.6 Validiteit

Bij deze bouwsteen nemen we even afstand en overlopen we of wel voldaan werd aan alle voorwaarden om de toetsscores te kunnen generaliseren naar het toetsdomein en deze scores vervolgens ook te extrapoleren naar het beoogde competentiedomein (zie ook 3.1.). De voorwaarden in deze bouwsteen houden verband met 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 aanzien van het toetsdomein en het beoogde competentiedomein.

Voowaarde 15: Generaliseren:

- **de PA-taken zijn representatief voor het toetsdomein;**
- **de PA is zodanig opgezet dat toevallige ruis zoveel mogelijk onder controle gehouden wordt; en**

- **de effectiviteit van maatregelen om toevallige ruis onder controle te houden wordt nagegaan.**

Generaliseerbaarheid heeft te maken met het gegeven dat scores van leerlingen variëren overheen replicaties van de toets getrokken uit het toetsdomein. Dit wil zeggen dat de score van een leerling varieert naargelang van bijvoorbeeld de specifieke taak, het specifieke afnamemoment en de specifieke beoordelaar. Deze steekproefvariabiliteit stelt grenzen aan de mogelijkheid om scores op een welbepaalde toets te generaliseren naar het toetsdomein. Algemeen geldt dat naarmate de steekproefomvang (van bv. taken, afnamemomenten, en beoordelaars) groter wordt, de generaliseerbaarheid van scores toeneemt. We kunnen de consistentie dus verbeteren door het aantal onafhankelijke observaties met betrekking tot elk facet (bv. taken, afnamemomenten en beoordelaars) te vergroten. Een andere maatregel om de generaliseerbaarheid te bevorderen houdt in om kenmerken van de ‘performance assessment’-taken, de administratieprocedures en het beoordelingsproces te standaardiseren. Toevallige meetfouten kunnen we dus onder controle houden door grotere steekproeven in de toets op te nemen en door standaardisering van de meetprocedure.

Een belangrijke voorwaarde in verband met de mogelijkheid om scores te generaliseren naar het toetsdomein houdt in dat de ‘performance assessment’-taken representatief zijn voor het toetsdomein. We verwijzen hiervoor naar de bouwstenen ‘domeinbeschrijving’ en ‘opzet en ontwikkeling’. Een andere voorwaarde houdt verband met het minimaliseren van toevallige ruis, die het resultaat is van variabiliteit in bijvoorbeeld taken of beoordelaars (zie bouwstenen ‘opzet en ontwikkeling’ en ‘scoren’). Aan deze voorwaarde is meteen ook een derde voorwaarde verbonden, met name die in verband met het nagaan van de effectiviteit van de maatregelen die getroffen werden om toevallige ruis te minimaliseren. Het nagaan van deze effectiviteit is o.a. mogelijk via betrouwbaarheidsstudies, generaliseerbaarheidsstudies of via IRT-gebaseerde informatiefuncties.

Voowaarde 16: Extrapoleren:

- **de PA-taken zijn representatief voor het beoogde competentiedomein (d.i. het vermijden van systematische ruis);**
- **de effectiviteit van maatregelen om systematische ruis te vermijden wordt nagegaan.**

Extrapolerbaarheid van scores naar het beoogde competentiedomein impliceert dat de prestaties op taken in een toets een goede indicator zijn voor prestaties op criteriumtaken uit de alledaagse context. De mate waarin toetstaken op taken in de criteriumsituatie lijken, kan zich op verschillende manieren veruitwendigen. Bijvoorbeeld via de inhoud en vorm van de taak, of de fysieke omgeving en sociale context waarin de taak wordt uitgevoerd. Met het oog op de extrapolerbaarheid van toetsscores is het met andere woorden noodzakelijk dat de taken in de toets een prestatie uitlokken die een weerspiegeling vormt van de beoogde competentie. Alles hangt dus af van de gelijkenis tussen het toetsdomein en het

beoogde competentiedomein. Indien er geen grote verschillen zijn, is de extrapoleerbaarheid waarschijnlijk.

Een eerste belangrijke voorwaarde in verband met de mogelijkheid tot het extrapoleren van toetsscores is dat de 'performance assessment'-taken uit het toetsdomein representatief zijn voor het competentiedomein waaruit het toetsdomein werd afgeleid (zie bouwstenen 'domeinbeschrijving' en 'opzet en ontwikkeling'). Dit betekent zoveel als het vermijden van systematische ruis.

Een tweede voorwaarde houdt verband met het checken van de effectiviteit van de maatregelen die getroffen werden om systematische ruis te vermijden. In dit verband kunnen verschillende soorten bewijzen aangeleverd worden: analytische bewijzen en/of empirische bewijzen. Analytisch bewijs wordt voornamelijk gegenereerd tijdens de ontwikkelingsfase van de toets. Voorbeelden van analytisch bewijs zijn:

- het nagaan van de cognitieve complexiteit van 'performance assessment'-taken via 'cognitive labs';
- het bevragen van experts wat betreft de afstemming van de taken op het beoogde competentiedomein;
- specifieke statistische analyses naar aanleiding van een pilootonderzoek om het model te valideren.

Empirisch bewijs wordt vooral verzameld tijdens de pilotering van de toets en op grond van (statistische) analyses op de verzamelde gegevens. Correlatiecoëfficiënten die, in het kader van convergente, discriminerende of predictieve validiteit, de relatie tussen de toetsscore en andere scores geassocieerd met het beoogde competentiedomein in kaart brengen, vormen een voorbeeld van empirisch bewijs.

Met betrekking tot 'generaliseren' en 'extrapoleren' stoten we, zeker in het geval van 'performance assessments', op een onvermijdelijke paradox. Het delicate evenwicht tussen generaliseren en extrapoleren wordt duidelijk vanuit 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, enz.), blijken in realiteit moeilijk te combineren met maatregelen die de extrapoleerbaarheid van de scores beogen te vergroten, zoals het uitwerken van authentieke taken die recht doen aan de criteriumsituatie. Het standaardiseren van aspecten van de toetsprocedure die niet als dusdanig vastgelegd zijn in de criteriumsituatie, heeft tot gevolg heeft dat toetsscores niet geëxtrapolerd kunnen worden naar het beoogde competentiedomein.

💡 *In de praktijkvoorbeelden zien we dat er met betrekking tot de afweging 'generaliseren'- 'extrapoleren' twee mogelijke sporen worden gevolgd. Enerzijds is er een groep systemen waarbij maximale standaardisering het uitgangspunt vormt. De primaire doelstelling is bij deze toetsprogramma's om toetsen op te zetten die scores opleveren die vergelijkbaar zijn tussen verschillende toetsafnames, 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 de taakhoud en de afname wordt begrensd door de primaire doelstelling van vergelijkbaarheid.

Anderzijds zijn er praktijkvoorbeelden die de criteriumsituatie als primair vertrekpunt nemen. 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 de criteriumsituatie 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.

Een belangrijke slotbemerking heeft betrekking op het fenomeen van toevallige meetfouten, in relatie tot het niveau waarop men de toetsresultaten rapporteert en interpreteert. Toevallige meetfouten kan men enerzijds onder controle houden door grotere steekproeven te betrekken, wat zich bijvoorbeeld kan uiten in het voorzien van meer taken, beoordelaars en/of afnamemomenten. Anderzijds biedt het standaardiseren van de meetprocedure een mogelijkheid om toevallige meetfouten onder controle te houden.

Bij rapportering op een hoger aggregatieniveau (bv. nationaal niveau) zijn meetfouten op het niveau van individuele leerlingen ten gevolge van dergelijke vertekeningen namelijk een minder grote zorg. Onderzoek van Hill and DePascale (2003) toont immers aan dat betrouwbaarheidsscores die een bron van zorg kunnen zijn bij rapportering op individueel niveau, nog steeds vaststellingen op hogere niveaus, kunnen ondersteunen. Ze toonden in hun onderzoek aan dat met name de omvang van de steekproef leerlingen een grotere impact heeft op de betrouwbaarheid van het schoolgemiddelde, dan de betrouwbaarheid van de individuele leerlingresultaten. Een toetssysteem kan dus betrouwbaar zijn, zelfs als de individuele leerlingenscores maar matig betrouwbaar zijn. Een van de aanbevelingen die uit de studie voortkomt, is om authentieke toetsvormen te gebruiken, zelfs indien dit ten koste gaat van lagere betrouwbaarheid op het niveau van de leerling. Bovendien weten we uit de 'Standards' (AERA APA & NCME 2014) 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').

5.2.7 Niveaubepaling en rapportering

Deze bouwsteen omvat het vastleggen van de prestatiestandaarden of cesuren waartegen de scores worden afgezet enerzijds en het rapporteren van de toetsscores anderzijds.

Voorwaarde 17: De methode van cesurbepaling is transparant en ligt in lijn met de doelstellingen (zie doelbepaling).

Om een valide interpretatie en gebruik van scores te ondersteunen, moet men scores betekenisvol maken. Het vastleggen van prestatiestandaarden of cesuren is een vaak gebruikte manier om betekenis te verlenen aan toetsscores. Op grond van deze standaarden kunnen leerlingen in bepaalde categorieën worden ingedeeld.

💡 *NAPLAN (zie 4.1.) werkt bijvoorbeeld met één prestatiestandaard per leerjaar en deelt de leerlingen op basis hiervan vervolgens op in drie groepen: 'onder de nationale minimumstandaard', 'op de nationale minimumstandaard' en 'boven de nationale minimumstandaard'. Ook NAP-ICTL (zie 4.2.) legde voor elk van beide leerjaren een prestatiestandaard ('Proficient Standard') vast. Deze valt voor leerjaar 6 bijvoorbeeld samen met schaalscore '409': de grens tussen bekwaamheidsniveau 2 en bekwaamheidsniveau 3.*

Prestatiestandaarden hangen best af van de kennis, vaardigheden en attitudes die noodzakelijk zijn voor een aanvaardbaar niveau van de beoogde competentie. Het komt er met andere woorden op neer de prestatiestandaarden te verankeren in de inhoud en in datgene wat leerlingen moeten kennen en kunnen. De focus ligt bij de cesuurbepaling op het vastleggen van een correcte prestatiestandaard, op basis van een geschikte procedure, in lijn met de eerder geformuleerde doelstellingen.

Er zijn verschillende methoden beschikbaar om prestatiestandaarden vast te leggen. Hambleton (2006) presenteren een interessante benadering om de verschillende methoden om prestatiestandaarden vast te leggen, te classificeren. Het betreft een uitbreiding van de typische opdeling in methoden die zich richten op de toets en methoden die zich richten op de beoordeelde. De focus van de oordelen van de panellisten brengen ze onder in vier categorieën van methoden. Het gaat om methoden die een beoordeling omvatten van:

- toetsitems en scoringsrubrieke (bv. (extended) Angoff);
- kandidaten (bv. contrasting groups);
- het werk van kandidaten (bv. 'body of work method' en 'analytic judgment method');
- van scoreprofielen (bv. dominant profile method).


💡 *Bij NAP-ICTL (4.2.) werden de prestatiestandaarden vastgelegd na een consultatieproces van twee jaar. Men hanteerde een 'empirical judgmental technique', waarbij stakeholders de toetsitems en de resultaten van de toets beoordeelden en vervolgens een standaard overeen kwamen voor elk van beide jaargroepen. De panels bestonden uit leerkrachten met specifieke ICT expertise, ICT- en toetsexperten.*

Het vastleggen van prestatiestandaarden is een proces dat uit verschillende stappen bestaat en behelst steeds een mix van inschattingen ('human judgment'), psychometrie en uitvoerbaarheid ('practicality') (zie o.a. Hambleton (2006)). De subjectieve aard is een gegeven waar men niet omheen kan; het menselijk oordeel speelt altijd een kritieke rol. Dit geldt bv. bij de keuze van de methode die gehanteerd zal worden, het samenstellen van het panel en niet in het minst bij de resulterende prestatiestandaarden. Het doel van al het

werk omheen het zetten van prestatiestandaarden is om de beoordelingen zo geïnformeerd mogelijk te laten gebeuren.

Als gevolg van de unieke karakteristieken van 'performance assessment' in vergelijking met klassieke toetsen, zijn traditionele methoden om prestatiestandaarden vast te leggen niet zonder problemen. Dit komt o.a. door:


- het gebruik van complexe en polytome rubrics;
- de multidimensionaliteit in de data (taken doen beroep op meerdere vaardigheden);
- de onderlinge afhankelijkheid in de rubrics (bijv. niet in staat zijn om een taak af te werken omdat men een stuk ervan niet kon);
- de beperkte generaliseerbaarheid van scores op taakniveau (het goed kunnen van een bepaalde groep taken impliceert nog niet dat men andere taken ook goed kan).

 Hambleton et al. (2000) beschrijven kenmerken van performance assessment die het standaardzettingsproces

In vergelijking met traditionele manieren om prestatiestandaarden vast te leggen, zijn de methoden voor het vastleggen van prestatiestandaarden bij 'performance assessment' echter nog niet zo goed ontwikkeld en gevalideerd. Het advies in dit verband is om een degelijk veldonderzoek op te zetten en uit te voeren alvorens de betreffende standaardzettingmethode te implementeren.

Voorwaarde 18: De methode van rapportering is transparant en ligt in lijn met de doelstellingen (zie doelbepaling).

Het doel van de toets vormt het raamwerk waarbinnen wordt gerapporteerd. Indien er sprake is van tegenstrijdigheid in die doelstellingen, kan dit leiden tot problemen bij de interpretatie die aan de toetsscores wordt gehecht.

 *De toets in het kader van NAPLAN-Persuasive Writing (zie 4.1.) heeft als primaire doelstelling om informatie te verzamelen die beslissingen op systeemniveau en het niveau van de scholen kan informeren (zgn. low-stakes toets). Op deze niveaus geldt dat de foutenmarges aanvaardbaar zijn. Er worden echter ook resultaten aangeleverd op het niveau van de individuele leerlingen. Hoewel er voor de individuele leerling formeel niks van de toetsresultaten afhangt (low-stakes toets), percipiëren leerlingen en hun ouders dit vaak anders. Het gevaar bestaat, volgens Hornsby and Wu (2012), dat deze groepen gebruikers vaak geen rekening houden met de aanzienlijke onnauwkeurigheden en onzekerheden waaraan deze resultaten onderhevig zijn.*

In de rapportering worden de prestaties van de leerlingen ten aanzien van de gemeten competentie in kaart gebracht. Tegemoet komen aan de interesses en behoeften van een uiteenlopend lezerspubliek vereist een zekere variëteit aan rapporten. Het brede publiek is

bijvoorbeeld meer gebaat met een eenvoudige, inzichtelijke brochure; experts en opdrachtgevers (overheden) hebben vooral behoefte aan uitgebreide rapporten. Doorgaans rapporteert men samenvattende resultaten zoals bijvoorbeeld gemiddelde (schaal)scores en/of percentages leerlingen die een bepaald prestatieniveau halen. Scores kunnen overigens niet alleen gerapporteerd worden voor de volledige populatie, maar ook uitgesplitst naar relevante leerlingkenmerken (bv. regio, geslacht, SES, ...).

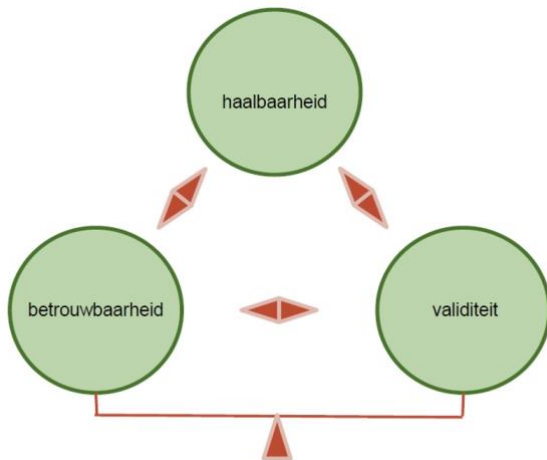
Daarenboven moet de betekenis van de scores voor een divers publiek zodanig vatbaar worden gemaakt, dat ze een valide interpretatie ondersteunen. Met het oog op het geven van betekenis aan scores, kan men naast het formuleren van prestatiestandaarden ook andere werkwijzen inzetten, zoals bv. 'item mapping' en 'scale anchoring'. 'Item mapping' houdt in dat, nadat de schaal ontwikkeld werd, taken worden gezocht die representatief zijn voor verschillende punten of locaties op de schaal. 'Scale anchoring' heeft als doel om voor geselecteerde punten op de schaal een toelichting te geven van wat een leerling op dat welbepaalde punt kan.

Ook belangrijk met het oog op het maken van valide interpretaties en vergelijkingen is het rapporteren van de standaardfout. Het rapporteren van deze standaardfout op zich is echter niet voldoende, het lezerspubliek dient ook aangemoedigd te worden om die in rekening te brengen bij de interpretatie van de resultaten.

5.2.8 Haalbaarheid van de bouwstenen

Elk van de vorige zeven bouwstenen wordt geflankeerd door de 'haalbaarheidsvoorwaarde': hoe kwaliteitsvol de opzet van een grootschalige competentietoets ook mag zijn, uiteindelijk kan die enkel maar in de praktijk worden gerealiseerd indien er voldoende tijd en middelen (o.a. personeel, logistiek, materiaal) beschikbaar zijn.

Centraal in het opzetten van 'performance assessment' staat het delicate evenwicht tussen de bouwstenen 'generaliseren' en 'extrapoleren': de zogenaamde betrouwbaarheid-validiteit paradox (zie ook 3.1 en 5.2.6.). In het kader van het opzetten van grootschalige toetsen die gebruik maken van 'performance assessment', stelt zich daarenboven ook de vraag of de gekozen opzet 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 5.4](#) maakt de evenwichtsoefening tussen de componenten betrouwbaarheid (generaliseerbaarheid), validiteit en haalbaarheid, visueel duidelijk.



Figuur 5.4: Evenwicht tussen componenten van 'performance assessment'.

Het opzetten van een toets is steeds een zoektocht 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: 'kiezen is verliezen'.

Childs and Jaciw (2019) onderscheiden verschillende kosten die in overweging genomen dienen te worden bij het uitwerken van een toetssysteem of -programma. Het gaat meer concreet om kosten verbonden aan:

- ontwikkeling: tijd en geld besteed aan de ontwikkeling (schrijven, redigeren, herzien, piloot- en veldonderzoek) van nieuwe items;
- materiaal: tijd en geld besteed aan printen en verzenden van toetsboekjes en andere materialen;
- afname: tijd die leerkrachten en ander schoolpersoneel besteden aan de afname van de toets;
- onderwijs: tijd die leerlingen steken in (het voorbereiden van) de toets, wijzigingen in de klaspraktijk van leerkrachten en het verwerken van het curriculum omwille van de toets, wijzigingen in de allocatie van schoolmiddelen omwille van de toets;
- scoring: tijd en geld besteed aan het scoren van de toetsen, zij het elektronisch, zij het door getrainde beoordelaars;
- betrouwbaarheid: wijzigingen in de accuraatheid en de consistentie van de toetsscores;
- vergelijkbaarheid: wijzigingen in de mate waarin de toetscores van verschillende leerlingen vergeleken kunnen worden;
- validiteit: wijzigingen in de mate waarin de toetsscores het construct reflecteren dat de toets beoogt te meten;
- rapportering: wijzigingen in het gemak waarmee de toetsscores uitgelegd kunnen worden aan leerkrachten, ouders en het publiek algemeen.

In een ideale situatie - met ongelimiteerde middelen - kan aan al deze kosten tegemoet gekomen worden en een ideaal toetssysteem of -programma worden opgesteld. Omdat dit in realiteit vaak niet het geval is, dienen toegevingen te worden gemaakt. Al deze kosten zijn aan elkaar gelieerd: besparen op de ene kost, leidt mogelijk tot extra uitdagingen ten aanzien van de andere kost. Zo is het afnemen van een beperkte steekproef van taken aan een beperkte steekproef van leerlingen weliswaar kostenbesparend, maar heeft het ook een invloed op de betrouwbaarheid en validiteit van de toets. Afhankelijk van de doelstelling van de toets, varieert ook het relatieve belang van de ene kost tegenover de andere.

5.2.9 Tot slot

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 vaak 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', is het ook belangrijk om de financiële en logistieke kosten zorgvuldig af te wegen.

6 Uitdagingen voor grootschalige toetsen die ‘performance assessment’ inschakelen

De systematische literatuurstudie in combinatie met een doorgedreven analyse van internationale 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. Hieronder vatten we beknopt samen over welke uitdagingen het gaat en welke alternatieve oplossingen zich kunnen aandienen.

6.1 Uitdaging 1: Voldoende taken voorzien

Een toets waarbij gebruik wordt gemaakt van ‘performance assessment’, moet een voldoende groot aantal verschillende taken omvatten om tot betrouwbare en valide conclusies te leiden. Dan pas kunnen scores gegeneraliseerd worden overheen verschillende taken uit de ontwikkelde pool van taken (zie ook 3.1.) in plaats van louter toegeschreven worden aan een enkele uitgevoerde taak.

Aan de basis van de noodzaak om meerdere taken in te zetten liggen enerzijds de tussen-takenvariabiliteit en anderzijds het brede domein dat een competentietoets vaak moet bestrijken. De tussen-takenvariabiliteit (‘task sampling variability’) houdt in dat prestaties van leerlingen substantieel variëren tussen taken, omwille van de unieke kenmerken van de taak en de interactie van deze kenmerken met de kennis en ervaring van de leerling (National Research Council (2014)). Om een hypothetisch voorbeeld te geven: het geven van een presentatie en de vaardigheid om dit te doen, kan sterk variëren naargelang het onderwerp waarover het moet gaan. Of het onderwerp aansluit bij de persoonlijke levenssfeer of interesse van de leerling kan de uiteindelijke prestatie sterk beïnvloeden. Uit de geraadpleegde literatuur blijkt dat de tussen-takenvariabiliteit een van de facetten is die het meeste bijdraagt tot de (Shavelson 2010). Tussen-takenvariabiliteit is een belangrijkere bron van meetfouten dan tussen-beoordelaarsvariabiliteit. Of anders gesteld: het aantal taken heeft een groter effect op de generaliseerbaarheid van scores dan het aantal beoordelaars (Brennan and Johnson 1995). De mogelijkheid tot generaliseren van scores wordt bijgevolg groter naarmate er meer performance taken worden voorzien. De nood aan voldoende taken wordt ook ingegeven door het feit dat het domein dat toetsen met het oog op systeemmonitoring dienen te bestrijken, doorgaans veel breder is dan dat van toetsen binnen de klascontext (National Research Council 2014). Het domein ‘informatieverwerving en –verwerking’ bijvoorbeeld is erg ruim en omvat zowel het zelfstandig en op systematische wijze gebruiken van verschillende informatiebronnen, als het systematisch verwerven en gebruiken van samenhangende informatie (ook andere dan teksten). Een toets samenstellen die representatief zou zijn voor dit beoogde domein, kan doorgaans alleen indien er voldoende taken worden voorzien. Immers, hoe minder taken, hoe minder mogelijkheden er zijn om het beoogde domein volledig te bestrijken.

Praktisch is het afnemen van meerdere ‘performance assessment’-taken echter een grote uitdaging door de vereiste inzet van middelen die dit met zich meebrengt. Een ‘performance

assessment'-taak is vaak complex, wat maakt dat de afname ingewikkelder is en het ook langer duurt om de taak af te ronden. Indien men ernaar streeft om de toetsduur niet al te sterk te verlengen, betekent dit dat er een grens is aan hoeveel 'performance assessment'-taken men een leerling kan voorleggen. Het terugvallen op een beperktere set 'performance assessment'-taken verhoogt echter weer het risico op een grote(re) meetfout en heeft een negatieve impact op de mogelijkheid tot generaliseren:

Many authors have observed that limited sampling of relevant performances from a target domain, owing to issues of practicality, safety and fairness as well as the complexity and/or length of the performance tasks, poses the main challenge for the validity of performance assessment in particular. (Curcin et al. 2014, 40)

Uit onze analyse van de praktijkvoorbeelden blijkt dat men doorgaans enkele honderden items en/of taken nodig heeft om een voldoende dekking van het brede, complexe toetsraamwerk te verzekeren. Bij een authentieke, thematische insteek is bovendien sowieso een ruim aantal taken nodig omdat gebruik gemaakt wordt van realistische scenario's. De set taken moet daarbij een verhaal helpen neerzetten dat betekenisvol is voor de leerlingen (zie 5.2.4 voor een illustratie vanuit een praktijkvoorbeeld).

Samenvattend kunnen we stellen dat toetsen met het oog op systeemmonitoring vaak een breed domein moeten bestrijken. Dit leidt er, in combinatie met het fenomeen van de tussen-taken-variabiliteit, toe dat deze toetsen een aanzienlijk aantal taken dienen te omvatten om valide en betrouwbare toetsscores op te leveren. Dit is echter praktisch vaak niet haalbaar in termen van kosten voor de ontwikkeling van de toets, de tijd die leerlingen moeten spenderen aan de toets, en de tijd die gaat kruipen in het scoren van de performances. Haalbare en kwaliteitsvolle werkwijzen die op deze problematiek inspelen, die in de praktijkvoorbeelden en de wetenschappelijke literatuur aan bod kwamen, zijn (zie 5.2.3.):

- het inperken 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.

6.2 Uitdaging 2: Standaardisering van toetsafname en scoren

Scores van toetsen zijn onderhevig aan meetfouten. Toevallige meetfouten kunnen enerzijds onder controle gehouden worden door de steekproef te vergroten (zie 6.1.), anderzijds door de meetprocedure te standaardiseren. Wat dit tweede aspect betreft, is het met het oog op de vergelijkbaarheid van de interpretaties die aan de scores gehecht worden, noodzakelijk dat dezelfde gedetailleerde procedures gevolgd worden op vlak van richtlijnen, omstandigheden van de toets en scoren (AERA APA & NCME 2014).

Standaardisering is een kwestie die bij alle (grootschalige) toetsen aan de orde is, ook toetsen die geen gebruik maken van 'performance assessment'. Eigen aan 'performance assessment' is echter dat het grote(re) risico's op variabiliteit ten gevolge van de toetsafname en het scoren van de 'performance assessment'-taken in zich draagt. Op het

vlak van toetsafname heeft dit bijvoorbeeld te maken met de grotere complexiteit van de taken in vergelijking met een toets die bestaat uit meerkeuzevragen en het risico dat leerlingen op de ene locatie meer begeleiding krijgen bij het oplossen van de taak dan elders. In dat geval heeft standaardisering (of het gebrek eraan) dus gevolgen op vlak van vergelijkbaarheid. Nog een uitdaging komt voort uit het feit dat ‘performance assessment’ vaak bestaat uit open opdrachten en dat de producten die uit deze opdracht ontstaan zo uiteenlopend zijn dat het het scoringsproces bemoeilijkt en de resultaten niet of minder vergelijkbaar zijn (Stecher 2015). Omwille van de complexiteit van de taken is het daarenboven in het geval van ‘performance assessment’ moeilijker om voldoende consistent en accuraat te scoren (Johnson, Penny, and Gordon 2009; Shavelson 2010). Naarmate beoordelaars de beoordelingscriteria verschillend toepassen wordt er ‘judgement uncertainty’ geïntroduceerd (National Research Council 2014). Beoordelaars hebben de neiging doorheen de tijd minder consistent te gaan beoordelen (‘rater drift’) (Shavelson 2010). Ook hier kan standaardisering een oplossing bieden.

Net omwille van de aard van ‘performance assessment’, is het standaardiseren van de toetsafname en het proces van scoren en beoordelen dus niet zomaar eenvoudig geklaard. Daar komt bovenop dat aangereikte oplossingen om de vergelijkbaarheid van scores te garanderen ook haalbaar moeten zijn; een kwestie die zeker bij grootschalige toetsen opspeelt. Kwaliteitsvolle werkwijzen om de toetsafname te standaardiseren, die in de praktijkvoorbeelden aan bod kwamen (zie ook 5.2.4.) zijn: - het lokaal inzetten van centraal getrainde toetsassistenten (duur en logistiek vaak omslachtig) of van lokale leerkrachten (in combinatie met centraal aangestuurde controle en kwaliteitszorg); - het terugvallen op digitale systemen die de omgeving waarin leerlingen hun toets afleggen duidelijk af te bakenen en tegelijkertijd een rijkere en meer authentieke context bieden.

Met betrekking tot het beperken van het risico op beoordelaarseffecten reiken empirische studies o.a. volgende oplossingen aan (zie ook 5.2.5.):

- het voorzien van een degelijke training aan de beoordelaars;
- het inzetten van verschillende beoordelaars;
- het zodanig ontwerpen van taken, o.a. op grond van een ‘evidence centered design’ (5.2.3.), dat ze consistent gescoord kunnen worden;
- het zodanig ontwerpen van scoringstools (analytische, dan wel holistische) dat ze dit proces ondersteunen.

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 (Lesterhuis et al. 2015, 2017; van Daal et al. 2019). 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.

6.3 Uitdaging 3: Vermijden van construct-irrelevante variantie

Construct-irrelevante variantie (CIV) treedt op als naast het construct dat men beoogt te meten, nog één of meerdere andere constructen worden gemeten (S. Messick 1989; Samuel Messick 1994). Als een leerling er bijvoorbeeld niet in slaagt om een bepaalde wiskundetaak op te lossen, kan dit het resultaat zijn van het feit dat de taak ook de competentie 'begrijpend lezen' meet. Het zorgt ervoor dat er systematische ruis in de scores van de toets wordt geïntroduceerd.

Construct-irrelevante variantie dient in alle soorten toetsen vermeden te worden, maar door de specifieke kenmerken van 'performance assessment' is het risico op construct-irrelevante variantie groter. Zo kan bepaalde voorkennis van leerlingen, omwille van de complexiteit van de taken en het gebruik van hulpmiddelen zoals bijvoorbeeld een computer, een belangrijke bron van CIV zijn. Ook de inzet van beoordelaars in het scoringsproces doet het risico stijgen dat er systematisch aandacht uitgaat naar irrelevante kenmerken van prestaties van leerlingen (S. Lane 2015). Beoordelaarseffecten kunnen het resultaat zijn van toevalsfouten (bijv. in het geval de beoordelaar een 'slechte dag' heeft), maar kunnen ook een systematische oorzaak hebben, bijvoorbeeld wanneer beoordelaars systematisch milder zijn in hun beoordelingen. In het verleden werd bijvoorbeeld al vastgesteld dat handgeschreven schrijftaken hogere scores krijgen dan schrijftaken die met een woordprocessor zijn afgewerkt (Powers et al. 1994). Kane (2006) wijst erop dat de keuze voor een welbepaalde toetsvorm, of het nu een set meerkeuzevragen of een 'performance assessment'-taak is, ook een bron van CIV kan zijn, omdat bepaalde groepen beter presteren op bepaalde toetsvormen. In de literatuur stelden we vast dat motivatie van leerlingen een belangrijke bron van CIV is, die specifiek opspeelt in 'low-stakes'-toetsen, zoals bijvoorbeeld toetsen met het oog op kwaliteitsmonitoring op systeemniveau. Bij toetsen waar voor de leerling in kwestie weinig op het spel staat ('low stakes') zijn leerlingen vaak minder gemotiveerd, waardoor geen juist beeld gevormd kan worden van het reële prestatieniveau. Deze problematiek rond lage motivatie speelt sterker bij 'performance assessment' dan bij klassieke toetsen samengesteld uit meerkeuzevragen, zo stellen Suzanne Lane and Stone (2006). Hoewel de literatuur ons op het spoor bracht van deze bron van CIV, bleek dit als dusdanig niet op te duiken in onze selectie praktijkvoorbeelden.

We hebben vastgesteld dat de praktijkvoorbeelden erg verschillend omspringen met construct-irrelevante variantie. 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 en laat men authenticiteit primeren. Concreet wordt construct-irrelevante variantie o.m. tegengegaan door (zie ook 5.2.3., voorwaarde 10):

- het uitgebreid testen van de taken, o.m. vanuit een evidence-centered design (5.2.3.); en
- het inzetten van verschillende meetmethoden (o.a. S. Messick (1989); zie ook 5.2.3.).

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

Hoe sterker de opdrachten in de toets lijken op taken die voorkomen in de reële situaties waarin men de te toetsen competentie moet inzetten (i.e. de criteriumsituatie of criteriumtaken), hoe beter de toetsscores de prestatie in het competentiedomein voorspellen (Straetmans 2014). De mate waarin toetstaken lijken op taken in de criteriumsituatie kan zich op verschillende manieren veruitwendigen, onder meer via de inhoud en vorm van de taak of via de fysieke omgeving en sociale context waarin de taak wordt uitgevoerd (Gulikers and Benthum 2017).

‘Performance assessments’ hebben het potentieel om, via authentieke taken, complexe vaardigheden en competenties te meten. 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. In realiteit is het vaak zo dat complexe taken onderverdeeld worden in verschillende componenten en voor elk van deze componenten vervolgens een aparte toets wordt uitgewerkt. Aan het einde worden de scores opgeteld en deze finale score representeert dan de in kaart gebrachte ‘performance’. Zoals Pecheone and Kahl (2015) aangeven, is dit de praktijk die in veel toetsen waar standaardisering om de hoek komt kijken, wordt gevolgd. De auteurs pleiten voor een andere, meer geïntegreerde aanpak die zij ‘criterion sampling’ noemen. Het begrip ‘criterion sampling’ is op zich eenvoudig:

“(...) if you want to know what a person knows and can do, sample tasks from the domain in which she is to act, observe her performance and infer competence and learning” (Pecheone and Kahl 2015, 72)

Verduidelijkend: deze aanpak veronderstelt dat het geheel meer is dan de optelsom van de onderdelen en dat complexe taken een integratie van bekwaamheden vereisen die niet gevat kunnen worden als ze verdeeld en gemeten worden als aparte componenten. Dat authentieke taken, of taken die recht doen aan de criteriumsituatie, vatbaar zijn voor construct-irrelevante variantie stipten we hierboven reeds aan (zie uitdaging 3). Een andere moeilijkheid is dat het opzetten van taken die recht doen aan de criteriumsituatie vaak in conflict komt met de noodzaak om te standaardiseren (zie uitdaging 2). Wanneer men in de toetsprocedure bijvoorbeeld aspecten gaat standaardiseren die niet vastgelegd zijn in de criteriumsituatie, vormt dit een bron van systematische ruis. Het gevolg is dat de resultaten niet geëxtrapoleerd kunnen worden naar het volledige competentiedomein (M. T. Kane 2013), met andere woorden: de taken doen geen recht doen de criteriumsituatie. We moeten dus steeds waakzaam zijn voor een (te) ver doorgedreven standaardisering. Standaardisering en authenticiteit moeten steeds onderling afgewogen. Met betrekking tot deze afweging lijkt de oplossing er op 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 geven.

Om representatief te zijn voor het beoogde competentiedomein is het belangrijk dat de omstandigheden van de observatie representatief zijn voor deze in het beoogde domein (Kane, Crooks, and Cohen 1999). Het meenemen van de criteriumsituatie in het opzetten van de taak, betekent dus dat zowel product als proces in kaart worden gebracht. Bij de opzet van een toets schrijfvaardigheden, bijvoorbeeld, houdt dit in dat er ook ruimte moet zijn voor aspecten als voorafgaande studie van de literatuur, planning en revisie achteraf; procesgerelateerde elementen dus. Powers and Fowles (1998) wijzen er met betrekking tot een toets schrijfvaardigheid echter op dat leerlingen bij een schrijftaak vaak enkel tijd hebben om een eerste ontwerp uit te schrijven, niet voor een uitvoerige planning en volgende fasen van revisie en herwerking. De taken belichten met andere woorden onvoldoende de vele processen die schrijvers gebruiken en representeren dus niet volledig de beoogde competentie (de schrijfvaardigheid in de criteriumsituatie). Ook in de praktijkvoorbeelden die we analyseerden worden procescomponenten momenteel nog in zeer beperkte mate meegenomen. Bij computergebaseerde toetsen lijkt het inzetten van 'tracking software' een beloftevolle piste om zicht te krijgen op het proces. Tracking is echter een middelenintensief proces, dat niet steeds die bepaalde gegevens oplevert 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. Zoals S. Lane (2015) onderstreept is het immers niet omdat 'performance assessment' bijzonder geschikt is voor het meten van complexe constructen, dat elke ontwikkelde toetsvorm die 'performance assessment' omvat dit ook werkelijk doet: bewijsmateriaal is nodig om te illustreren dat de taken en rubrics werkelijk gericht zijn op het meten van dit beoogde construct. Haalbare en kwaliteitsvolle werkwijzen voor het vergaren van evidentie tijdens de pilootfase die in de praktijkvoorbeelden en de wetenschappelijke literatuur aan bod kwamen, zijn bijvoorbeeld (zie ook 5.2.7.) :

- het gebruik maken van cognitieve interviews;
- het uitvoeren van een piloottest om vervolgens statistisch na te gaan of de taken voldoen.

6.5 Uitdaging 5: Conform de doelstellingen rapporteren

De redenen waarom een toets ontwikkeld wordt, vormen ook het raamwerk waarbinnen wordt gerapporteerd (Cohen and Wollack 2006): resultaten dienen te worden gecommuniceerd in een vorm die overeenstemt met het doel van de test. Zo is het bijvoorbeeld mogelijk om resultaten op verschillende agregatieniveaus te rapporteren: individuele leerlingen, deelnemende scholen en/of systeemniveau. Tegenover de vereisten die voortvloeien uit de doelstellingen, komt ook het aspect haalbaarheid te staan, en wel op twee manieren. Enerzijds dient het rapport klaargestoomd te worden binnen een bepaalde termijn, opdat de opdrachtgevers van de toets ook tijdig aan de slag kunnen gaan met de informatie waarover wordt gerapporteerd (Cohen and Wollack 2006). Anderzijds wordt de manier waarop men kan rapporteren ook begrensd door de kwaliteit van de toetsscores. Zo is het bijvoorbeeld niet mogelijk om betrouwbare feedback op schoolniveau te genereren

indien men daar bij het vastleggen van de doelstellingen en bij de toetsopzet, (m.n. bij het bepalen van de omvang van de steekproef leerlingen) geen rekening mee hield.

Ruwe scores worden getransformeerd en onder de vorm van geschaalde scores gerapporteerd (Tan and Michel 2011). Onder 'schalen' verstaan we *"the process of associating numbers or other ordered indicators with the performance of examinees"* (Kolen and Brennan 2014, 329). Een schaal wordt initieel meestal ontwikkeld voor één toets. Indien men een schaal opnieuw wil gebruiken voor de afname van een andere toets, dient men over te gaan tot equivalering: een statistisch proces dat gebruikt wordt om scores op twee of meer toetsen aan te passen, zodat de scores onderling inwisselbaar worden (Kolen and Brennan 2014), zelfs indien de toetsen (deels) uit verschillende items en/of taken bestaan. Op die manier kunnen de ruwe scores van opeenvolgende toetsafnames op de ontwikkelde scoreschaal worden geplaatst (Kolen and Brennan 2014). In de literatuur vinden we tal van equivaleringsmodellen en -procedures terug (o.a. Kolen and Brennan 2014; Holland and DePascale 2006).

Bij de analyse van de praktijkvoorbeelden viel op dat men zeer frequent item respons theorie (IRT) gebruikt, zowel om te schalen als met het oog op equivalering. IRT-modellen zijn statistische modellen die gebruikt kunnen worden om de 'performance' op een toets te schatten, waarbij gebruik wordt gemaakt van karakteristieken, van zowel personen als items, waarop de performance verondersteld gebaseerd te zijn (Suzanne Lane and Stone 2006). De focus op IRT is in de bestudeerde praktijkvoorbeelden soms zelfs bepalend voor het toetsdesign.

Davey et al. (2015) verwijzen naar een paper van Gorin and Mislevy (2013) waarin de auteurs twee centrale, psychometrische uitdagingen ten aanzien van het gebruik van IRT samenvatten. Een eerste uitdaging houdt verband met de gewenste lokale onafhankelijkheid ('local independence') van items en/of taken, wat impliceert dat toetsvragen/activiteiten idealiter niet met elkaar in verband mogen staan. Typisch voor 'performance assessment' is echter dat het taaktypes inschakelt, waarvan de toetsvragen/activiteiten net verband houden met elkaar, om op die manier voor de leerling een betekenisvol geheel te kunnen vormen. De tweede centrale psychometrische knoop die met het oog op het inzetten van IRT ontward dient te worden, houdt verband met de assumptie van 'unidimensionaliteit'. Dit betekent dat het psychometrische model best werkt wanneer een toets slechts één construct meet. 'performance assessment' houdt ook op dat punt bepaalde risico's in, in die zin dat het ingezet wordt om bredere constructen te meten waarbinnen veel verschillende elementen onderscheiden kunnen worden. Naast deze beide centrale psychometrische uitdagingen, is het feit dat de meeste equivaleringsdesigns steunen op het hergebruik van minstens een aantal van de gebruikte taken ('ankertaken'), eveneens problematisch voor de toepassing van IRT op toetsen die een 'performance assessment'-component bevatten. 'performance assessment'-taken zijn immers vaak makkelijk te memoriseren door leerlingen en kunnen daarom moeilijker gebruikt worden als link tussen verschillende afnames van een toets. In het geval de voorwaarden van IRT geschonden (zullen) worden, stelden we in de praktijkvoorbeelden overigens vast dat men zich beperkt tot het rapporteren van beschrijvende resultaten. Gegeven dat IRT, specifiek voor 'performance assessment', enkele psychometrische uitdagingen met zich meebrengt, kan men overwegen om (bijkomend) andere technieken in

te zetten, zoals bijvoorbeeld comparatieve beoordeling of paarsgewijze vergelijking (Heldsinger and Humphry 2010; S. Heldsinger and Humphry 2013; Lesterhuis et al. 2017).

Om een valide interpretatie (en gebruik) van scores te ondersteunen, moeten er ook beslissingen genomen worden over hoe de geschaalde scores ook betekenisvol gemaakt kunnen worden. De literatuur onderscheidt hiertoe drie soorten procedures: 'item mapping', 'scale anchoring' en 'standard setting' (Kolen 2006; Mazzeo and Zieky 2006) (zie 5.2.7.). Het gebruik van prestatiestandaarden is wat dit betreft een interessante manier, met name om te kunnen inschatten welk aandeel van de leerlingenpopulatie een bepaalde cesuur of minimumstandaard haalt. De nood aan nieuwe, empirisch onderbouwde methoden voor het vastleggen van prestatiestandaarden met betrekking tot 'performance assessment' wordt al lang signaleerd. Onder andere uit de literatuurstudie valt echter 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 werden.

7 Implicaties

Uit het onderzoek dat we voerden, kunnen lessen getrokken worden voor het beleid en de praktijk. In dit laatste deel formuleren we bijgevolg acht aanbevelingen met het oog op de ontwikkeling en evaluatie van (toekomstige) grootschalige 'performance assessments'.

1. 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.

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.

Het inzetten van 'performance assessment' bij het grootschalig toetsen van competenties dient echter weloverwogen te gebeuren. De evaluatiematrix die we 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 'performance assessment'-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 tijdsperiode te toetsen.

2. Reserveer als opdrachtgever van grootschalige competentietoetsen met een 'performance assessment'-component, gericht op monitoring op systeemniveau, voldoende tijd en middelen voor een heldere en volledige doelbepaling.

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 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.

3. Wees bij grootschalige competentietoetsen met een 'performance assessment'-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.

Uit de analyse van de praktijkvoorbeelden leren we dat toetsprogramma's en toetsen multiële 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.

Waakzaamheid is echter geboden 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.

4. Maak gebruik van de bouwstenen en voorwaarden geïdentificeerd in de evaluatiematrix om te bepalen of grootschalige competentietoetsen met een 'performance assessment'-component, gericht op monitoring op systeemniveau, kwaliteitsvol zijn.

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 'performance assessment'-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.

5. Maak bij het realiseren van grootschalige competentietoetsen met een 'performance assessment'-component, gericht op monitoring op systeemniveau, een weloverwogen afweging tussen generaliseerbaarheid van scores, extrapoleerbaarheid van scores en haalbaarheid (in termen van tijd en middelen).

Door in te spelen op de voorwaarden uit de evaluatiematrix kan in principe een kwaliteitsvolle toets worden uitgewerkt. Voor het realiseren van een kwaliteitsvolle toets, moeten echter keuzes gemaakt 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.

6. Sta bij het opzetten van grootschalige competentietoetsen met een 'performance assessment'-component, gericht op monitoring op systeemniveau, open voor andere opties dan strikt gestandaardiseerde toetssystemen.

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 'performance assessments' in principe aanspraak kunnen maken. Dit doet de vraag rijzen welke mate van standaardisering in feite wenselijk en noodzakelijk is.

Er is echter ook een alternatieve piste mogelijk, 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.

Deze 'alternatieve' werkwijze heeft zowel voor- als tegenstanders. 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.

7. Overweeg bij grootschalige competentietoetsen met een 'performance assessment'-component, gericht op monitoring op systeemniveau, om 'matrix sampling' te gebruiken.

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 hoofdstuk 6 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.

8. Blijf oog hebben voor nieuwe ontwikkelingen in onderzoek naar en de praktijk van grootschalige competentietoetsen met een 'performance assessment'-component, gericht op monitoring op systeemniveau.

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

Paarsgewijze vergelijking lijkt een valide, betrouwbaar en haalbaar alternatief te zijn voor scores van 'performance assessment'-taken via specifieke scoringstools, zeker in combinatie met nieuwe technologische mogelijkheden. Daarnaast doet geautomatiseerd scores 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 scores, 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 hier de vinger aan de pols te houden. Tijdens het onderzoek dat aan de basis van deze publicatie lag viel het ook op dat vele van de buitenlandse praktijkvoorbeelden bereid zijn om inzichten

en ideeën te delen en dat 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.

Woordenlijst

<i>Afgeleide score</i>	Afgeleide scores worden uit ruwe scores getransformeerd. Ruwe scores geven enkel aan hoeveel vragen/taken een leerling correct heeft opgelost en zijn daarom niet geschikt als basis voor vergelijkingen. Afgeleide scores, zoals bijvoorbeeld percentielrankings en schaalscores, laten daarentegen toe om vergelijkingen te maken tussen toetsscores.
<i>Betrouwbaarheid</i>	Verwijst naar de consistentie van scores over replicaties van een toets of van beoordelingen heen. De aard en kwaliteit van de respons van een leerling op een toets kunnen variëren van de ene steekproef van taken naar de andere, of van het ene moment van toetsafname naar het andere, zelfs onder gecontroleerde omstandigheden. Verschillende beoordelaars kunnen bovendien andere scores toekennen aan dezelfde prestatie.
<i>Competentie</i>	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.
<i>Construct</i>	Het theoretische concept dat men door middel van de toets wenst te meten.
<i>Construct-irrelevante variantie</i>	Verwijst naar variantie in een score die resulteert uit iets anders (één of meerdere irrelevante constructen) dan het construct dat men beoogde te meten en zorgt ervoor dat systematische ruis in de toetsscores wordt geïntroduceerd.
<i>Construct-Onderrepresentatie</i>	Dit houdt in dat de toets belangrijke aspecten (inhoud en/of processen) van het beoogde construct niet vat. Het gevolg is dat de betekenis die aan de toetsscores gehecht kan worden, verengd wordt.
<i>Criteriumgerefereerde toets</i>	Is een toets waarin inhoudsstandaarden - datgene wat leerlingen moeten kennen en kunnen - de maatstaf vormen voor een het al of niet behalen van een bepaalde prestatie- of competentieniveau (in tegenstelling tot normgerefereerde beoordelingen waarbij een vooraf vastgelegde slaagratio de maatstaf vormt).
<i>Criteriumsituatie</i>	Is de reële context waarin de competentie die men beoogt te meten, vorm krijgt. Criteriumtaken zijn de taken die in die reële context worden uitgevoerd.
<i>Extrapoleren</i>	De mogelijkheid tot het extrapoleren van scores impliceert dat de prestaties op de toetstaken een goede indicator zijn van prestaties op criteriumtaken uit de alledaagse context.
<i>Generaliseren</i>	De mogelijkheid tot het generaliseren van toetsscores houdt in dat de betekenis van een specifieke toetsscore zich uitstrekt overheen replicaties (bv. naar taak, beoordelaar en/of afnamemoment), getrokken uit het toetsdomein. Algemeen geldt dat naarmate het

	aantal onafhankelijke observaties met betrekking tot elk van deze facetten (d.i. de steekproefomvang) toeneemt en naarmate de meetprocedure gestandaardiseerd verloopt, de generaliseerbaarheid toeneemt.
<i>IRT (item response theorie)</i>	Is een statistische theorie die gebruikt maakt van modellen om de prestatie op een toets te schatten. Dit gebeurt op basis van karakteristieken van zowel personen als items, waarop de performance verondersteld is gebaseerd te zijn.
<i>Kwaliteit</i>	Vatten we op 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.
<i>'Low stakes'-toetsen</i>	Zijn toetsen met een lage inzet (bv. voor de leerlingen of voor de school). Toetsen met het oog op kwaliteitsmonitoring op systeemniveau, die alleen geaggregeerde resultaten en dus geen informatie op individueel leerling- of schoolniveau opleveren, zijn hier een voorbeeld van. Hiertegenover staan 'high stakes'-toetsen, waarbij de inzet net hoog is, zoals bijvoorbeeld een toets die bepaalt of een leerling slaagt of niet in een bepaald leerjaar.
<i>Monitoring op systeemniveau</i>	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.
<i>Performance assessment</i>	Betreft beoordeling (van competenties) waarbij gebruik wordt gemaakt van (levensechte) taken, relevant voor de beoogde competenties.
<i>Recalibratiesets</i>	Zijn vooraf gescoorde antwoorden die in een opfrissingssessie gebruikt worden om de beoordelaars de standaarden opnieuw in herinnering te brengen.
<i>Standaardisering</i>	Het zorgen voor uniformisering van afname- en scoringsprocedures met het oog op vergelijkbaarheid van toetsscores overheen contexten.
<i>Steekproefvariabiliteit</i>	Verwijst naar veranderlijkheid in de toetsscores, veroorzaakt door variaties in taken, beoordelaars en/of afnamemomenten; met andere woorden naar de mate waarin de toetsscore varieert van steekproef tot steekproef (i.c. van taken, beoordelaars en/of afnamemomenten). Hoe groter de steekproefvariabiliteit, hoe groter de meetfout.

Toevallige ruis

Toevallige ruis of toevallige meetfout wordt veroorzaakt door factoren die de toetsscores op toevallige wijze beïnvloeden. Toevallige ruis heeft geen systematisch effect op de toetsscores op de volledige steekproef. Bij herhaalde metingen worden de toetsscores op toevallige wijze nu eens de hoogte, dan weer de laagte ingestuurd. Er zit dus een zekere spreiding op. Hiertegenover staat systematische ruis (of systematische meetfout), die de toetsscores systematisch de hoogte ofwel de laagte in stuurt. Bij herhaling van de meting (met dezelfde toets) zal dezelfde afwijking geconstateerd worden.

Validiteit

Verwijst hier specifiek naar de mogelijkheid om scores op een toets te generaliseren naar het toetsdomein en vervolgens te extrapoleren naar het beoogde competentiedomein.

Referenties

ACARA - Australian Curriculum, Assessment and Reporting Authority. 2014. "NAPLAN Achievement in Reading, Persuasive Writing, Language Conventions and Numeracy: National Report for 2014." Sydney: ACARA.

AERA APA & NCME. 2014. *Standards for Educational and Psychological Testing*. Washington D.C.: American Educational Research Association.

<https://blackwells.co.uk/bookshop/product/Standards-for-Educational-and-Psychological-Testing-by-American-Educational-Research-Association-American-Psychological-Association-National-Council-on-Measurement-in-Education-Joint-Committee-on-Standards-for-Educational-and-Psychological-Testing-U-S-/9780935302356>.

Baartman, Liesbeth. 2008. "Assessing the Assessment: Development and Use of Quality Criteria for Competence Assessment Programmes." Doctoral Thesis, Utrecht University.

Baartman, Liesbeth, Theo Bastiaens, Paul Kirschner, and Cees van der Vleuten. 2006. "The Wheel of Competency Assessment: Presenting Quality Criteria for Competency Assessment Programs." *Studies in Educational Evaluation* 32 (2): 153–70.
<https://doi.org/10.1016/j.stueduc.2006.04.006>.

Baartman, Bastiaens, Kirschner, and van der Vleuten. 2007. "Evaluating Assessment Quality in Competence-Based Education: A Qualitative Comparison of Two Frameworks." *Educational Research Review* 2 (2): 114–29. <https://doi.org/10.1016/j.edurev.2007.06.001>.

Basturk, Ramazan. 2008. "Applying the Many-facet Rasch Model to Evaluate PowerPoint Presentation Performance in Higher Education." *Assessment & Evaluation in Higher Education* 33 (4): 431–44. <https://doi.org/10.1080/02602930701562775>.

Ben-Simon, Anat, and Randy Elliot Bennett. 2007. "Toward More Substantively Meaningful Automated Essay Scoring." *The Journal of Technology, Learning and Assessment* 6 (1, 1). <https://ejournals.bc.edu/index.php/jtla/article/view/1631>.

Biggs, John. 1996. "Enhancing Teaching Through Constructive Alignment." *Higher Education* 32 (3): 347–64. <https://doi.org/10.1007/BF00138871>.

Biggs, John B., and Catherine So-kum Tang. 2011. *Teaching for Quality Learning at University: What the Student Does*. 4th edition. SRHE and Open University Press Imprint. Maidenhead, England New York, NY: McGraw-Hill, Society for Research into Higher Education & Open University Press.

Brennan, Robert L., ed. 2006. *Educational Measurement*. 4. ed. Series on Higher Education. New York: American Council on Education [u.a.].

Brennan, Robert L., and Eugene G. Johnson. 1995. "Generalizability of Performance Assessments." *Educational Measurement: Issues and Practice* 14 (4): 9–12.
<https://doi.org/10.1111/j.1745-3992.1995.tb00882.x>.

Chapelle, Carol A. 2012. "Validity Argument for Language Assessment: The Framework Is Simple...." *Language Testing* 29 (1): 19–27. <https://doi.org/10.1177/0265532211417211>.

Chapelle, Carol A., Mary K. Enright, and Joan Jamieson. 2010. "Does an Argument-Based Approach to Validity Make a Difference?" *Educational Measurement: Issues and Practice* 29 (1): 3–13. <https://doi.org/10.1111/j.1745-3992.2009.00165.x>.

Childs, Ruth, and Andrew Jaciw. 2019. "Matrix Sampling of Items in Large-Scale Assessments." *Practical Assessment, Research, and Evaluation* 8 (1). <https://doi.org/10.7275/gwvh-4z51>.

Cohen, Allan, and James Wollack. 2006. "Test Administration, Security, Scoring, and Reporting." In *Educational Measurement*, edited by Robert L. Brennan, 4th ed., 355–86. American Council on Education/Praeger.

Cronbach, Lee. 1971. "Test Validation." In *Educational Measurement*, edited by L. Thorndike, 2nd ed., 443–507. Washington D.C.: American Council on Education/Praeger.

Cronbach, Lee J., and Goldine C Gleser. 1965. *Psychological Tests and Personnel Decisions*. Urbana: University of Illinois Press.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52 (4): 281–302. <https://doi.org/10.1037/h0040957>.

Crooks, Terry J., Michael T. Kane, and Allan S. Cohen. 1996. "Threats to the Valid Use of Assessments." *Assessment in Education: Principles, Policy & Practice* 3 (3): 265–86. <https://doi.org/10.1080/0969594960030302>.

Curcin, Milja, Andrew Boyle, Tom May, and Zeeshan Rahman. 2014. "A Validation Framework for Work-Based Observational Assessment in Vocational Qualifications." Coventry: Office of Qualifications and Examinations Regulation.

Darling-Hammond, Linda, and Frank Adamson. 2014. *Beyond the Bubble Test: How Performance Assessments Support 21st Century Learning*. First edition. San Francisco, CA: Jossey-Bass & Pfeiffer Imprints, Wiley.

Davey, Tim, Steve Ferrara, P. W. Holland, Rich Shavelson, Noreen M. Webb, and Laurens L. Wise. 2015. "Psychometric Considerations for the Next Generation of Performance Assessment. Princeton." Educational Testing Service.

De Maeyer, Sven, Vincent Donche, Jan Vanhoof, Peter Van Petegem, Liesje Coertjens, Jetje De Groof, and Alexia Deneire. 2016. "Hoe Zijn Competenties Grootschalig Te Toetsen? Ontwikkeling van Een Evaluatiematrix Voor Toetsprogramma's En Een Inventarisatie van 'Good Practices.'" Eindrapport. Departement Onderwijs.

Der Vleuten, Cees P M van, and Lambert W T Schuwirth. 2005. "Assessing Professional Competence: From Methods to Programmes." *Medical Education* 39 (3): 309–17. <https://doi.org/10.1111/j.1365-2929.2005.02094.x>.

Dienst Beroepsopleiding. 2008. "Competentieleren: Een Gedachte-Experiment: Rapport." Brussel: Dienst Beroepsopleiding, Departement Onderwijs en Vorming.

Educational Assessment Research Unit & NZCER - New Zealand Council for Educational Research, EARU -. 2014. "National Monitoring Study of Student Achievement (Wanangatia Te Putanga Tauira) - Health and Physical Education 2013." New Zealand: Ministry of Education.

Eisner, Elliot W. 1999. "The Uses and Limits of Performance Assessment." *The Phi Delta Kappan* 80 (9): 658–60. <https://www.jstor.org/stable/20439532>.

Engelhard, George Jr. 2002. "Monitoring Raters in Performance Assessments." In *Large-Scale Assessment Programs for All Students*, edited by Gerald Tindal and Thomas M. Haladyna. Routledge.

Figel, J. 2007. "Key Competences for Lifelong Learning-European Reference Framework." Luxembourg: Office for Official Publications of the European Communities.

Fitzpatrick, R., and E. Morrison. 1971. "Performance and Product Evaluation." In *Educational Measurement*, edited by L. Thorndike, 2nd ed., 443–507. Washington D.C.: American Council on Education/Praeger.

Gorin, Joanna S, and Robert J Mislevy. 2013. "Inherent Measurement Challenges in the Next Generation Science Standards for Both Formative and Summative Assessment." New Jersey: Educational Testing Service. <https://www.ets.org/Media/Research/pdf/gorin-mislevy.pdf>.

Gulikers, Judith, and Niek van Benthum. 2017. "Toetsen van competenties." In *Toetsen in het hoger onderwijs*, edited by Henk van Berkel, Anneke Bax, and Desirée Joosten-ten Brinke, 227–39. Houten: Bohn Stafleu van Loghum. https://doi.org/10.1007/978-90-368-1679-3_18.

Haertel, E. 2006. "Reliability." In *Educational Measurement*, by Robert L. Brennan, 4th ed. Westport: Praeger Publishers.

Hambleton, Ronald K. 2006. "Setting Performance Standards." In *Educational Measurement*, by B. S. Pitoniak and Robert L. Brennan, 4th ed. Westport: Praeger Publishers.

Hambleton, Ronald K., Richard M. Jaeger, Barbara S. Plake, and Craig Mills. 2000. "Setting Performance Standards on Complex Educational Assessments." *Applied Psychological Measurement* 24 (4): 355–66. <https://doi.org/10.1177/01466210022031804>.

Heldsinger, S., and Humphry. 2013. "Using Calibrated Exemplars in the Teacher-Assessment of Writing: An Empirical Study." *Educational Research* 55 (3): 219–35. <https://doi.org/10.1080/00131881.2013.825159>.

Heldsinger, and Humphry. 2010. "Using the Method of Pairwise Comparison to Obtain Reliable Teacher Assessments." *The Australian Educational Researcher* 37 (2): 1–19. <https://doi.org/10.1007/BF03216919>.

Hill, Richard K., and Charles A. DePascale. 2003. "Reliability of No Child Left Behind Accountability Designs." *Educational Measurement: Issues and Practice* 22 (3): 12–20. <https://doi.org/10.1111/j.1745-3992.2003.tb00133.x>.

- Holland, P. W., and Charles A. DePascale. 2006. "Linking and Equation." In *Educational Measurement*, by Robert L. Brennan, 4th ed., 187–220. Westport: Praeger Publishers.
- Hornsby, D., and M. Wu. 2012. "Misleading Everyone with Statistics." http://sydney.edu.au/education_social_work/news_events/resources/No_NAPLAN.pdf.
- Johnson, Robert L., James A. Penny, and Belita Gordon. 2009. *Assessing Performance: Designing, Scoring, and Validating Performance Tasks*. New York: The Guilford Press.
- Kane. 2006. "Validation." In *Educational Measurement*, by Robert L. Brennan, 4th ed. Westport: Praeger Publishers.
- Kane, M. T. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50 (1): 1–73. <https://doi.org/10.1111/jedm.12000>.
- Kane, Crooks, and Cohen. 1999. "Validating Measures of Performance." *Educational Measurement: Issues and Practice* 18 (2): 5–17. <https://doi.org/10.1111/j.1745-3992.1999.tb00010.x>.
- Kimbell, Richard, Tony Wheeler, Soo Miller, and Alastair Pollitt. 2007. *E-Scape Portfolio Assessment - Phase 2 Report*. London: Goldsmiths.
- Kish, Leslie. 2005. *Statistical Design for Research*. <https://nbn-resolving.org/urn:nbn:de:101:1-20141021261>.
- Kolen. 2006. "Scaling and Norming." In *Educational Measurement*, by Robert L. Brennan, 4th ed. Westport: Praeger Publishers.
- Kolen, and Brennan. 2014. *Test Equating, Scaling, and Linking: Methods and Practices*. 3d edition. Statistics for Social Science and Public Policy. New York: Springer.
- Kuhlemeier, Hans, Bas Hemker, and Huub van den Bergh. 2013. "Impact of Verbal Scale Labels on the Elevation and Spread of Performance Ratings." *Applied Measurement in Education* 26 (1): 16–33. <https://doi.org/10.1080/08957347.2013.739425>.
- Kuhlemeier, Hans, A. van Til, Bas Hemker, W. de Klijn, and H. Feenstra. 2013. "Balans van de Schrijfvaardigheid in Het Basis- En Speciaal Basisonderwijs 2. Uitkomsten van de Peiling in 2009 in Groep 5, Groep 8 En de Eindgroep van Het SBO." 53. PPON-reeks. Arnhem: Cito.
- Lane, S. 2015. "Performance Assessment: The State of the Art." In *Beyond the Bubble Test*, edited by Linda Darling-Hammond and Frank Adamson, 131–84. San Francisco: John Wiley & Sons, Inc. <https://doi.org/10.1002/9781119210863.ch5>.
- Lane, Suzanne. 2010. *Performance Assessment: The State of the Art*. SCOPE Student Performance Assessment Series. Stanford, CA: Stanford University, Stanford Center of Opportunity Policy in Education. https://edpolicy.stanford.edu/sites/default/files/publications/performance-assessment-state-art_1.pdf.

Lane, Suzanne, and C. Stone. 2006. "Performance Assessment." In *Educational Measurement*, edited by Robert L. Brennan, 4th ed., 387–432. American Council on Education/Praeger.

Lesterhuis, Donche, De Maeyer, van Daal, Van Gasse, Coertjens, Verhavert, Mortier, Coenen, and Vlerick. 2015. "Compententies Kwaliteitsvol Beoordelen: Brengt Een Comparatieve Aanpak Soelaas?" *Tijdschrift Voor Hoger Onderwijs* 33 (2): 55–67.

Lesterhuis, Verhavert, Coertjens, Donche, and De Maeyer. 2017. "Comparative Judgement as a Promising Alternative to Score Competences." In *Innovative Practices for Higher Education Assessment and Measurement*, by E. Cano and G. Ion, 119–36. <https://doi.org/10.4018/978-1-5225-0531-0.ch007>.

Linn, Robert, Eva Baker, and Stephen B. Dunbar. 1991. "Complex, Performance-Based Assessment: Expectations and Validation Criteria." *Educational Researcher* 20 (8): 15–21. <https://doi.org/10.3102/0013189X020008015>.

Lissitz, Robert W, and Feifei Li. 2011. "Standard Setting in Complex Performance Assessments: An Approach Aligned with Cognitive Diagnostic Models." *Psychological Test and Assessment Modeling* 53 (4): 461–85.

Lizzio, Alf, and Keithia Wilson. 2004. "Action Learning in Higher Education: An Investigation of Its Potential to Develop Professional Capability." *Studies in Higher Education* 29 (4): 469–88. <https://doi.org/10.1080/0307507042000236371>.

Lu, L. R. 2012. *A Validation Framework for Automated Essay Scoring Systems*. Unpublished Doctoral Dissertation. Australia: Faculty of Education, University of Wollongong.

Mazzeo, J., and M. J. Zieky. 2006. "Monitoring Educational Progress with Group-Score Assessments." In *Educational Measurement*, by Robert L. Brennan, 4th ed., 681–99. Westport: Praeger Publishers.

Messick. 1996. "Validity of Performance Assessments." In *Technical Issues in Large-Scale Performance Assessment*, edited by G. Phillips, 198–258. Washington D.C.: National Center for Education Statistics.

Messick, S. 1989. "Validity." In *Educational Measurement, 3rd Ed*, edited by R. L. Linn, 13–103. The American Council on Education/Macmillan Series on Higher Education. American Council on Education.

Messick, Samuel. 1994. "The Interplay of Evidence and Consequences in the Validation of Performance Assessments." *Educational Researcher* 23 (2): 13–23. <https://doi.org/10.3102/0013189X023002013>.

Moss, Pamela A. 1994. "Can There Be Validity Without Reliability?" *Educational Researcher* 23 (2): 5–12. <https://doi.org/10.3102/0013189X023002005>.

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

- Newhouse, C. Paul. 2011. "Using IT to Assess IT: Towards Greater Authenticity in Summative Performance Assessment." *Computers & Education* 56 (2): 388–402. <https://doi.org/10.1016/j.compedu.2010.08.023>.
- Newhouse, Paul. 2013. "Literature Review and Conceptual Framework." In *Digital Representations of Student Performance for Assessment*, edited by P. John Williams and C. Paul Newhouse, 9–28. Rotterdam: SensePublishers. https://doi.org/10.1007/978-94-6209-341-6_2.
- Pecheone, Raymond, and Stuart Kahl. 2015. "Where We Are Now." In *Beyond the Bubble Test*, 53–91. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119210863.ch3>.
- Powers, Donald E., and Mary E. Fowles. 1998. "Effects of Preexamination Disclosure of Essay Topics." *Applied Measurement in Education* 11 (2): 139–57. https://doi.org/10.1207/s15324818ame1102_2.
- Powers, Donald E., Mary E. Fowles, Marisa Farnum, and Paul Ramsey. 1994. "Will They Think Less of My Handwritten Essay If Others Word Process Theirs? Effects on Essay Scores of Intermingling Handwritten and Word-Processed Essays." *Journal of Educational Measurement* 31 (3): 220–33. <https://www.jstor.org/stable/1435267>.
- Prodromou, Luke. 1995. "The Backwash Effect: From Testing to Teaching." *ELT Journal* 49 (1): 13–25. <https://doi.org/10.1093/elt/49.1.13>.
- Rubin, D. 1996. "A Preface Relating Alternative Assessment, Test Fairness, and Assessment Utility to Communication." In *Large Scale Assessment of Oral Communication: K–12 and Higher Education*, by S. Morreale and P. Backlund, 1–4. Annandale: Speech Communication Association. <https://files.eric.ed.gov/fulltext/ED399578.pdf>.
- Schmeiser, C., and C. Welch. 2006. "Test Development." In *Educational Measurement*, by Robert L. Brennan, 4th ed., 307–54. Westport: Praeger Publishers.
- Shavelson, Richard J. 2010. "On the Measurement of Competency." *Empirical Research in Vocational Education and Training* 2 (1, 1): 41–63. <https://doi.org/10.1007/BF03546488>.
- Shaw, Stuart, Victoria Crisp, and Nat Johnson. 2012. "A Framework for Evidencing Assessment Validity in Large-Scale, High-Stakes International Examinations." *Assessment in Education: Principles, Policy & Practice* 19 (2): 159–76. <https://doi.org/10.1080/0969594X.2011.563356>.
- Sireci, Stephen G. 2009. "Packing and Unpacking Sources of Validity Evidence: History Repeats Itself Again." In *The Concept of Validity: Revisions, New Directions, and Applications*, 19–37. Charlotte, NC, US: IAP Information Age Publishing.
- Stecher, Brian. 2015. "Looking Back." In *Beyond the Bubble Test*, 15–52. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119210863.ch2>.

Steedle, Jeffrey T., and Steve Ferrara. 2016. "Evaluating Comparative Judgment as an Approach to Essay Scoring." *Applied Measurement in Education* 29 (3): 211–23.
<https://doi.org/10.1080/08957347.2016.1171769>.

Straetmans, G. 2014. "Toetsen met performance assessment methodieken." In *Toetsen in het hoger onderwijs*, edited by Henk van Berkel, Anneke Bax, and Desiree Joosten-ten Brinke. Bohn Stafleu van Loghum.

Tan, Xuan, and Rochelle Michel. 2011. "Why Do Standardized Testing Programs Report Scaled Scores?" *ETS R&D Connections*, no. 16: 6.

Toulmin, Stephen. 2003. *The Uses of Argument*. Updated ed. Cambridge, U.K. ; New York: Cambridge University Press.

van Daal, Lesterhuis, Coertjens, Donche, and De Maeyer. 2019. "Validity of Comparative Judgement to Assess Academic Writing: Examining Implications of Its Holistic Character and Building on a Shared Consensus." *Assessment in Education: Principles, Policy & Practice* 26 (1): 59–74. <https://doi.org/10.1080/0969594X.2016.1253542>.

Weigel, Tanja, Martin Mulder, and Kate Collins. 2007. "The Concept of Competence in the Development of Vocational Education and Training in Selected EU Member States." *Journal of Vocational Education & Training* 59 (1): 53–66.
<https://doi.org/10.1080/13636820601145549>.

Wools, Saskia. 2015. "All About Validity - An Evaluation System for the Quality of Educational Assessment." Enschede: University of Twente.

Wools, Saskia, P. Sanders, and E. Roelofs. 2007. *Beoordelingsinstrument: Kwaliteit van Competentie Assessment*. Arnhem: Cito.