Kwaliteit van toetsen binnen handbereik: Reviewstudie van onderzoek en onderzoeksresultaten naar de kwaliteit van toetsen[1]

N. A. M. Maassen, D. den Otter, S. Wools, B. T. Hemker, G. J. J. M. Straetmans en T. J. H. M. Eggen

Samenvatting

Toetsing is in het onderwijs aan de orde van de dag. De uitslagen van deze toetsen kunnen zeer bepalend zijn voor de onderwijscarrière van studenten. Het is daarom van belang dat de kwaliteit van de gebruikte toetsen goed is. Toch blijkt het complex om te bepalen wat goede kwaliteit is. Het doel van deze reviewstudie is inzicht te geven in wat men op dit moment beschouwt als toetskwaliteit en aan te tonen waar hiaten in kennis liggen. Op systematische wijze zijn 242 artikelen verzameld die ingaan op kwaliteitsaspecten van toetsen in het onderwijs. De genoemde kwaliteitsaspecten zijn ondergebracht in een begrippenkader, bestaande uit vijf hoofdcategorieën. De kwaliteitsaspecten die binnen de hoofdcategorie betrouwbaarheid vallen komen het meest frequent voor. De resultaten laten zien dat de frequentie waarmee de kwaliteitsaspecten worden genoemd afhangt van een aantal factoren: het doel van de toets, de fase in de toetscyclus en of een onderzoek in de praktijk is uitgevoerd of dat het een theoretische beschrijving betreft. Deze resultaten geven aanleiding tot vervolgonderzoek en aanbevelingen voor de praktijk, zodat de werkwijze in de praktijk kan worden verbeterd en toetsen van goede kwaliteit zijn.

Kernwoorden: toets, kwaliteit, onderwijs, reviewstudie

1 Inleiding

Het gebruik van toetsen is een belangrijk hulpmiddel om na te gaan in hoeverre studenten de benodigde kennis en vaardigheden beheersen om actief te kunnen participeren in de samenleving. Enerzijds worden toetsen ingezet voor het certificeren van studenten: beheersen zij de doelstellingen van het onderwijsprogramma voldoende om dit onderdeel definitief af te sluiten? Anderzijds worden toetsen gebruikt om het leren van studenten en het onderwijzen van docenten tijdig bij te kunnen stellen met het oog op de verwerving van de doelstellingen. Vanwege deze belangrijke rollen is de kwaliteit van toetsing in alle geledingen van het onderwijs van belang. Afhankelijk van de inzet en het gebruik wordt er immers een kwaliteitseis gesteld aan de toets. Uiteindelijk moet het eindniveau van studenten, en daarmee het civiel effect van het diploma, worden geborgd.

De afgelopen tijd hebben verschillende partijen, zowel binnen als buiten het onderwijs, vraagtekens gezet bij de kwaliteit van toetsen in het onderwijs (Inspectie van het Onderwijs, 2009; Onderwijsraad, 2006). Mede daardoor is de wetenschappelijke aandacht voor toetskwaliteit en wat dit precies inhoudt verder toegenomen (Joosten-ten Brinke & Sluijsmans, 2012). Hoewel er steeds meer geïnvesteerd wordt in de kwaliteit van toetsen blijven de vragen: wat wordt onder de kwaliteit van een toets verstaan? En waar hangt deze kwaliteit van af?

Het blijkt complex om te bepalen wat kwaliteit precies inhoudt. Er zou gezegd kunnen worden dat er in Nederland ongeveer 16 miljoen deskundigen zijn op het gebied van toetsing. Iedereen heeft ervaringen met toetsen en een eigen idee over wat goede en slechte toetsen zijn (Eggen, 2009). Daarnaast wordt de onduidelijkheid over toetskwaliteit versterkt doordat toetskwaliteit vanuit verschillende invalshoeken benaderd kan worden. Deze verschillende perspectieven leiden tot een onduidelijk begrippenkader. Enerzijds worden er verschillende termen gebruikt voor hetzelfde begrip, anderzijds worden dezelfde begrippen gehanteerd maar bedoelt men iets anders.

Dit is tevens terug te zien in de diversiteit aan beoordelingssystemen om toetsen op

kwaliteit te beoordelen. Zowel in Nederland (Baartman, Bastiaens, Kirschner & Van der Vleuten, 2006; Evers, Lucassen, Sijtsma & Meijer, 2010; Sanders & Hemker, 2011; Sluijsmans, 2013) als internationaal (Association for Educational Assessment - Europe, 2012; AERA, APA & NCME, 1999) zijn er verschillende systemen die veel overeenkomstige criteria bevatten maar tevens onderling verschillen, onder andere voor wat betreft het prescriptieve karakter van de systemen en de strengheid van de eisen (Wools, 2009). Mogelijk zijn deze verschillen in inzicht het gevolg van bepaalde factoren zoals het doel van de toets, de onderwijssector, de rollen in de beoordeling, de toetscyclus en of onderzoek in de praktijk is uitgevoerd of dat het een theoretische beschrijving betreft.

Doel van de toets

Aangezien onderwijskundige meetinstrumenten vanuit verschillende behoeften worden ingezet, lopen de doelen hiervan zeer uiteen. Een gangbare indeling om toetsdoelen van elkaar te onderscheiden is de indeling in summatieve en formatieve toetsen (Eggen, 2013). Summatieve toetsen worden ingezet met het doel een eindoordeel te geven over het niveau van de student. Formatieve toetsen worden ingezet om feedback te geven over het hiaat tussen het huidige niveau van de student en het gewenste niveau, met als doel de student te helpen zijn prestatie te verbeteren (Van der Kleij, Vermeulen, Schildkamp & Eggen, 2015). Eenzelfde toets kan voor verschillende doeleinden worden gebruikt (Gioka, 2009; Taras, 2005). Het is begrijpelijk dat er (deels) verschillende kwaliteitsaspecten worden benadrukt voor toetsen met een summatieve en formatieve functie. Bij een toets die bijvoorbeeld ingezet wordt voor certificeringsdoeleinden zal de focus op andere kwaliteitsaspecten kunnen liggen dan bij een voortgangstoets voor studenten.

Onderwijssector

In de onderwijssectoren primair onderwijs, voortgezet onderwijs, beroepsgericht onderwijs en hoger onderwijs vervullen toetsen verschillende doelen, worden verschillende vaardigheden gemeten (Straetmans, 2006), is de wijze van toetsing verschillend en is de rol die docenten bij het toetsproces vervullen verschillend. Vanwege het feit dat toetsen verschillende rollen vervullen in de onderwijssectoren, is het voorstelbaar dat het belang dat gehecht wordt aan bepaalde kwaliteitsaspecten verschilt per onderwijssector.

Rollen in de beoordeling

Volgens traditioneel model beoordeelt de docent de studenten zelf. Recent is er aandacht gekomen voor alternatieve beoordelingsmodellen. Daarbij neemt niet de docent, maar nemen anderen de rol van beoordelaar op zich. Voorbeelden daarvan zijn peer assessment waarbij studenten elkaars werk beoordelen, self-assessment waarbij studenten hun eigen werk beoordelen en co-assessment waarbij de docent en student samen het werk van de student beoordelen (De Grez, Valcke & Roozen, 2012; Dochy, Segers & Sluijsmans, 1999). Bij deze alternatieve beoordelingsmethodieken is het denkbaar dat andere kwaliteitsaspecten een rol spelen (Ploegh, Tillema & Segers, 2009).

Toetscyclus

Afhankelijk van de fase in de toetscyclus waarin iemand betrokken is kunnen verschillende aspecten van toetskwaliteit relevant zijn (Tillema, Leenknecht & Segers, 2011). Tijdens het construeren van een toets is het voorstelbaar dat er op andere aspecten gelet wordt dan tijdens de afname, beoordeling of evaluatie van een toets.

In de praktijk

Het verschil in de manier waarop onderzoek naar toetskwaliteit wordt gedaan kan tot verschillende inzichten op toetskwaliteit leiden. Een veronderstelling is dat artikelen waarin onderzoek in de praktijk is uitgevoerd de werkelijkheid meer benaderen dan artikelen die het concept toetskwaliteit theoretisch beschouwen. Er kan hierbij onderscheid gemaakt worden tussen empirisch onderzoek en theoretische beschouwingen. Onder empirisch onderzoek worden artikelen verstaan waarbij observaties of toetsanalyses zijn uitgevoerd in de school (o.a. Gioka, 2006) of

waarbij interviews of vragenlijsten zijn afgenomen om zicht te krijgen in de praktijksituatie (o.a. Maclellan, 2004). Onder theoretische beschouwingen worden artikelen verstaan waarin het concept toetskwaliteit vanuit een theoretisch kader wordt beschreven. Er is in deze artikelen geen onderzoek in praktijksituaties uitgevoerd (o.a. Newton, 2012).

De verschillende factoren laten zien dat de aspecten die samenhangen met kwaliteit en kwaliteitsborging van toetsen talrijk zijn en dat het begrip toetskwaliteit vanuit veel verschillende invalshoeken kan worden benaderd. Als gevolg van deze complexiteit kan het lastig zijn om de juiste beoordelingskaders te selecteren en te gebruiken om toetsen te beoordelen op hun kwaliteit.

Het doel van deze reviewstudie is om bestaande kennis en informatie over toetskwaliteit te verzamelen, te classificeren en beschikbaar te stellen voor personen in zowel de praktijk- als de onderzoekswereld. De onderzoeksvraag die in dit artikel centraal staat is: *Wat beschouwt men op dit moment als kwaliteit van toetsen in het onderwijs*? Om deze vraag te kunnen beantwoorden is een systematisch literatuuronderzoek uitgevoerd, waarbij is uitgegaan van de veronderstelling dat de veel versus weinig onderzochte kwaliteitsaspecten en het aanvullende oordeel van een expertpanel een indicatie geven over wat men als toetskwaliteit beschouwt.

2 Methode

De reviewstudie bestaat uit vier fasen: (1) het uitvoeren van een systematisch literatuuronderzoek, (2) het construeren van een begrippenkader met behulp van een expertpanel, (3) het coderen van de literatuur en (4) het analyseren van de data om de resultaten vervolgens te valideren met behulp van klankbordgroepen.

2.1 Systematisch literatuuronderzoek

In de eerste fase is een systematisch literatuuronderzoek uitgevoerd om de huidige inzichten ten aanzien van de kwaliteit van toetsen binnen het onderwijs te verzamelen. Artikelen die gaan over toetsen in andere vakgebieden, zoals psychologische toetsen, behoren niet tot de reviewstudie. Ook worden er alleen artikelen verzameld die kwaliteitscriteria van toetsen benoemen.

Om tot een juiste selectie van artikelen te komen zijn de volgende woorden als kernwoorden gebruikt naar aanleiding van de onderzoeksvraag: “kwaliteitscriteria” en “toetsen binnen het onderwijs”. Met behulp van een thesaurus zijn de mogelijke relevante variaties op deze kernwoorden achterhaald. Uiteindelijk leidde de volgende combinatie van zoektermen tot een brede selectie met relevante bronnen: (quality standard* OR quality guideline* OR quality criteri* OR evaluation criteri*) AND (educational test* OR student evaluation* OR educational assessment* OR classroom assessment*). Deze zoekopdracht is in de databases ERIC, PsychINFO, Scopus en Web of Science uitgevoerd. In ERIC en PsychINFO leidde deze zoekopdracht tot 4915 artikelen. Hierbij is geselecteerd op peer-reviewed artikelen van de afgelopen vijftien jaar (2000-2015), zodat recente, wetenschappelijk betrouwbare inzichten werden verkregen. Dit leidde tot 614 artikelen. Dezelfde zoekopdracht is tevens in de databases Scopus en Web of Science uitgevoerd, maar leidde niet tot nieuwe relevante artikelen. Van de 614 artikelen zijn de titels gelezen om te bepalen of het artikel voldeed aan de inclusiecriteria. Wanneer dit onduidelijk bleef is tevens het abstract gelezen of hebben de auteurs de relevantie van het artikel samen besproken. Inclusiecriteria voor geschikte artikelen waren: (1) het bevatten van kwaliteitsaspecten van een toets, (2) het betrekking hebben op het onderwijs en (3) het richten op het toetsprogramma, de toets zelf of het itemniveau van de toets. De overige artikelen zijn verwijderd omdat deze artikelen ingingen op de evaluatie van de kwaliteit van de docent of de evaluatie van het onderwijs in het algemeen, de kwaliteit van het toetsbeleid of toetsing op de werkplek. Van de 134 artikelen die op basis van deze inclusiecriteria zijn geselecteerd zijn er bij codering van de inhoud nog 43 artikelen verwijderd omdat ze niet binnen de scope van het onderzoek bleken te vallen. Op basis van de overgebleven

91 artikelen zijn nog 56 artikelen toegevoegd door middel van de sneeuwbalmethode. Hiertoe behoren ook enkele artikelen van voor 2000. Dit zijn vaak geciteerde en daarmee belangrijk geachte artikelen.

Aanvullend is er gezocht naar praktijkgericht onderzoek in de Nederlandse toetspraktijk dat vanaf het jaar 2000 is verschenen. Bronnen hiervoor waren vaktijdschriften (*Didactief, Examens, OnderwijsInnovatie, Pedagogische Studiën, Tijdschrift voor Hoger Onderwijs* en *Toets!)*, mastertheses en proefschriften van meerdere universiteiten, bibliotheken en informatie van een aantal lectoren werkzaam op het terrein van onderwijskundig meten. Dit leverde 95 relevante artikelen op. In totaal zijn er 242 artikelen geselecteerd (Maassen et al., 2014).

2.2 Van expertpanel naar begrippenkader

In de tweede fase is een begrippenkader geconstrueerd met behulp van een expertpanel bestaande uit zes experts. Deze experts zijn geselecteerd op basis van hun deskundigheid op het gebied van toetsing. De zes experts zijn allen verbonden aan verschillende lectoraten en toetsinstanties. Om tot een begrippenkader van kwaliteitsaspecten te komen is aan deze experts een vragenlijst voorgelegd. Hiervoor zijn de kwaliteitsaspecten voorgelegd die op basis van een eerste globale analyse uit de literatuur naar voren zijn gekomen. Experts beoordeelden hun bekendheid met en het belang van de verschillende kwaliteitsaspecten. Ook is hen gevraagd om de lijst met kwaliteitsaspecten aan te vullen. Tot slot is er gevraagd welke indeling in hoofd- en subcategorieën er volgens de experts gemaakt kan worden.

Op basis van de antwoorden van de experts op de vragenlijst en een eerste oriëntatie op de literatuur is door de auteurs een begrippenkader opgesteld waarin alle kwaliteitsaspecten verwerkt zijn. Nadat een volgende selectie van artikelen is gelezen, is het begrippenkader aangepast om er zeker van te zijn dat alle relevante kwaliteitsaspecten zijn opgenomen. Het uiteindelijke begrippenkader bestaat uit drie niveaus: hoofdcategorieën, subcategorieën en onderdelen. Het niveau van de onderdelen is het laagste niveau. Een aantal onderdelen kunnen worden gegroepeerd onder één subcategorie. Meerdere subcategorieën vormen één hoofdcategorie. Deze categorieën zijn gevormd op basis van de indelingen die gemaakt zijn door de experts, de indelingen die in de literatuur zijn gegeven en inzichten van de auteurs.

2.3 Coderen van literatuur

In de derde fase van de reviewstudie is de inhoud van alle artikelen gecodeerd aan de hand van de kwaliteitsaspecten in het begrippenkader. Voor elk artikel is bepaald op welke kwaliteitsaspecten het artikel betrekking heeft. Eén artikel kon betrekking hebben op meerdere kwaliteitsaspecten. Als een artikel een bepaald kwaliteitsaspect benoemde, kreeg dit aspect waarde 1. Als een kwaliteitsaspect niet werd genoemd kreeg het waarde 0.

Deze codering maakte echter nog geen onderscheid in de drie niveaus uit het begrippenkader. Daarom is de codering omgezet in scores die wel rekening houden met de hoofdcategorieën, subcategorieën en onderdelen. Deze hercodering ging als volgt: Wanneer een kwaliteitsaspect, bijvoorbeeld *meerdere beoordelaars*, in een artikel voorkwam kreeg dit onderdeel waarde 1. De subcategorie waar dit onderdeel onder viel, in dit geval de subcategorie *objectiviteit*, kreeg hierdoor tevens waarde 1. Tot slot kreeg ook de hoofdcategorie waar dit onderdeel deel van uitmaakt, in dit geval de hoofdcategorie *betrouwbaarheid*, waarde 1.

In veel artikelen zijn meerdere kwaliteitsaspecten genoemd, bijvoorbeeld *meerdere beoordelaars* én *beoordelingsvoorschrift*. De bijbehorende subcategorie *objectiviteit* en hoofdcategorie *betrouwbaarheid* kregen in dat geval toch waarde 1 en niet waarde 2 of 3. Zo is vergelijking mogelijk tussen verschillende hoofdcategorieën met een verschillende hoeveelheid subcategorieën, en subcategorieën met een verschillend aantal onderdelen.

Om de betrouwbaarheid van het coderen van de artikelen te borgen zijn een aantal artikelen dubbel gecodeerd. Hieruit bleek dat 60% van de artikelen precies dezelfde kwaliteitsaspecten toegekend hebben gekregen. Na overleg tussen de twee beoordelaars bleken

er slechts kleine verschillen in codering van kwaliteitsaspecten van de overige artikelen te zijn.

2.4 Analyseren van data en valideren met klankbordgroepen

In de laatste fase zijn de kwaliteitsaspecten op de drie niveaus (hoofdcategorie, subcategorie en onderdeel) geanalyseerd. Allereerst is een beschrijvende analyse uitgevoerd waarbij de frequenties van de kwaliteitsaspecten bij het totaal aantal artikelen (n=242) is onderzocht.

Vervolgens zijn er analyses uitgevoerd die specifiek ingingen op de genoemde factoren: het toetsdoel, de onderwijssectoren, de rollen in de beoordeling, de fase in de toetscyclus en of een onderzoek in de praktijk is uitgevoerd of dat het artikel een theoretische beschrijving betreft. Met behulp van de Pearson Chi-Kwadraattoets is getoetst of er verschillen in kwaliteitsaspecten tussen de groepen waren. Zo zijn de artikelen over summatieve toetsing vergeleken met de artikelen over formatieve toetsing. Vanwege inhoudelijke argumenten is er in het geval van de factoren onderwijssectoren en de toetscyclus gestart met de totale groep als basisgroep waar achtereenvolgens de subgroepen tegen zijn afgezet. Zo is er voor elke subgroep gekeken of deze significant afwijkt van de totale groep. Artikelen over het hoger onderwijs zijn bijvoorbeeld vergeleken met alle artikelen die betrekking hebben op één specifieke onderwijssector. Tot deze totale groep behoren dus wederom de artikelen over het hoger onderwijs, omdat de onderwijssectoren zonder het hoger onderwijs als groep geen eenheid vormen. Bovendien vindt er nu geen kunstmatige toename van verschillen tussen de groepen plaats, aangezien de geselecteerde groep deel uitmaakt van de totale groep en deze groep dus nog sterker moet afwijken voordat er daadwerkelijk significante verschillen gevonden zullen worden.

Om de validiteit van de resultaten te borgen is ervoor gekozen om binnen de specifieke analyses alleen artikelen mee te nemen die betrekking hebben op precies één groep binnen een factor. Ter illustratie: een artikel dat betrekking heeft op het hoger onderwijs is meegenomen in de analyse, maar een artikel dat de onderwijssector niet heeft gespecificeerd of betrekking heeft op meerdere onderwijssectoren is niet meegenomen. Zo kunnen aspecten die betrekking hebben op een andere genoemde onderwijssector niet interfereren in de uitspraak over kwaliteitsaspecten van het hoger onderwijs.

Tot slot zijn de resultaten van de codering en het gecreëerde overzicht van de resultaten ter validering voorgelegd aan verschillende klankbordgroepen bestaande uit docenten werkzaam in verschillende sectoren van het onderwijs. Met behulp van semigestructureerde interviews is nagegaan in hoeverre de resultaten uit het literatuuronderzoek in de praktijk werden herkend.

3 Resultaten

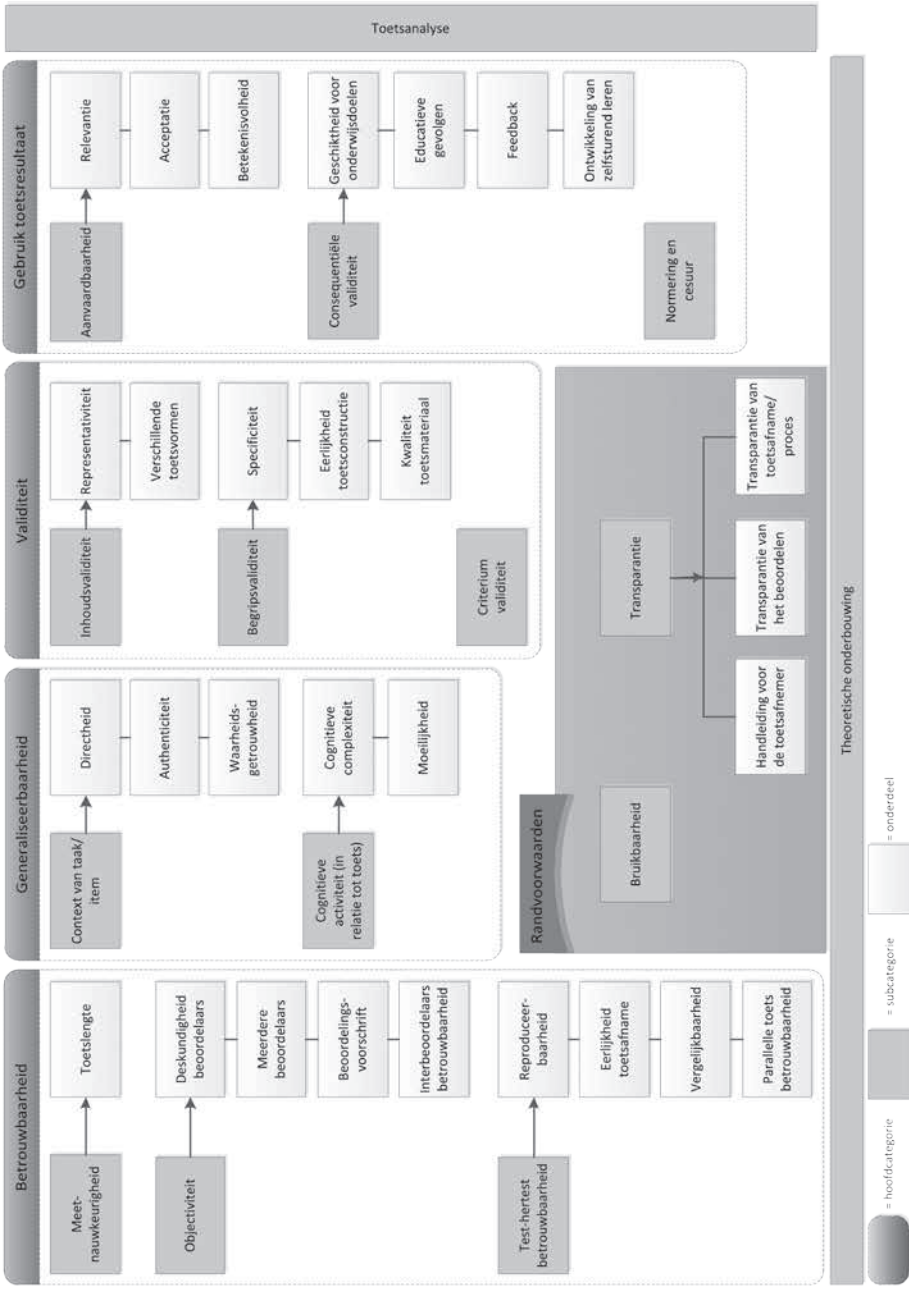
Allereerst wordt er ingegaan op de resultaten van het expertpanel, waarna het begrippenkader wordt gepresenteerd. Vervolgens worden de bevindingen beschreven met betrekking tot de frequentie van de kwaliteitsaspecten en de verschillen in frequentie binnen de factoren zoals in de inleiding genoemd.

3.1 Expertpanel

Uit de vragenlijst die bij zes experts op het gebied van toetsing is afgenomen bleek dat de voorgelegde kwaliteitsaspecten uit de eerste literatuuroriëntatie bijna allemaal bekend zijn bij de experts. Er werden enkele aanvullingen gegeven op de lijst van kwaliteitsaspecten, maar deze aanvullingen vielen met hun definitie onder kwaliteitsaspecten die al opgenomen zijn in het begrippenkader. Alle experts zien validiteit en betrouwbaarheid als twee hoofdcategorieën. Daarnaast zien drie experts gebruiksgemak of bruikbaarheid ook als hoofdcategorie. Geen van de indelingen van de experts is echter gelijk. Dit bevestigt de aanname dat er geen eenduidige visie is op toetskwaliteit.

3.2 Begrippenkader

De indeling van kwaliteitsaspecten door het expertpanel en de gehanteerde indelingen in de geselecteerde literatuur hebben geresulteerd in een begrippenkader dat de veelheid van kwaliteitsaspecten samenvat (Figuur 1).



Figuur 1. Begrippenkader. Aspecten van toetskwaliteit volgens de geselecteerde literatuurbronnen en experts.

Er zijn hierin vijf hoofdcategorieën onderscheiden: betrouwbaarheid, generaliseerbaarheid, validiteit, gebruik van het toetsresultaat en randvoorwaarden. In Tabel 1 worden definities en voorbeelden van deze hoofdcategorieën genoemd. De meeste kwaliteitsaspecten uit de artikelen hebben betrekking op het niveau van de toets, zoals toetslengte en transparantie van de toetsafname. Slechts enkele aspecten gaan in op het itemniveau, zoals eerlijkheid van toetsconstructie (de mate waarin toetsitems discriminerende aspecten bevatten), of hebben betrekking op het toetsprogrammaniveau, zoals verschillende toetsvormen.

Tabel 1
Definities en voorbeelden van hoofdcategorieën uit begrippenkader

Definitie	Voorbeeld
Betrouwbaarheid is de mate waarin de scores op een toets consistent, nauwkeurig en reproduceerbaar zijn. In dat geval is het meetresultaat vrij van meetfouten.	Als een student op maandag een toets maakt, zou het resultaat hetzelfde moeten zijn als wanneer hij op dinsdag de toets maakt (ervan uitgaande dat het kennisniveau gelijk is gebleven).
Generaliseerbaarheid is de mate waarin datgene wat een student in de toets laat zien (in deze specifieke omstandigheden), ook opgaat in andere omstandigheden.	Als een student verpleegkunde aantoont opgaven te beheersen die te maken hebben met het toedienen van vloeibare medicatie, mag er dan van uit worden gegaan dat deze student voldoende vaardig is op het gebied van verpleegkundig rekenen?
Validiteit is de eigenschap dat de toets meet wat de constructeur bedoeld heeft ermee te meten. Welke conclusie kan er getrokken worden uit een toetsresultaat?	Een student die minder taalvaardig is maakt een rekentoets die uit veel verhaalsommen bestaat. Zijn lage score wordt verklaard door zijn slechte rekenvaardigheid. Of heeft hij de taal in de sommen niet goed begrepen en daardoor een lage score behaald?
Gebruik toetsresultaat gaat over de vraag hoe het toetsresultaat wordt verwerkt en wat er vervolgens mee wordt gedaan.	Als de student 50 punten heeft gehaald, krijgt hij een onvoldoende. Hij krijgt hulp op de onderdelen die hij niet goed heeft gemaakt.
Randvoorwaarden zijn voorwaarden om te komen tot toetskwaliteit.	Een student moet op de hoogte zijn van de toetsafname en aspecten die daarbij komen kijken, zodat hij goed voorbereid is.

Binnen elke hoofdcategorie zijn subcategorieën onderscheiden. Deze subcategorieën zijn aan de linkerzijde van elk hoofdcategoriekader weergegeven. Binnen de subcategorieën bestaan één of meerdere onderdelen. Deze onderdelen zijn aan de rechterzijde van elk hoofdcategoriekader weergegeven, waarbij de pijl aangeeft onder welke subcategorie dit valt. Bij de hoofdcategorie randvoorwaarden zijn de subcategorieën horizontaal in het kader weergegeven en zijn de onderdelen hieronder geplaatst. Binnen de hoofdcategorie *betrouwbaarheid* bijvoorbeeld, is de subcategorie *objectiviteit* te bewerkstelligen door een helder *beoordelingsvoorschrift*, *deskundige beoordelaars* en/of *meerdere beoordelaars* in te zetten, waarbij deze beoordelaars onderling een goede overeenstemming over de toe te kennen scores weten te bereiken zodat er sprake is van *interbeoordelaarsbetrouwbaarheid*.

De kwaliteitsaspecten *toetsanalyse* en *theoretische onderbouwing* van de toets hebben betrekking op meerdere aspecten binnen het begrippenkader en zijn om die reden aan de zijkanten van het begrippenkader weergegeven. In de analyses zijn zij als subcategorie meegenomen.

3.3 Veel en weinig voorkomende kwaliteitsaspecten

Tabel 2 beschrijft de verdeling van het totaal aantal kwaliteitsaspecten dat in de 242 artikelen voorkomt. De verschillende kwaliteitsaspecten in de hoofdcategorieën worden in totaal 664 keer genoemd in deze 242 artikelen en de verschillende kwaliteitsaspecten in de subcategorieën worden in totaal 1001 keer genoemd. De beschrijvende analyse laat zien dat de hoofdcategorie *betrouwbaarheid* het meest frequent voorkomt in de artikelen: ongeveer 28% van de genoemde kwaliteitsaspecten valt onder deze hoofdcategorie (Tabel 2). Op het niveau van de subcategorieën komt *objectiviteit* het meest voor (14.4%). Het meest genoemde onderdeel is het *beoordelingsvoorschrift* (9.8%). De hoofdcategorieën *generaliseerbaarheid* en *randvoorwaarden* komen het minst vaak voor (11.4% resp. 16.6%).

3.4 Invloed van factoren op kwaliteitsaspecten

Doel van de toets

De frequentie van kwaliteitsaspecten is significant anders in artikelen over summatieve toetsing (n=64) vergeleken met de artikelen over formatieve toetsing (n=24; $\chi^2(4)=20.27$,

Tabel 2
Frequentie waarmee de verschillende kwaliteitsaspecten in de hoofd- en subcategorieën worden genoemd in de 242 artikelen

Hoofdcategorie	**Frequentie**		**Subcategorie**	**Frequentie**	
	N	*%*		*N*	*%*
Betrouwbaarheid	188	28.3			
			Betrouwbaarheid	54	5.4
			Meetnauwkeurigheid	27	2.7
			Objectiviteit	144	14.4
			Test-hertest betrouwbaarheid	60	6.0
Generaliseerbaarheid	76	11.4			
			Generaliseerbaarheid	24	2.4
			Context van taak / item	48	4.8
			Cognitieve activiteit	48	4.8
Validiteit	146	22.0			
			Validiteit	73	7.3
			Inhoudsvaliditeit	65	6.5
			Begripsvaliditeit	73	7.3
			Criterium validiteit	21	2.1
Gebruik toetsresultaat	144	21.7			
			Aanvaardbaarheid	41	4.1
			Consequentiële validiteit	100	10.0
			Normering en cesuur	47	4.7
Randvoorwaarden	110	16.6			
			Bruikbaarheid	47	4.7
			Transparantie	94	9.4
			Theoretische onderbouwing	20	2.0
			Toetsanalyse	15	1.5
Totaal	664	100	Totaal	1001	100

$p<.001$). In Tabel 3 zijn de verschillen weergegeven. Zo wordt de hoofdcategorie *validiteit* in 73.4% van de artikelen over summatieve toetsing genoemd, terwijl slechts 41.7% van de artikelen over formatieve toetsing deze hoofdcategorie benoemen. Ook de subcategorie *normering en cesuur* komt bij artikelen over summatieve toetsing meer frequent voor dan bij artikelen over formatieve toetsing (35.9% resp. 4.2%).

Daarentegen laten de beschrijvende analyses zien dat de subcategorieën *bruikbaarheid* (6.3% resp. 25.0%) en *consequentiële validiteit* (20.3% resp. 70.8%; de gewenste en ongewenste effecten van een toets op het leren van studenten) in de artikelen over summatieve toetsing minder vaak genoemd worden. Wellicht zijn dit aspecten die meer met formatieve toetsing te maken hebben. Het is echter niet mogelijk om dit op significantie te toetsen aangezien er relatief weinig artikelen betrekking hadden op formatieve toetsing.

Onderwijssector

Tussen de onderwijssectoren zijn geen significante verschillen gevonden wat betreft de frequenties van de kwaliteitsaspecten. Dit resultaat is gebaseerd op 148 artikelen waarin de onderwijssector werd gespecificeerd. Er lijkt wel een verschillende mate van aandacht voor toetskwaliteit te zijn: van relatief weinig aandacht in het primair onderwijs tot relatief veel aandacht in het hoger onderwijs.

Tabel 3
Percentage artikelen waarin kwaliteitsaspecten worden beschreven per toetsdoel

Kwaliteitsaspect beschreven in artikel	Toetsdoel beschreven in artikel	
	Summatief	Formatief
Validiteit	73.4%	41.7%
Normering en cesuur	35.9%	4.2%
Bruikbaarheid	6.3%	25.0%
Consequentiële validiteit	20.3%	70.8%

Dit resultaat blijkt zowel uit de klankbordgesprekken als uit het verschil in aantal artikelen die het primair onderwijs (n=9) en het hoger onderwijs (n=102) betroffen.

Rollen in de beoordeling

Er is slechts een beperkt aantal artikelen dat inging op kwaliteitsaspecten van alternatieve beoordelingsmethodieken zoals peer, self- en co-assessment (n=17). De leden van de klankbordgroep bevestigen deze beperkte aandacht vanuit de praktijk. Er kunnen door het beperkt aantal artikelen geen analyses worden uitgevoerd naar de kwaliteitsaspecten die bij de verschillende beoordelingsmethodieken een rol spelen.

Toetscyclus

De frequentie van de kwaliteitsaspecten hangt af van de fasen in de toetscyclus. De frequenties van de kwaliteitsaspecten van de drie niveaus (hoofdcategorie, subcategorie en onderdelen) wijken bij de artikelen over de fase van afname en beoordeling significant af van de hele groep ($\chi^2(4)=10.97$, $p=.027$; $\chi^2(17)=45.64$, $p<.001$; $\chi^2(28)=65.90$, $p<.001$). De hoofdcategorie *generaliseerbaarheid* lijkt minder frequent voor te komen in de fase van afname en beoordeling (20.0%) dan in de constructiefase (50.0%) en evaluatiefase (47.1%). Aspecten rondom betrouwbaarheid, zoals de subcategorie *objectiviteit* (69.8% resp. 43.8% en 44.7%) en de onderdelen *beoordelingsvoorschrift* (40.5% resp. 25% en 28.2%) en *deskundigheid van de beoordelaar* (33.6% resp. 12.5% en 20.0%) blijken een hogere frequentie te hebben in de afname- en beoordelingsfase.

Bij de evaluatiefase is er op het niveau van de onderdelen tevens een significant verschil ($\chi^2(28)=56.27$, $p<.001$). De twee onderdelen *beoordelingsvoorschrift* (28.2% resp. 40.5%) en *deskundigheid van de beoordelaar* (20.0% resp. 33.6%) komen in artikelen over de evaluatiefase (n=85) juist minder voor dan in de afname- en beoordelingsfase (n=116). Artikelen met betrekking tot de constructiefase (n=16) tonen geen significant verschil met alle artikelen die de fase in de toetscyclus specificeerden (n=217).

In de praktijk

In ruim 60% van het totaal aantal artikelen is geen onderzoek in de praktijk uitgevoerd (n=146). Deze artikelen geven een theoretische beschouwing van toetskwaliteit. Tabel 4 laat zien dat in theoretische beschouwingen op hoofdcategorieniveau significant andere kwaliteitsaspecten worden genoemd vergeleken met wat wordt verwacht als de empirische onderzoeken als basisgroep worden genomen ($\chi^2(4)=11.69$, $p\leq.05$). De hoofdcategorie *generaliseerbaarheid* lijkt vaker in theoretische beschouwingen genoemd te worden dan in artikelen waarin empirische onderzoeken zijn uitgevoerd (35.6% resp. 25.0%). Het verschil op hoofdcategorieniveau is echter enkel het geval wanneer de empirische onderzoeken als basisgroep in de analyse worden meegenomen: andersom is er geen verschil op hoofdcategorieniveau gevonden ($\chi^2(4)=6.55$, $p=.162$).

Verder is er in zowel empirische onderzoeken als in theoretische beschouwingen veel aandacht voor *betrouwbaarheid* en aspecten die daarmee samenhangen (82.3%

Tabel 4
Toetsingsgrootheden (χ^2) van empirisch onderzoek vergeleken met theoretische beschouwingen.

	Empirisch onderzoek			Theoretische beschouwing		
	Hoofd-categorie	**Sub-categorie**	**Onder-deel**	**Hoofd-categorie**	**Sub-categorie**	**Onder-deel**
Empirisch onderzoek [a]	-	-	-	11.69*	96.26**	- [b]
Theoretische beschouwing [a]	6.55	34.20**	92.41**	-	-	-

Noot: [a] basisgroep; [b] niet uitvoerbare analyse vanwege een onderdeel met waarde 0; * $p \leq .05$; ** $p \leq .01$.

resp. 74.7%). De subcategorie *validiteit* wordt in theoretische beschouwingen daarentegen relatief vaker genoemd dan in empirische onderzoeken (39.0% resp. 16.7%). De groepen verschillen op subcategorieniveau significant van elkaar ($\chi^2(17)=34.20$, $p=.008$; $\chi^2(17)=92.26$, $p<.001$).

Tot slot zijn ook op het niveau van de onderdelen significante verschillen gevonden ($\chi^2(28)=92.41$, $p \leq .001$). Vooral het kwaliteitsaspect *meerdere beoordelaars* wordt in theoretische beschouwingen relatief vaker genoemd dan in empirische onderzoeken (15.1% resp. 6.3%). Uit de klankbordgesprekken blijkt dat dit in de praktijk wel belangrijk werd gevonden, maar dat het vaak niet haalbaar was. Meerdere beoordelaars inzetten voor de scoring van toetsen is niet efficiënt en vaak te duur.

4 Conclusies en discussie

Deze reviewstudie heeft de huidige kennis over toetskwaliteit in het onderwijs in kaart gebracht. Met behulp van een begrippenkader is weergegeven welke aspecten een rol spelen bij toetskwaliteit. Er is aangetoond welke kwaliteitsaspecten vaak of minder vaak in de literatuur voorkomen. Daarnaast blijkt het belang van de kwaliteitsaspecten te verschillen afhankelijk van het doel van de toets, de fase in de toetscyclus en of een onderzoek in de praktijk is uitgevoerd of dat het een theoretische beschrijving betreft.

Deze resultaten zijn uiteraard mede bepaald door de focus van het onderzoek en daarmee de keuze van de zoektermen. Het is onwaarschijnlijk dat werkelijk alle artikelen die betrekking hebben op toetskwaliteit zijn gevonden. Er zijn relatief veel artikelen gevonden op toetsniveau en weinig op item- of toetsprogrammaniveau. Tevens zijn de gekozen indelingen binnen de factoren medebepalend geweest voor de resultaten. In dit onderzoek is gekozen voor de meest gangbare indelingen in de geselecteerde literatuur, zoals het onderscheid tussen een summatief en formatief toetsdoel. Andere indelingen zouden mogelijk tot andere resultaten kunnen leiden. Tot slot laat deze reviewstudie zien naar welke kwaliteitsaspecten veel of weinig onderzoek is gedaan. De hoeveelheid onderzoek kan de gepercipieerde waarde van een kwaliteitsaspect aantonen, maar hier kunnen eveneens andere verklaringen voor worden gegeven. Zo zou een weinig onderzocht kwaliteitsaspect moeilijk te onderzoeken kunnen zijn, terwijl het om een voor de toetskwaliteit cruciaal begrip gaat.

De resultaten en kanttekeningen geven aanleiding voor vervolgonderzoek. Hoewel er geen opvallende verschillen zijn gevonden tussen de onderwijssectoren bleek er wel verschil in de mate van aandacht voor toetskwaliteit te zijn. Dit suggereert dat er binnen de onderwijssectoren sprake is van verschillende behoeften aan informatie over toetskwaliteit. Een eerste aanbeveling is dan ook om toekomstig onderzoek en praktijkgerichte interventies aan te passen aan de desbetreffende sector. Zo zou het in het primair en voortgezet onderwijs relevant kunnen zijn om meer bewustwording te creëren omtrent het belang van kwalitatief goede toetsen. Binnen het beroepsgericht en hoger

onderwijs zouden daarentegen meer specifieke vraagstukken kunnen worden onderzocht wat betreft manieren om een goede toetskwaliteit te bereiken.

Ten tweede kwam naar voren dat de fase in de toetscyclus bepalend is voor welke kwaliteitsaspecten van belang zijn. Vooral binnen de fase van afname en boordeling van toetsen zijn specifieke kwaliteitsaspecten gevonden. Omdat er een beperkt aantal artikelen over de constructiefase was gevonden kon hier geen uitspraak over worden gedaan. Mogelijk hangt het beperkt aantal artikelen over deze fase samen met het lage aantal artikelen over het niveau van toetsitems. Een aanbeveling op grond van dit resultaat is dan ook om nader onderzoek te doen naar de kwaliteitsaspecten per fase in de toetscyclus. Is er een (betere) koppeling te maken tussen de kwaliteitsaspecten en procesfase, waardoor het proces beter ondersteund wordt met als uiteindelijk resultaat dat de kwaliteit verbeterd wordt?

Ten derde bleek dat validiteit relatief weinig voorkomt in empirische artikelen, terwijl dit in theoretische beschouwingen veelvuldig wordt beschreven. Een mogelijk onderzoeksthema is daarom validiteitsaspecten in de praktijk. Een vraag die hierbij gesteld kan worden is: hoe kunnen validiteitsbedreigingen van het goed meten van specifieke vaardigheden concreet worden gemaakt en hoe kunnen zij worden opgelost? Het onderzoeksthema zou zich ook kunnen richten op het adresseren van specifieke vragen uit de praktijk ten aanzien van de validiteit van toetsen, zoals het beoordelen van individuele bijdragen in groepsprestaties. Er is bijvoorbeeld aangetoond dat het inzetten van peer assessment een oplossing kan zijn bij het beoordelen van individuele prestaties bij groepswerkstukken (Cheng & Warren, 2010). Dit is echter in Hongkong onderzocht en niet in Nederlandse context. Bax (2004) heeft wel onderzoek in Nederland gedaan naar beoordeling van individuele bijdragen in groepsprestaties, maar men was hier voorzichtiger met de inzet van peer assessment.

Daarnaast is er vooral onderzoek gevonden op het niveau van de toets. Er kan ook worden ingezoomd op de opgaven (itemniveau) of juist worden uitgezoomd naar het toetsprogrammaniveau. Het komt namelijk in alle onderwijssectoren voor dat toetsresultaten of metingen van een vaardigheid van studenten met elkaar worden gecombineerd om te komen tot een beoordeling. De manier waarop deze verschillende toetsresultaten gecombineerd worden is minstens zo belangrijk als de toetsen die worden ingezet om tot de resultaten te komen (Chester, 2003). Het inrichten van een kwalitatief goed toetsprogramma kan hierbij helpen (Baartman, Kloppenburg & Prins, 2013). Daarnaast geldt dat toetsen verweven kunnen worden in het leerproces door ze op een meer formatieve manier in te zetten (Shepard, 2009). Voor zowel het combineren van toetsresultaten als toetsing die zich richt op het leerproces geldt dat er onderzoek nodig is om inzicht te krijgen in de effectiviteit van de inrichting van een toetsprogramma.

Naast aanbevelingen voor vervolgonderzoek kunnen er ook praktische aanbevelingen gedaan worden. Uit het onderzoek bleek dat het toetsdoel een beïnvloedende factor is op de frequentie van de kwaliteitsaspecten. Bij summatieve toetsing is het immers meer dan bij formatieve toetsing van belang dat er een juiste (valide) beslissing wordt genomen op basis van de toetsscore. Als er belangrijke beslissingen over studenten worden genomen op basis van de toets is het van belang dat de toets datgene meet wat getoetst moet worden (Tanilon, Segers, Vedder & Tillema, 2009) en dat de normering op een juiste manier tot stand komt en wordt gebruikt (Van Berkel, 2004; Dalbert, Schneidewind & Saalbach, 2007) zodat er een betekenisvolle interpretatie van de toetsscore gegeven kan worden. In aansluiting op de resultaten van dit onderzoek lijkt consequentiële validiteit, zoals feedback (Black & Wiliam, 1998; Gioka, 2006), ontwikkeling van zelfregulerend leren (Nieweg, 2002) en educatieve gevolgen (Young & Kim, 2010) een kwaliteitsaspect te zijn dat meer van belang is bij formatieve toetsing. Voor docenten is het dus van belang dat zij eerst het toetsdoel vaststellen. Aan de hand daarvan kunnen zij nagaan met welke kwaliteitsaspecten zij vooral rekening moeten houden.

Tot slot kunnen de aspecten van

toetskwaliteit zoals uit deze reviewstudie is gebleken worden toegepast in de praktijk. De resultaten zijn daarom verwerkt in een praktijkgericht boekje voor docenten: 'Eerste Hulp Bij Toetsen' (Maassen & Den Otter, 2014). In dit boekje zijn de kwaliteitsaspecten in concrete situaties vertaald en worden checklists gegeven die ondersteuning bieden bij het in de praktijk brengen van deze kennis. De resultaten van de reviewstudie kunnen zodoende helpen de werkwijze van de praktijk te verbeteren, zodat de toetsen op grond waarvan belangrijke beslissingen over studenten tot stand komen van goede kwaliteit zijn.

Noot

[1] Dit onderzoek is gefinancierd door de NRO Programmaraad voor Praktijkgericht Onderwijsonderzoek (projectnummer 405-14-535).

Literatuur

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing*. Washington: American Psychological Association.

Association for Educational Assessment – Europe. (2012). *European Framework of Standards for Educational Assessment*.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2006). The wheel of competency assessment: Presenting quality criteria for Competency Assessment Programs. *Studies in Educational Evaluation, 32*(2), 153-170. doi:10.1016/j.stueduc.2006.04.006

Baartman, L. K. J., Kloppenburg, R., & Prins, F. J. (2013). *Kwaliteit van toetsprogramma's*. In Van Berkel, H., Baks, A., & Joosten-ten Brinke, D. (red.), *Toetsen in het Hoger Onderwijs, 3e druk*. Houten: Bohn Stafleu van Loghum.

Bax, A. E. (2004). Beoordelingsmethoden voor het toekennen van individuele cijfers aan groepsproducten: Loon naar werken. *Examens, 4*, 18-21.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. doi:10.1080/0969595980050102

Cheng, W., & Warren, M. (2010). Making a difference: Using peers to assess individual students' contributions to a group project. *Teaching in Higher Education, 5*(2), 243-255. doi:10.1080/135625100114885

Chester, M. D. (2003). Multiple measures and high stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice, 22*(2), 32-41. doi:10.1111/j.1745-3992.2003.tb00126.x

Dalbert, C., Schneidewind, U., & Saalbach, A. (2007). Justice judgments concerning grading in school. *Contemporary Educational Psychology, 32*, 420-433. doi:10.1016/j.cedpsych.2006.05.003

De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self- and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education, 13*(2), 129-142. doi:10.1177/1469787412441284

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24*(3), 331-350. doi:10.1080/03075079912331379935

Eggen, T. J. H. M. (2009). *De kwaliteit van Toetsen*. Oratie Universiteit Twente, 9 april 2009.

Eggen, T. J. H. M. (2013). *Computerized adaptive testing serving educational testing purposes*. Paper presented at IAEA Conference, Tel Aviv, Israel.

Evers, A., Lucassen, W., Sijtsma, K., & Meijer, R. R. (2010). *COTAN beoordelingssysteem.* NIP, Utrecht.

Gioka, O. (2006). Assessment for learning in physics investigations: Assessment criteria, questions and feedback in marking. *Physics Education, 41*(4), 341-346. doi:10.1088/0031-9120/41/4/009

Gioka, O. (2009). Teacher or examiner? The tensions between formative and summative assessment in the case of science coursework. *Research in Science Education, 39*(4), 411-428. doi:10.1007/s11165-008-9086-9

Inspectie van het Onderwijs (2009). *Boekhouder of wakend oog. Verslag van een onderzoek bij examencommissies in het hoger onderwijs over de garantie van het niveau*. Inspectierapport 2009-16 (april). Verkregen van: http://www.onderwijsinspectie.nl/actueel/publicaties/Boekhouder+of+wakend+oog.html

Joosten-ten Brinke, D., & Sluijsmans, D. M. A. (2012). Tijd voor toetskwaliteit: het borgen van toetsdeskundigheid van examencommissies. *TH&MA, 19*(4), 16-21. Verkregen van: http://hdl.handle.net/ 1820/4759

Maassen, N. A. M., & Den Otter, D. (2014). *Eerste hulp bij toetsen: Grip op toetskwaliteit*. Verkregen van: http://www.nro.nl/wp-content/uploads/2014/12/RCEC_Kwaliteit_Toets_Checklist_2014.pdf

Maassen, N. A. M., Den Otter, D., Wools, S., Hemker, B. T., Straetmans, G. J. J. M., & Eggen, T. J. H. M. (2014). *Kwaliteit van toetsen binnen handbereik. Een reviewstudie van onderzoek en onderzoeksresultaten naar de kwaliteit van toetsen.* Verkregen van: http://www.nro.nl/wp-content/uploads/2014/12/Eindrapportage-PPO-Reviewstudie-Kwaliteit-van-toetsen-binnen-handbereik-Eggen-et-al.pdf

Maclellan, E. (2004). Initial knowledge states about assessment: Novice teachers' conceptualisations. *Teaching and Teacher Education, 20*, 523-535. doi:10.1016/j.tate.2004.04.008

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 1-29. doi:10.1080/15366367.2012.669666

Nieweg, M. R. (2002). Leren van toetsen: Op weg naar een nieuw model. *Tijdschrift voor Hoger Onderwijs, 20*(1), 42-59.

Onderwijsraad (2006). *Advies Examinering: draagvlakken toegankelijkheid, uitgebracht aan de staatssecretaris van Onderwijs, Cultuur en Wetenschap*. *Nr. 20060320/86*5. Den Haag (november). Verkregen van: www.onderwijsraad.nl/upload/publicaties/316/documenten/examinering__draagvlak_en_toegankelijkheid.pdf

Ploegh, K., Tillema, H. H., & Segers, M. S. R. (2009). In search of quality criteria in peer assessment practices. *Studies in Educational Evaluation, 35*, 102-109. doi:10.1016/j.stueduc.2009.05.001

Sanders, P. F., & Hemker, B. T. (2011). De kwaliteit van toetsen en examens. In: P.F. Sanders (Ed.) *Toetsen op school* (pp. 157-174). Arnhem: Cito.

Shepard, L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement Issues and Practice, 28*(3), 32-37. doi:10.1111/j.1745-3992.2009.00152.x

Sluijsmans, D. M. A. (2013). *Verankerd in leren: Vijf bouwstenen voor professioneel beoordelen in het hoger beroepsonderwijs*. Lectorale rede. Heerlen: Hogeschool Zuyd.

Straetmans, G. J. J. M. (2006). *Bekwaam beoordelen en beslissen*. Lectorale rede. Deventer: Saxion Hogescholen.

Tanilon, J., Segers, M., Vedder, P., & Tillema, H. (2009). Development and validation of an admission test designed to assess samples of performance on academic tasks. *Studies in Educational Evaluation, 35*, 168-173. doi:10.1016/j.stueduc.2009.12.003

Taras, M. (2005). Assessment - summative and formative - some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466-478. doi:10.1111/j.1467-8527.2005.00307.x

Tillema, H., Leenknecht, M., & Segers, M. (2011). Assessing assessment quality: Criteria for quality assurance in design of (peer)assessment for learning - A review of research studies. *Studies in Educational Evaluation, 37*, 25-34. doi:10.1016/j.stueduc.2011.03.004

Van Berkel, H. J. M. (2004). Zoeken naar normen: het geven van cijfers blijft een probleem. *Examens, 1*(4), 9-11.

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making: Assessment for learning and diagnostic testing in formative assessment. *Assessment in education, 22*(3), 324 - 343.

Wools, S. (2009). Is dit assessment kwalitatief goed genoeg? Over de ontwikkeling van een beoordelingsinstrument voor competentie assessment. *Examens, 4,* 10-14.

Young, V. M., & Kim, D. H. (2010). Using assessments for instructional improvement: A literature review. *Education Policy Analysis Archives, 18*(19), 1-37. Verkregen van: http://epaa.asu.edu/ojs/article/view/809

Auteurs

Nathalie A. M. Maassen en **Dorien den Otter** zijn werkzaam als junior-onderzoeker bij de Universiteit Twente. **Saskia Wools** is manager Prototyping CitoLab bij Cito. **Bas T. Hemker** is Toetsdeskundige bij Cito. **Gerard J. J. M. Straetmans** is lector Assessment bij Saxion Hogeschool en Toetsdeskundige bij Cito. **Theo J. H. M. Eggen** is bijzonder hoogleraar Psychometrie bij de Universiteit Twente, wetenschappelijk directeur bij RCEC en senior Toetsdeskundige bij Cito.

Correspondentieadres: T.J.H.M. Eggen. Universiteit Twente, Faculteit der Gedragswetenschappen, Secretariaat RCEC, Postbus 217, 7500 AE Enschede. rcec@utwente.nl

Abstract

Quality of assessments within reach: Review study of research and results of the quality of assessments

Educational tests and assessments are important instruments to measure a student's knowledge and skills. The question that is addressed in this review study is: "which aspects are currently considered as important to the quality of educational assessments?" Furthermore, it is explored how this information can be made available for both researchers and practitioners. Based on a systematic literature review, a conceptual framework was developed. The quality aspects in the framework were: reliability, generalizability, validity, the use of test and assessment results, and boundary conditions. The results were validated by focus groups. It was concluded that the different aspects of quality mentioned within articles were dependent on the authors' perspective on assessments. Perspectives differ based on purposes of assessments, phase in the assessment process an article addressed, and whether the article was theoretically or practically oriented. Overall, results show that the quality aspect reliability is discussed most frequently.