Rapportage

Kwaliteit van toetsen binnen handbereik

Een reviewstudie van onderzoek en onderzoeksresultaten naar de kwaliteit van toetsen

Nathalie Maassen, Dorien den Otter,
Saskia Wools, Bas Hemker,
Gerard Straetmans, Theo Eggen

September 2014 – RCEC, Cito

COLOFON

De literatuurreview *'Kwaliteit van toetsen binnen handbereik: Een reviewstudie van onderzoek naar kwaliteit van toetsen'* is gefinancierd door de NRO Programmaraad voor Praktijkgericht Onderwijsonderzoek.

Nathalie Maassen, RCEC, Universiteit Twente
Dorien den Otter, RCEC, Universiteit Twente
Saskia Wools, Cito
Bas Hemker, Cito
Gerard Straetmans, Saxion Hogescholen, Cito
Theo Eggen, RCEC, Universiteit Twente, Cito

www.rcec.nl

INHOUDSOPGAVE

SAMENVATTING

BELANG

Eén van de belangrijkste doelen van het onderwijs is leerlingen of studenten die kennis, attitude en vaardigheden te leren om actief te kunnen participeren in de samenleving. Een belangrijk hulpmiddel hierbij is het gebruik van toetsen. Toetsen kunnen helpen bij het certificeren (hebben studenten genoeg geleerd om de volgende stap in hun ontwikkeling te nemen?) en bij het leren (waar heeft de student meer onderwijs over nodig?). Vanwege deze belangrijke rollen is de kwaliteit van toetsing in alle geledingen van het onderwijs van belang: voor studenten, docenten, schoolleiders en beleidsmedewerkers.

PROBLEEMSTELLING

De afgelopen tijd hebben verschillende partijen, zowel binnen als buiten het onderwijs, vraagtekens gezet bij de in de onderwijspraktijk gerealiseerde kwaliteit van toetsing. Er wordt steeds meer geïnvesteerd in de kwaliteit van toetsing, waarbij men stuit op de vraag: wat is een goede toetskwaliteit? Vanwege de verschillende mogelijke invalshoeken is er een onduidelijk begrippenkader dat bepalend is voor de kwaliteit van een toets. Deze reviewstudie heeft daarom het doel de bestaande kennis en informatie over toetskwaliteit te verzamelen, te classificeren en beschikbaar te stellen voor verschillende doelgroepen. De centrale vraag is:

Wat beschouwt men op dit moment als kwaliteit van toetsen in het onderwijs?

METHODE

De reviewstudie bestaat uit vier fasen. In de eerste fase is op systematische wijze gezocht naar bruikbare wetenschappelijke en praktijkgerichte literatuur. In totaal zijn er 242 artikelen geselecteerd. In de tweede fase is er met behulp van experts een begrippenkader opgesteld. Dit begrippenkader vormde de basis voor de derde fase, waar de inhoud van de artikelen is gecodeerd. In de vierde fase is de data geanalyseerd en ter validering voorgelegd aan verschillende klankbordgroepen, bestaande uit docenten van alle onderwijssectoren.

Er is gestart met een beschrijvende analyse naar de frequentie van de kwaliteitsaspecten. Vervolgens zijn er specifieke analyses gedaan die ingaan op de factoren: doel van de toets, onderwijssector, rollen in de beoordeling, fase in de toetscyclus en de mate waarin de auteurs van een artikel onderzoek in de praktijk hebben uitgevoerd. Binnen deze factoren zijn verschillende groeperingen gemaakt, zoals de groepering in summatief en formatief bij de factor ‘doel van de toets’. Met behulp van een statistische toets is nagegaan of een specifieke groep afwijkt van het totaal aantal artikelen (inclusief de geselecteerde groep) of afwijkt van een andere specifieke groep.

RESULTATEN

Deze review heeft de onderzoekresultaten op het gebied van kwaliteit van toetsen in kaart kunnen brengen aan de hand van een begrippenkader met drie niveaus (Figuur 1, p. 13). Dit begrippenkader bestaat allereerst uit vijf hoofdcategorieën: betrouwbaarheid, generaliseerbaarheid, validiteit, gebruik toetsresultaat en randvoorwaarden. Vervolgens zijn er binnen elke hoofdcategorie meerdere subcategorieën te onderscheiden. De pijlen geven tot slot aan uit welke onderdelen de subcategorieën bestaan.

De 242 artikelen gaan vooral in op kwaliteitsaspecten van een toets en in mindere mate op kwaliteitsaspecten van toetsitems of toetsprogramma's. Uit de analyses blijkt dat de hoofdcategorie *betrouwbaarheid* het meest frequent voorkomt in de artikelen, evenals de subcategorie *objectiviteit* en de onderdelen *beoordelingsvoorschrift* en *deskundigheid van de beoordelaar*. De hoofdcategorieën *generaliseerbaarheid* en *randvoorwaarden* blijken het minst voor te komen. Het belang van de kwaliteitsaspecten verschilt tot slot afhankelijk van een aantal factoren.

- TOETSDOEL: Artikelen over summatieve toetsing tonen een hogere frequentie voor de kwaliteitsaspecten uit de subcategorie *validiteit* en *normering en cesuur*, terwijl artikelen over formatieve toetsen vooral nadruk leggen op de *consequentiële validiteit*.

- ONDERWIJSSECTOR: Er zijn geen verschillen in genoemde kwaliteitsaspecten tussen de onderwijssectoren. Uit zowel het literatuuronderzoek als de klankbordgesprekken blijkt wel dat de mate van aandacht voor toetskwaliteit verschilt van weinig aandacht in het primair onderwijs tot veel aandacht in het hoger onderwijs.

- ROLLEN IN DE BEOORDELING: Er is nauwelijks aandacht voor alternatieve beoordelingsmethodieken zoals peer-, co-, en self-assessment. Hoewel de enkele gevonden artikelen hierover zich richten op het hoger onderwijs, blijkt uit klankbordgesprekken dat er ook binnen deze sector geen aandacht voor is. Het is daardoor onduidelijk of er verschillen zijn in kwaliteitsaspecten bij deze beoordelingsmethodieken.

- TOETSCYCLUS: Het merendeel van de artikelen gaat over de fase van afname en beoordeling. In deze fase blijkt *betrouwbaarheid*, en de daarmee samenhangende begrippen *objectiviteit*, *deskundigheid van de beoordelaar* en het gebruik van een *beoordelingsschema* vaker voor te komen dan in de constructie- en evaluatiefase.

- IN DE PRAKTIJK: Zowel in empirisch onderzoek als in theoretische beschouwingen is er veel aandacht voor *betrouwbaarheid* en aspecten die daarmee samenhangen, zoals *objectiviteit* en *deskundigheid van de beoordelaar*. Het kwaliteitsaspect *validiteit* wordt vooral in theoretische beschouwingen vaak genoemd, terwijl het in empirische onderzoeken nauwelijks voorkomt.

CONCLUSIE EN DISCUSSIE

Uit de reviewstudie is gebleken welke aspecten een rol spelen bij toetskwaliteit in het onderwijs. Het belang van de kwaliteitsaspecten verschilt afhankelijk van een aantal factoren. Uiteraard zijn de resultaten mede bepaald door de keuze van de zoektermen, de gekozen indeling binnen de factoren en het feit dat een veelgenoemd aspect niet direct relatie hoeft te hebben met de waarde van het kwaliteitsaspect. Toch geeft dit onderzoek inzicht in wat men op dit moment beschouwt als toetskwaliteit en laat het zien waar hiaten in het onderzoek naar dit thema liggen. Op basis van deze resultaten kan vervolgonderzoek opgezet worden en kan de werkwijze in de praktijk verbeterd worden, zodat de manier waarop belangrijke beslissingen over studenten tot stand komen van goede kwaliteit zijn.

AANBEVELINGEN

TOETSDOEL: Het is voor docenten van belang dat zij eerst het toetsdoel vaststellen. Aan de hand daarvan kan worden vastgesteld met welke kwaliteitsaspecten men extra rekening moet houden, aangezien er verschillende kwaliteitsaspecten een rol spelen bij de verschillende toetsdoelen.

ONDERWIJSSECTOR: Hoewel het erop lijkt dat in alle onderwijssectoren dezelfde kwaliteitsaspecten een rol spelen, zal toekomstig onderzoek en praktijkgerichte hulp moeten worden aangepast aan de desbetreffende sector. Door het verschil in aandacht voor de kwaliteit van toetsen zijn er verschillende behoeften naar informatie over de toetskwaliteit.

TOETSCYCLUS: De fase in de toetscyclus is bepalend voor welke kwaliteitsaspecten van belang zijn. 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?

IN DE PRAKTIJK: Validiteit komt nauwelijks voor in praktijkgerichte artikelen, terwijl dit in theoretische beschouwingen veelvuldig wordt beschreven. Hoe kunnen bedreigingen van het goed meten van specifieke vaardigheden concreet worden gemaakt en hoe worden zij opgelost? Of hoe kunnen individuele bijdragen in groepsprestaties worden beoordeeld?

ONDERZOEKSNIVEAU: Het meeste onderzoek is gedaan op toetsniveau. Er kan ook meer gedetailleerd worden ingezoomd op de items of juist worden uitgezoomd naar het toets-programma, omdat toetsen of metingen vaak gecombineerd worden om tot een beoordeling te komen. Daarnaast geldt dat toetsen verweven kunnen worden in het leerproces door ze meer formatief in te zetten. Hoe kan een toetsprogramma zo effectief mogelijk worden ingericht om dit op een juiste manier uit te voeren?

1. INLEIDING

Eén van de belangrijkste doelen van het onderwijs is studenten kennis, attitude en vaardigheden te leren. Daarmee worden zij in staat gesteld om actief en kritisch te participeren in vervolgonderwijs, beroep en andere maatschappelijke contexten. Voor de realisatie van dat doel speelt de inzet van toetsen een belangrijke rol. Toetsen zijn nodig voor het certificeren van studenten: beheersen zij de doelstellingen van het onderwijsprogramma voldoende om dit onderdeel definitief af te sluiten? Daarnaast zijn toetsen nodig 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. Studenten zijn bijvoorbeeld gebaat bij kwalitatief goede toetsen omdat die hen de mogelijkheid bieden hun kennis en vaardigheden te demonstreren. Voor docenten zijn kwalitatief goede toetsen van belang omdat die uitspraken mogelijk maken over wat de studenten kennen en kunnen, zodat zij de juiste beslissingen over studenten en/of onderwijsprogramma's kunnen nemen. Schoolleiders of beleidsmedewerkers zullen erop gericht zijn om een kwalitatief goed toetsprogramma samen te stellen, om het eindniveau van studenten, en daarmee het civiel effect van het diploma, te borgen.

De afgelopen tijd hebben verschillende partijen, zowel binnen als buiten het onderwijs, vraagtekens gezet bij de in de onderwijspraktijk gerealiseerde kwaliteit van toetsing in alle sectoren van 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). Er wordt steeds meer geïnvesteerd in de kwaliteit van toetsing. Daarbij komen echter veel vragen op: Wat wordt er onder de kwaliteit van een toets verstaan? Waar hangt deze kwaliteit van af? Hoe kan deze kwaliteit worden bereikt? In deze review zal de kwaliteit van toetsen centraal staan.

2. THEORETISCH KADER

Het belang van kwalitatief goed toetsen is helder, maar het blijkt complex om te bepalen wat kwaliteit inhoudt. Er zou gezegd kunnen worden dat er in Nederland ongeveer 16 miljoen deskundigen zijn op toetsgebied. Iedereen heeft ervaringen met toetsen en een eigen idee over wat goede en slechte toetsen zijn (Eggen, 2009). Kwaliteit van toetsen is een begrip dat vanuit verschillende invalshoeken benaderd kan worden. De verschillende perspectieven leiden ertoe dat er een onduidelijk begrippenkader is, wat de onduidelijkheid versterkt over wat de kwaliteit van een toets bepaalt. Enerzijds worden er verschillende termen gebruikt voor hetzelfde begrip, anderzijds worden dezelfde begrippen gehanteerd maar bedoelt men iets anders.

Het gevolg van deze onduidelijkheid is dat de kwaliteit van een toets niet makkelijk is vast te stellen en kennisdeling bemoeilijkt kan worden door de diversiteit aan begrippen. De diversiteit blijkt bijvoorbeeld uit het feit dat er verschillende evaluatiesystemen bestaan 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). Deze systemen bevatten veel overeenkomstige criteria, maar verschillen onderling ook, onder andere voor wat betreft het prescriptieve karakter van de systemen en de strengheid van de eisen (Wools, 2009). De verschillen van inzicht kunnen het gevolg zijn van een aantal factoren die in dit hoofdstuk zullen worden besproken. Deze factoren zijn niet volledig onafhankelijk van elkaar en kunnen elkaar dus beïnvloeden.

2.1 DOEL VAN DE TOETS

Onderwijskundige meetinstrumenten worden vanuit verschillende behoeften ingezet. De doelen om een toetsinstrument af te nemen lopen dan ook 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 daarentegen geven feedback 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. Eenzelfde toets kan voor verschillende doeleinden worden gebruikt (Gioka, 2009; Taras, 2005). Het is begrijpelijk dat er (deels) verschillende kwaliteitsaspecten worden aangelegd voor toetsen met een summatieve en formatieve functie. Bij een voortgangstoets voor studenten zal de focus bijvoorbeeld op andere kwaliteitsaspecten kunnen liggen dan bij een toets die ingezet wordt voor certificerings-doeleinden.

2.2 ONDERWIJSSECTOR

Een andere belangrijke factor betreft verschillen in de toetscultuur binnen de verschillende sectoren van het onderwijs: primair onderwijs, voortgezet onderwijs, beroepsgericht onderwijs en hoger onderwijs. In de genoemde sectoren vervullen toetsen verschillende doelen, worden verschillende vaardigheden gemeten (Straetmans, 2006), is de wijze van toetsing verschillend (papier/digitaal, praktijk, simulatie) en is de rol die docenten bij het toetsproces vervullen verschillend. Het is daarom voorstelbaar dat de aspecten die een rol spelen bij kwaliteit van toetsen verschillen per sector.

2.3 ROLLEN IN DE BEOORDELING

Volgens traditioneel model kan de docent zelf de studenten beoordelen. Recent is er aandacht gekomen voor alternatieve beoordelingsmodellen, waarbij ook anderen dan de docent de rol van beoordelaar op zich nemen. 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 kwaliteits-aspecten een rol spelen dan in het traditionele model met de docent als beoordelaar (Ploegh, Tillema & Segers, 2009).

2.4 TOETSCYCLUS

Welke aspecten van toetskwaliteit relevant zijn voor een gebruiker is ook afhankelijk van de fase in de toetscyclus waarbij iemand betrokken is (Tillema, Leenknecht & Segers, 2011). Zo zijn er bij de selectie van een bestaand instrument vaak andere aspecten relevant dan bij de constructie van items of opdrachten. Ook zijn er verschillende kwaliteitsaspecten van belang bij de toetsafname en bij de verwerking en analyse van gegevens, of bij de rapportage van de resultaten naar de belanghebbende. In een breder perspectief: het opstellen van toetsbeleid vraagt om andere kwaliteitsaspecten dan het evalueren van een enkele toets.

2.5 IN DE PRAKTIJK

Er is verschil in de manier waarop onderzoek naar toetskwaliteit is gedaan. Dit leidt tot verschillende perspectieven en daardoor verschillende inzichten op toetskwaliteit. Er kan verondersteld worden dat artikelen met onderzoek in de praktijk de werkelijkheid meer benaderen dan artikelen die het concept enkel theoretisch beschouwen. Er kan onderscheid gemaakt worden tussen direct onderzoek, indirect onderzoek en theoretische beschouwingen. Onder direct onderzoek worden artikelen verstaan waarbij observaties of toetsanalyses zijn uitgevoerd in de school (o.a. Calhoon, Greenberg & Hunter, 2010; Gioka, 2006). Bij indirect onderzoek wordt gebruik gemaakt van interviews of vragenlijsten om zicht te krijgen op de praktijk (o.a. Baartman, Bastiaens, Kirschner & Van der Vleuten, 2007; Maclellan, 2004). In theoretische beschouwingen wordt geen onderzoek in praktijksituaties uitgevoerd (o.a. Newton, 2012; White, 2007) maar wordt de toetskwaliteit vanuit een theoretisch kader beschreven.

Deze factoren laten zien dat de aspecten die samenhangen met kwaliteit en kwaliteitsborging van toetsen talrijk zijn en dat er vanuit veel verschillende invalshoeken naar gekeken kan worden. Als gevolg van deze complexiteit kan het lastig zijn om bij de kwaliteitsbeoordeling van toetsen de juiste beoordelingskaders te selecteren en te gebruiken. Het doel van deze reviewstudie is daarom 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 centrale vraag 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. Dit literatuuronderzoek heeft zich gericht op het verzamelen van onderzoeksresultaten op het gebied van kwaliteit van toetsen. In hoofdstuk 3 wordt de methode in vier stappen beschreven. Hoofdstuk 4 beschrijft de resultaten, waarbij wordt ingegaan op de voorkomende kwaliteitsaspecten in de literatuur en op welke manier de factoren het belang van deze kwaliteitsaspecten beïnvloeden. In hoofdstuk 5 worden conclusies getrokken en aanbevelingen voor toekomstig onderzoek en de praktijk gedaan.

3. METHODE

In deze reviewstudie kunnen vier fasen worden onderscheiden: (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.

3.1 SYSTEMATISCH LITERATUURONDERZOEK

In de eerste fase is een systematisch literatuuronderzoek uitgevoerd om de huidige inzichten ten aanzien van de kwaliteit van toetsen te verzamelen. Hiervoor is zowel wetenschappelijke literatuur als praktijkgericht onderzoek geraadpleegd.

Uit wetenschappelijke databases zijn 91 peer-reviewed artikelen vanaf het jaar 2000 geselecteerd. Deze artikelen zijn verkregen door een zoekopdracht in de databases *ERIC, PsychINFO, Web of Science* en *Scopus* in maart 2014. Hierbij is de volgende combinatie van zoektermen gehanteerd: (quality standard* OR quality guideline* OR quality criteri* OR evaluation criteri*) AND (educational test* OR student evaluation* OR educational assessment* OR classroom assessment*). Inclusiecriteria waren: 1) het bevatten van kwaliteits-aspecten van een toets en 2) het betrekking hebben op het onderwijs. Verder moest het artikel zich richten op het toetsprogramma, de toets zelf of het itemniveau van de toets. Op basis van deze 91 artikelen zijn nog 56 artikelen toegevoegd door middel van een 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. De bronnen hiervoor waren vaktijdschriften (*Didactief, Examens, OnderwijsInnovatie, Tijdschrift voor Hoger Onderwijs* en *Toets!)*, mastertheses en proefschriften van meerdere universiteiten en informatie van een aantal lectoren werkzaam op het terrein van onderwijskundig meten. Tot slot is er gezocht naar geschikte literatuur in bibliotheken. Samen leverde dit uiteindelijk 95 relevante artikelen op, waarvan 58 artikelen uit vaktijdschriften, 22 artikelen van lectoren, acht mastertheses en proefschriften en zeven artikelen uit bibliotheken. In totaal zijn er 242 artikelen verzameld.

3.2 VAN EXPERTPANEL NAAR BEGRIPPENKADER

In de tweede fase van de reviewstudie werd een begrippenkader geconstrueerd met behulp van een expertpanel. Dit expertpanel bestond uit zes experts die werkzaam zijn op het gebied van toetsing, verbonden aan verschillende lectoraten en toetsinstanties. Er is een vragenlijst voorgelegd waarbij werd gevraagd naar de bekendheid met, en het belang van, verschillende kwaliteitsaspecten. Deze kwaliteitsaspecten zijn gevonden met behulp van een eerste globale analyse van de geselecteerde literatuur.

Op basis van de antwoorden op de vragenlijst is een begrippenkader opgesteld waarin alle kwaliteitsaspecten verwerkt zijn. Nadat een volgende selectie van artikelen werd gecodeerd met behulp van dit begrippenkader, is het begrippenkader aangepast om er zeker van te zijn dat alle relevante kwaliteitsaspecten zijn opgenomen.

Het definitieve begrippenkader (Figuur 1, p. 13) bestaat uit vijf hoofdcategorieën: betrouwbaarheid, generaliseerbaarheid, validiteit, gebruik van het toetsresultaat en randvoorwaarden. De verschillende kwaliteitsaspecten die in de literatuur werden gevonden, zijn ondergebracht binnen deze hoofdcategorieën en worden aangeduid als subcategorieën en onderdelen. De gehanteerde definities van deze kwaliteitsaspecten zijn weergegeven in Appendix A.

3.3 CODEREN VAN LITERATUUR

In de derde fase is de inhoud van alle artikelen gecodeerd. Voor elk artikel is bepaald op welke kwaliteitsaspecten het artikel betrekking heeft. Eén artikel kan betrekking hebben op meerdere kwaliteitsaspecten. Als een artikel een bepaald kwaliteitsaspect noemde, kreeg dit aspect waarde '1'. Als het kwaliteitsaspect niet werd genoemd kreeg het waarde '0'.

Deze codering maakt echter nog geen onderscheid in de drie niveaus van toetskwaliteit uit het begrippenkader. Daarom is de codering omgezet in scores, waarbij wel rekening wordt gehouden tussen hoofdcategorieën, subcategorieën en onderdelen. Deze hercodering ging als volgt: Wanneer een kwaliteitsaspect, bijvoorbeeld *meerdere beoordelaars*, in een artikel voorkomt, krijgt dit onderdeel waarde '1'. De subcategorie waar dit onderdeel onder valt, in dit geval de subcategorie *objectiviteit*, krijgt hierdoor tevens waarde '1'. Tot slot krijgt ook de hoofdcategorie waar dit onderdeel onder valt, in dit geval de hoofdcategorie *betrouwbaarheid*, waarde '1'.

In veel artikelen worden meerdere kwaliteitsaspecten genoemd, bijvoorbeeld *meerdere beoordelaars* én *beoordelingsvoorschrift*. De bijbehorende subcategorie *objectiviteit* en hoofdcategorie *betrouwbaarheid* krijgen in dat geval toch waarde '1' en niet waarde '2' of '3'. Zo is vergelijking tussen verschillende hoofdcategorieën met een verschillende hoeveelheid subcategorieën, en subcategorieën met een verschillend aantal onderdelen mogelijk. Om de betrouwbaarheid van het coderen van de artikelen te borgen, zijn een aantal artikelen door drie beoordelaars gecodeerd, waarbij een overeenstemming kon worden vastgesteld van 0.6.

3.4 ANALYSEREN VAN DATA EN VALIDEREN MET KLANKBORDGROEPEN

De kwaliteitsaspecten op de drie niveaus (hoofdcategorie, subcategorie en onderdeel) zijn op verschillende manieren geanalyseerd. Allereerst is er een beschrijvende analyse gedaan met betrekking tot de frequentie van de kwaliteitsaspecten bij het totaal aantal artikelen (n=242). Vervolgens is er gekeken naar verschillen in deze frequentie tussen wetenschappelijke en praktijkgerichte artikelen.

Daarna zijn er analyses gedaan die specifiek ingingen op de genoemde factoren uit hoofdstuk 2: het toetsdoel, de onderwijssectoren, de rollen in de beoordeling, de fase in de toetscyclus en de mate waarin een onderzoek in de praktijk is uitgevoerd. Met behulp van de Pearson Chi-Kwadraat toets is getoetst of er verschillen tussen groepen waren. Met deze toets wordt de aanname dat beide groepen aan elkaar gelijk zijn getoetst. Hierbij is telkens één van beide groepen als uitgangspunt genomen. Deze groep vormt de basisgroep. Op grond van deze basisgroep werd een verwachting gesteld, waarna de andere groep hiertegen kon worden afgezet. Zo kon er worden getoetst of er aan de verwachting werd voldaan. Omdat de verwachting dus van de basisgroep afhangt, zijn telkens beide groepen als basisgroep in de analyse benoemd en zijn er in feite twee analyses gedaan.

Om de kwaliteitsaspecten van een bepaalde groep artikelen te vergelijken met de rest van de artikelen, is deze groep vergeleken met het totaal aantal artikelen, inclusief de geselecteerde groep. Er is besloten om de geselecteerde groep niet te verwijderen uit het totaal aantal artikelen. Zo is de hoger onderwijssector vergeleken met alle onderwijssectoren, inclusief het hoger onderwijs. De onderwijssectoren vormen namelijk één groep. Wanneer de sector hoger onderwijs zou worden vergeleken met de onderwijssectoren exclusief hoger onderwijs, dan wordt er vergeleken met een niet-complete groep. Dit geeft geen relevante informatie. 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 dus nog sterker moet afwijken voordat er daadwerkelijk significante verschillen gevonden zullen worden.

Om de validiteit van de resultaten verder 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. Artikelen die geen waarde hebben bij een variabele omdat er in het artikel geen expliciete opmerking over gemaakt is, zijn niet meegenomen in de analyse. Artikelen die op meerdere groepen betrekking hadden zijn ook niet meegenomen. Ter illustratie: wanneer een artikel over formatieve én summatieve toetsing ging, is dit artikel niet meegenomen in de analyse waar formatieve en summatieve toetsing met elkaar worden vergeleken. De kwaliteitsaspecten uit dit artikel zijn namelijk gezamenlijk geclassificeerd, waardoor er geen onderscheid gemaakt kon worden tussen kwaliteitsaspecten die bij summatieve of juist bij formatieve toetsen genoemd werden. Als deze artikelen wel zouden worden meegenomen, zullen de aspecten met betrekking tot summatieve toetsen interfereren in de uitspraak over aspecten bij formatieve toetsen, en andersom.

De resultaten van de codering en het gecreëerde overzicht van de resultaten zijn ter validering voorgelegd aan verschillende klankbordgroepen bestaande uit docenten werkzaam in verschillende sectoren van het onderwijs. Met behulp van een semigestructureerd interview werd nagegaan in hoeverre de resultaten uit het literatuuronderzoek in de praktijk werden herkend (Appendix B).

4. RESULTATEN EN VERKLARINGEN

4.1 EXPERTPANEL

Uit de vragenlijst die bij zes experts op het gebied van toetsing zijn afgenomen, bleek dat de experts verschillende meningen hebben over de kwaliteit van toetsen. Dit heeft zich onder andere geuit in zes verschillende indelingen van de kwaliteitsaspecten in hoofdcategorieën.

Tevens waren er diverse kwaliteitsaspecten die zowel bij de top tien van meest belangrijke aspecten als bij de top tien van minst belangrijke aspecten werden genoemd. *Objectiviteit*, *meerdere toetsmomenten* en *feedback* zijn hier voorbeelden van. Kennelijk hebben de bevraagde experts elk vanuit hun eigen perspectief gerespondeerd naar toetskwaliteit, waardoor er geen directe eenduidigheid was over de indeling, de hiërarchie en het belang van de kwaliteitsaspecten.

4.2 BEGRIPPENKADER

Op basis van de resultaten van het expertpanel en de geselecteerde literatuur is er een nieuw begrippenkader ontwikkeld om zo de veelheid van kwaliteitsaspecten samen te vatten (Figuur 1). In dit begrippenkader is het onderscheid tussen de hoofdcategorieën, subcategorieën en onderdelen zichtbaar.

Er zijn vijf hoofdcategorieën: betrouwbaarheid, generaliseerbaarheid, validiteit, gebruik toetsresultaat en randvoorwaarden. Deze laatste categorie bevat voorwaarden om te komen tot toetskwaliteit.

Binnen elke hoofdcategorie zijn subcategorieën te onderscheiden. Deze subcategorieën zijn in gekleurde kaders weergegeven in de linker kolom van elke hoofdcategorie. Zo is binnen de hoofdcategorie *betrouwbaarheid* 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* 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. In Appendix A is het gehele begrippenkader toegelicht met de in dit onderzoek gehanteerde definities.

Betrouwbaarheid is de mate waarin de scores op een toets consistent, nauwkeurig en reproduceerbaar zijn. In dat geval is het meetresultaat vrij van meetfouten.

Voorbeeld: 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.

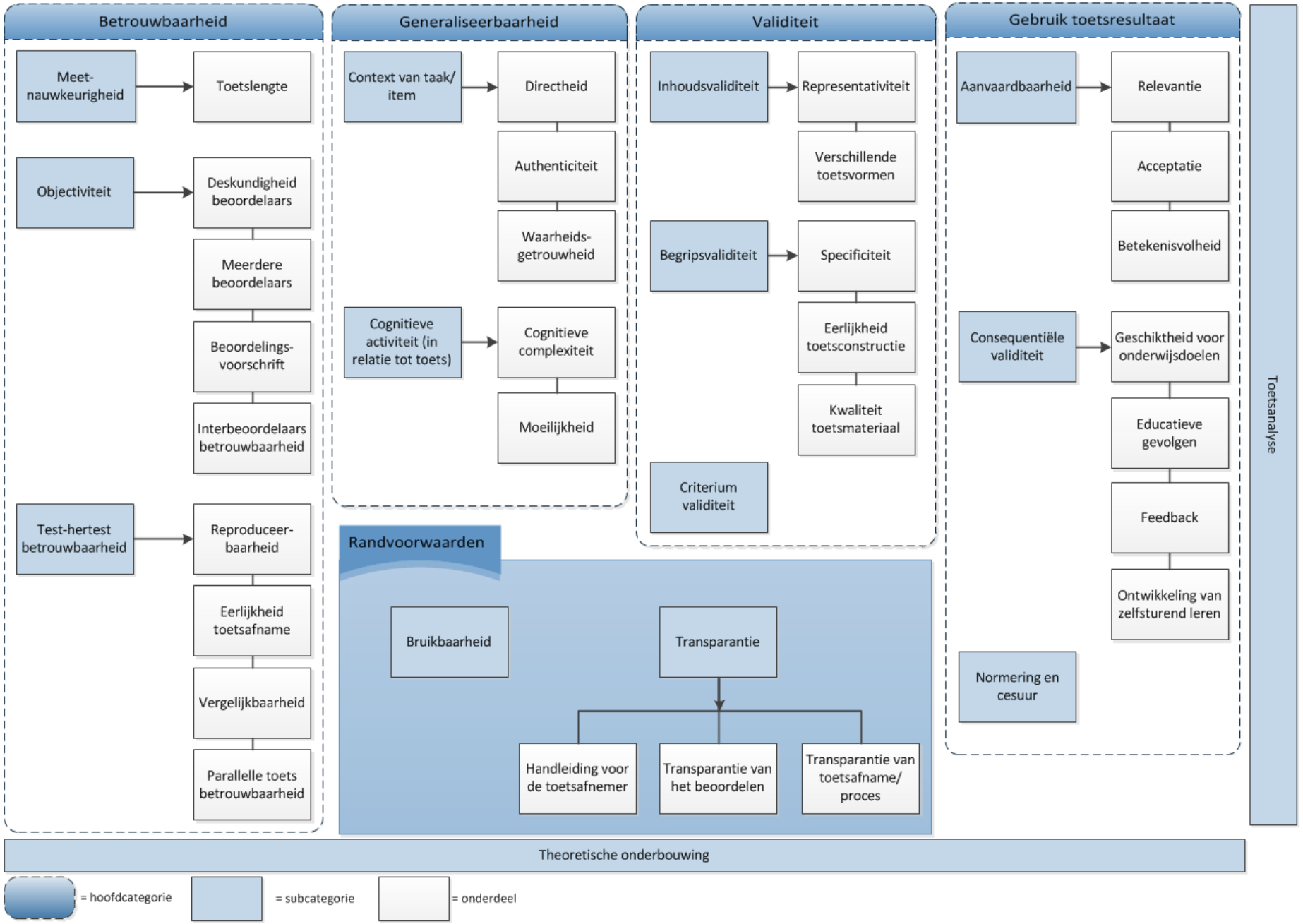
Voorbeeld: Als een student Verpleegkunde aantoont opgaven te beheersen die te maken hebben met het toedienen van vloeibare medicatie, mag er dan vanuit worden gegaan dat deze student voldoende vaardig is op het gebied van verpleegkundig rekenen?

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?

Voorbeeld: 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.

Voorbeeld: Als de student 50 punten heeft gehaald, krijgt hij een onvoldoende. Hij krijgt hulp op de onderdelen die hij niet goed heeft gemaakt.



Figuur 1. Aspecten van toetskwaliteit volgens de geselecteerde literatuurbronnen.

4.3 ANALYSES

De kwaliteitsaspecten op de drie niveaus van hoofdcategorieën, subcategorieën en onderdelen zijn op verschillende wijzen geanalyseerd. In paragraaf 4.3.1 is de beschrijvende analyse van de data gegeven. In de daaropvolgende paragrafen zijn de analyses beschreven die specifiek ingingen op de genoemde factoren uit hoofdstuk 2: het toetsdoel (§ 4.3.2), de onderwijssectoren (§ 4.3.3), de rollen in de beoordeling (§ 4.3.4), de toetscyclus (§ 4.3.5) en de mate waarin onderzoek in de praktijk is uitgevoerd (§ 4.3.6).

4.3.1 ALGEMEEN

Er zijn 242 artikelen beschouwd waarin de kwaliteit van toetsen centraal staat. In Tabel 1 is voor elke factor weergegeven hoeveel artikelen er zijn die ingaan op deze factor en op hoeveel aspecten er in deze artikelen is ingegaan. In de laatste kolom is het gemiddeld aantal kwaliteitsaspecten per artikel zichtbaar. Hieruit blijkt dat de artikelen gemiddeld op ongeveer vijf kwaliteitsaspecten ingaan.

Tabel 1. *Frequentie van aantal artikelen, kwaliteitsaspecten en de onderlinge verhouding voor elke factor.*

Factor	**Aantal artikelen**		**Aantal kwaliteitsaspecten**		**Gemiddeld aantal kwaliteitsaspecten per artikel**
	N	**%**	**N**	**%**	
Doel van de toets					
Formatief	24	9,9%	112	8,7%	4,7
Summatief	64	26,4%	316	24,7%	4,9
Beide	55	22,7%	382	29,8%	6,9
Niet gespecificeerd	99	40,9%	471	36,8%	4,8
Onderwijssector					
Primair onderwijs	9	3,7%	35	2,7%	3,9
Voortgezet onderwijs	14	5,8%	72	5,6%	5,1
Beroepsgericht onderwijs	23	9,5%	181	14,1%	7,9
Hoger onderwijs	102	42,1%	450	35,1%	4,4
Meerdere sectoren	8	3,3%	66	5,2%	8,3
Niet gespecificeerd	86	35,5%	477	37,2%	5,5
Rollen in de beoordeling					
Docent beoordelaar	105	43,4%	481	37,5%	4,6
Peer-assessment	12	5,0%	52	4,1%	4,3
Self-assessment	5	2,1%	23	1,8%	4,6
Co-assessment	0	0,0%	-	-	-
Meerdere rollen	34	14,0%	191	14,9%	5,6
Niet gespecificeerd	86	35,5%	534	41,7%	6,2
Toetscyclus					
Constructiefase	16	6,6%	79	6,2%	4,9
Afname en beoordelingsfase	116	47,9%	438	34,2%	3,8
Evaluatiefase	85	35,1%	605	47,2%	7,1
Meerdere fasen	15	6,2%	91	7,1%	6,1
Niet gespecificeerd	10	4,1%	68	5,3%	6,8
In de praktijk					
Direct onderzoek	51	21,1%	203	15,8%	4,0
Indirect onderzoek	45	18,6%	273	21,3%	6,1
Theoretische beschouwing	146	60,3%	805	62,8%	5,5
Totaal	242	100%	1281	100%	5,3

ALGEMENE ANALYSE

De beschrijvende algemene analyse laat zien dat de hoofdcategorie *betrouwbaarheid* het meest frequent voorkomt in de artikelen (Tabel 2). Ongeveer 28% van de genoemde kwaliteitsaspecten valt onder deze hoofdcategorie. Op het niveau van de subcategorieën kwam *objectiviteit* (14,4%) het meest voor en zijn de meest genoemde onderdelen (niet getoond in deze tabel) *beoordelingsvoorschrift* (9,8%) en *deskundigheid beoordelaar* (7,8%). Beiden vallen onder *betrouwbaarheid*. Uit Tabel 2 blijkt tevens dat twee andere veelgenoemde subcategorieën *consequentiële validiteit* (10,0%) en *transparantie* (9,4%) zijn. Hierbij moet overigens opgemerkt worden dat *consequentiële validiteit* gaat over het gebruik van de toets. Daardoor past dit aspect beter onder die hoofdcategorie en niet –ondanks dat de naam anders doet vermoeden– onder de hoofdcategorie validiteit. Bij *transparantie* ging het vaak over transparantie van de *scoring* van een toets (6,8%).

De hoofdcategorieën *generaliseerbaarheid* en *randvoorwaarden* werden het minst vaak genoemd. Daarnaast waren er een aantal onderdelen die nauwelijks voorkwamen, zoals *relevantie, toetslengte, parallelle toets betrouwbaarheid* en *directheid*. Deze minder vaak genoemde kwaliteitsaspecten kunnen enerzijds als minder belangrijk worden gezien. Anderzijds is het mogelijk dat deze aspecten moeilijker te onderzoeken zijn.

Tabel 2. *Frequentie van het aantal kwaliteitsaspecten in de hoofd- en subcategorieën.*

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%

Betrouwbaarheid is de meest voorkomende hoofdcategorie. Vooral over de bijbehorende subcategorie objectiviteit en bijbehorende onderdelen beoordelingsvoorschrift en deskundigheid van de beoordelaar is veel onderzocht en geschreven.

WETENSCHAPPELIJKE EN PRAKTIJKGERICHTE ARTIKELEN

Om de twee soorten artikelen in dit onderzoek –wetenschappelijke en praktijkgerichte (niet-wetenschappelijke) artikelen– te analyseren, is er uitgegaan van de verwachting dat de frequentie van de kwaliteitsaspecten van beide groepen artikelen aan elkaar gelijk zijn. Met behulp van de Pearson Chi-Kwadraat toets is geanalyseerd of aan deze verwachting is voldaan.

De twee soorten artikelen verschilden niet op hoofdcategorieniveau (p=.239 en p=.083). Op subcategorieniveau verschilden ze wel van elkaar (p<.001). De subcategorieën *betrouwbaarheid* en *inhoudsvaliditeit* werden namelijk vaker in praktijkgerichte artikelen genoemd dan verwacht op basis van de wetenschappelijke artikelen. Dit zou te verklaren kunnen zijn door het schijnbare dilemma dat in de praktijk heerst (Sijtsma, 2011): moeten toetsen een zinvolle inhoud hebben (inhoudsvaliditeit) of een betrouwbare score (betrouwbaarheid)? De balans hierin blijkt in de praktijk lastig te bereiken, waardoor veel onderzoeken mogelijk ingaan op deze twee factoren.

De subcategorieën *consequentiële validiteit* en *generaliseerbaarheid* werden daarentegen vaker genoemd in wetenschappelijke artikelen dan verwacht werd op basis van de praktijkgerichte artikelen. *Consequentiële validiteit* gaat over de gewenste en ongewenste effecten van een toets op het leren van studenten en het lesgeven van de docent. In 2004 was dit nog een relatief nieuw begrip (Segers, 2004). Het is mogelijk dat dit aspect nog niet geïmplementeerd is in de praktijk, terwijl er wel aandacht aan wordt gegeven in de wetenschap. Het is ook mogelijk dat men in de praktijk juist al (on)bewust rekening houdt met de consequentiële validiteit van een toets, waardoor dit aspect niet vaak voorkomt in praktijkgerichte artikelen. Wat betreft de subcategorie *generaliseerbaarheid*: dit zou ook een theoretisch begrip kunnen zijn wat niet, of in andere termen (zoals de term representativiteit), voorkomt in praktijkgerichte artikelen.

Op het meest specifieke niveau, de onderdelen, werden eveneens significante verschillen gevonden tussen wetenschappelijke en praktijkgerichte artikelen (p<.001). In praktijkgerichte artikelen werden diverse kwaliteitsaspecten vaker genoemd dan verwacht werd op basis van de wetenschappelijke artikelen, zoals de aspecten *deskundigheid van de beoordelaar* en *representativiteit*. Deze sluiten goed aan op de in de praktijkgerichte artikelen significant vaker genoemde subcategorieën *betrouwbaarheid* en *inhoudsvaliditeit*. Ook in wetenschappelijke artikelen werden diverse kwaliteitsaspecten vaker genoemd dan verwacht op basis van de praktijkartikelen, zoals de aspecten *beoordelingsvoorschrift*, *educatieve gevolgen*, *eerlijkheid van constructie* en *eerlijkheid van afname.* Hier sluit enkel het aspect *educatieve gevolgen* aan op de eerder frequent genoemde subcategorie *consequentiële validiteit.*

Wetenschappelijke en praktijkgerichte artikelen verschilden niet significant op hoofdcategorieniveau. De subcategorieën en onderdelen lieten echter wel verschillen tussen de twee soorten artikelen zien. Wetenschappelijke artikelen noemden vaker aspecten als eerlijkheid, consequentiële validiteit en generaliseerbaarheid, terwijl praktijkgerichte artikelen betrouwbaarheid en inhoudsvaliditeit vaker noemden.

4.3.2 DOEL VAN DE TOETS

Van de 242 artikelen zijn er 88 die het doel van de toets, **summatief** (*n*=64) óf **formatief** (*n*=24), hebben gespecificeerd. Het merendeel hiervan heeft zich dus gericht op summatieve toetsing.

De hoofdcategorieën van deze artikelen over summatieve en formatieve toetsdoelen bleken niet significant verschillend te zijn van het totaal aantal artikelen (respectievelijk p=.384 en p=.705). Artikelen over summatieve toetsing wijken op hoofdcategorieniveau wel significant af van artikelen over formatieve toetsing (p<.001). Hierbij zijn de artikelen over summatieve toetsing als basisgroep genomen. De hoofdcategorie *validiteit* werd vaker genoemd in artikelen over summatieve toetsing. 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. Hierdoor is het belangrijk dat de toets datgene meet wat getoetst moet worden, zodat er een betekenisvolle interpretatie van de toetsscore gegeven kan worden (Tanilon, Segers, Vedder & Tillema, 2009).

Tevens bleek dat de artikelen over summatieve toetsing significant andere subcategorieën noemden dan het totaal aantal artikelen (p=.004) en de artikelen over formatieve toetsing (p<.001). Zo werd de subcategorie *normering en cesuur* vaker genoemd in artikelen over summatieve toetsing. Aangezien er belangrijke beslissingen worden genomen op grond van de toetsscores, is het van belang dat de normering op een juiste manier tot stand komt en wordt gebruikt (Van Berkel, 2004; Dalbert, Schneidewind & Saalbach, 2007).

Consequentiële validiteit en *bruikbaarheid* werden daarentegen minder vaak genoemd. Wellicht zijn dit aspecten die meer met formatieve toetsing te maken hebben. Dit is echter niet vast te stellen met de uitgevoerde analyses, aangezien er relatief weinig artikelen over formatieve toetsing gingen. Toch zijn er enige trends waar te nemen. Zo blijkt een vijfde van de genoemde aspecten in de artikelen over formatieve toetsing te vallen onder de subcategorie *consequentiële validiteit*. Uit verschillende artikelen over formatieve toetsing blijken aspecten van consequentiële validiteit, zoals feedback (Black & Wiliam, 1998; Gioka, 2006), ontwikkeling van zelfsturend leren (Nieweg, 2002) en educatieve gevolgen (Young & Kim, 2010) inderdaad van belang te zijn bij formatieve toetsen.

> *Artikelen over summatieve en formatieve toetsdoelen tonen verschillende aspecten waar nadruk op werd gelegd: validiteit en normering en cesuur werden bij artikelen over summatieve toetsdoelen frequent genoemd, terwijl bij artikelen over formatieve toetsdoelen consequentiële validiteit vaak voorkwam. Dit strookt met de opvatting dat de kwaliteit van een toets afhangt van het doel van de toets: niet elk doel vraagt om dezelfde kwaliteitseisen.*

4.3.3 ONDERWIJSSECTOR

Een andere factor betreft verschillen in de toetscultuur tussen de verschillende sectoren van het onderwijs. In dit onderzoek is onderscheid gemaakt tussen primair onderwijs, voortgezet onderwijs (vmbo-theorie, havo en vwo), beroepsgericht onderwijs (vmbo-praktijk en mbo) en hoger onderwijs (hbo en wo). Er waren 148 artikelen waarin de onderwijssector werd gespecificeerd.

Er zijn slechts negen artikelen van toepassing op het **primair onderwijs**. Dit is te weinig om te analyseren of deze sector specifieke kwaliteitsaspecten kent. Gesprekken met de klankbordgroep bevestigden dat er in deze sector minimale aandacht is voor toetsen. Leerkrachten uit het primair onderwijs ontwerpen zelden zelf een toets, maar gebruiken meestal methodegebonden toetsen of toetsen die door externen gemaakt worden, zoals Cito-toetsen. Deze toetsen worden wel vaak door de leerkrachten zelf nagekeken.

De klankbordgroep vond *validiteit* en vooral *inhoudsvaliditeit* een belangrijk kwaliteitsaspect: vormen de opgaven/opdrachten een goede afspiegeling van de doelstellingen die verworven moeten worden? Ook werd de overeenkomst tussen de vorm waarin de opgaven wordt aangeboden en de vorm van de (oefen)opgaven tijdens de instructie belangrijk gevonden. De toetsen worden in de meeste gevallen formatief gebruikt, waarbij het resultaat tot remediëring leidt als de doelen niet zijn behaald.

In het **voortgezet onderwijs** is eveneens weinig onderzoek gedaan op het terrein van toetsing (*n*=14). De beschrijvende analyses laten zien dat het merendeel betrekking heeft op summatieve toetsing (*n*=8). Daarnaast gaan bijna alle artikelen in op de fase afname en beoordeling (*n*=13). Uit klankbordgesprekken bleek tevens dat er weinig aandacht is voor de kwaliteit van toetsen binnen het voortgezet onderwijs. Wanneer er wel aandacht voor was, werd er in de praktijk weinig naar gehandeld. Docenten maken regelmatig zelf toetsen, maar overleggen hierover vrijwel nooit met collega's. In de klankbordgesprekken zocht men de verklaring hiervoor in de gebrekkige onderlinge verstandhouding wat betreft het onderwerp toetsen die docenten in het voortgezet onderwijs weerhoudt om met anderen te overleggen over hun toets of collega's feedback te geven.

Van het **beroepsgericht onderwijs** zijn meer artikelen gevonden (*n*=23). Uit gesprekken met de klankbordgroep bleek dat er aandacht voor de kwaliteit van toetsen is, maar ook dat er nog veel werkzaamheden te verrichten zijn op dit gebied. Het is opvallend dat er nauwelijks artikelen zijn die enkel over formatieve toetsing gaan (*n*=3), terwijl uit klankbordgesprekken bleek dat toetsing in het beroepsgericht onderwijs naast de (eind)examens veelal van formatieve aard is. Uit de analyses waren geen van de specifieke kwaliteitsaspecten kenmerkend voor het beroepsgericht onderwijs op de drie niveaus van hoofdcategorie (*p*=.401), subcategorie (*p*=.162) en onderdelen (*p*=.069). Baartman et al. (2007) vonden eveneens dat docenten uit het beroepsgericht onderwijs de tien door hen onderzochte klassieke en competentiegerichte kwaliteitsaspecten even belangrijk vonden. Uit de klankbordgesprekken bleek echter wel dat *authenticiteit* een grote rol speelt en dat men dit kenmerkend vindt voor het beroepsgericht onderwijs.

Het meest gevonden onderzoek heeft betrekking op het **hoger onderwijs** (*n*=102). De klankbordgesprekken bevestigden dit met het feit dat er veel aandacht is voor de kwaliteit van toetsen binnen deze onderwijssector. Uit de analyses bleek dat zowel de verdeling van kwaliteitsaspecten in de hoofdcategorieën (*p*=.457), in de subcategorieën (*p*=.627), als in de onderdelen (*p*=.543) niet verschilt tussen de artikelen over hoger onderwijs en het totaal aantal artikelen. Doordat er echter 102 van de 148 artikelen het hoger onderwijs betroffen, zouden de verschillen tussen deze groep en de gehele groep heel groot moeten zijn voordat van significantie sprake zou kunnen zijn. Dit bleek niet het geval, zodat ook voor deze sector geconstateerd kon worden dat er geen kenmerkende kwaliteitsaspecten gelden.

> *Er zijn geen verschillen gevonden in genoemde kwaliteitsaspecten tussen de onderwijssectoren. Mogelijk spelen in elke sector nagenoeg dezelfde aspecten een rol of kunnen verschillen niet worden gevonden door het kleine aantal artikelen in bepaalde sectoren. Het resultaat laat wel zien dat de aandacht voor toetskwaliteit in elke sector anders is: van weinig aandacht in het primair onderwijs tot veel aandacht in het hoger onderwijs.*

4.3.4 ROLLEN IN DE BEOORDELING

Van alle artikelen (*n*=242) gingen de meeste uit van een traditioneel model van beoordelen met de nadruk op de rol van de docent als beoordelaar van individueel werk van studenten. Er is weinig tot niets gevonden over de kwaliteit van de alternatieve beoordelingsmethodieken **peer-assessment**, (*n*=12), **self-assessment** (*n*=5) en **co-assessment** (*n*=0).

De enkele artikelen die wel ingingen op alternatieve beoordelingsmethodieken bleken vooral aanwezig binnen het hoger onderwijs. Bij verificatie van dit resultaat in de klankbordgroep werd dit beeld echter niet herkend: ook in het hoger onderwijs worden deze vormen nauwelijks daadwerkelijk gebruikt. Mogelijk wordt er in de onderzoekswereld met deze nieuwe vormen geëxperimenteerd, maar speelt het in de dagelijkse praktijk nog geen rol.

Door het kleine aantal artikelen was het niet zinvol om te analyseren of verschillende kwaliteitsaspecten een rol spelen bij de verschillende beoordelingsmethodieken. In diverse artikelen werd bovendien geconcludeerd dat de kwaliteitsaspecten binnen deze beoordelingsmethodiek afhangt van het doel waarmee wordt getoetst (Gielen, Dochy, Onghena, Struyven & Smeets, 2011; Ploegh et al., 2009).

Er is weinig geschreven over kwaliteitsaspecten van alternatieve beoordelingsmethodieken zoals peer-, self-, en co-assessment. Uit de klankbordgesprekken bleek tevens dat deze beoordelingsmethodieken nauwelijks in de onderwijssectoren worden ingezet.

4.3.5 TOETSCYCLUS

In dit onderzoek is onderscheid gemaakt tussen drie fasen van de toetscyclus: (1) de **constructiefase**, waarbij het doel wordt vastgesteld, de toets wordt ontworpen en de vragen worden geconstrueerd, (2) de **afname en beoordelingsfase**, waarbij de toets wordt afgenomen, wordt beoordeeld en het resultaat wordt teruggekoppeld en tot slot (3) de **evaluatiefase**, waarbij de toets wordt geëvalueerd op zijn kwaliteit en kwaliteitsaspecten worden gemeten.

Er waren 217 artikelen die de fase in de toetscyclus beschreven. Slechts 16 artikelen gingen in op de constructiefase van een toets. Daarentegen had ruim de helft van de artikelen betrekking op de afname en beoordelingsfase (n=116). De frequenties waarmee de kwaliteitsaspecten van de drie niveaus voorkwamen, weken bij de artikelen over de fase van afname en beoordeling significant af van de hele groep (p=.027; p<.000; p<.000). De hoofdcategorie *betrouwbaarheid* kwam frequent voor. Vooral de subcategorie *objectiviteit* en de onderdelen *beoordelingsvoorschrift* en *deskundigheid van de beoordelaar* bleken hierbinnen een hoge frequentie te hebben. De hoofdcategorie *generaliseerbaarheid* werd daarentegen minder frequent genoemd. Vooral de subcategorie *cognitieve activiteit* kwam weinig voor. Ook de subcategorie *begripsvaliditeit* werd in deze fase minder genoemd. Het is aannemelijk dat dit aspect niet van toepassing is tijdens de afname en beoordeling van een toets. Er wordt immers vooraf bekeken hoe de constructeur van een toets een bepaald kenmerk kan toetsen, en achteraf wordt getoetst of dit waargemaakt is (Tanilon, et al., 2009).

Er waren tot slot 85 artikelen die betrekking hadden op de evaluatiefase. Deze artikelen verschilden op hoofd- en subcategorieniveau niet van de hele groep artikelen (respectievelijk p=.100 en p=.066). Op het niveau van de onderdelen was wel een significant verschil (p<.001). De onderdelen die een hoge frequentie hadden in de afname- en beoordelingsfase, bleken in de evaluatiefase juist minder vaak genoemd te worden.

De meeste artikelen hadden betrekking op de afname en beoordelingsfase binnen de toetscyclus. In deze fase bleek de hoofdcategorie betrouwbaarheid, en de daarmee samenhangende kwaliteitsaspecten objectiviteit, deskundigheid van de beoordelaar en het gebruik van een beoordelingsschema vaak voor te komen. Een aantal van deze aspecten werd bij de evaluatiefase juist minder frequent genoemd. Bij de constructiefase werden geen verschillen gevonden.

4.3.6 IN DE PRAKTIJK

In deze reviewstudie werd tot slot de mate waarin een onderzoek in de praktijk werd uitgevoerd geanalyseerd. De onderliggende gedachte was dat de resultaten van een onderzoek de werkelijke schoolpraktijk meer benaderden als er onderzoek was gedaan in de schoolpraktijk zelf. Er is onderscheid gemaakt tussen **directe onderzoeken**, **indirecte onderzoeken** en **theoretische beschouwingen**. De bijbehorende hypothese luidde dat de kwaliteitsaspecten verschilden afhankelijk van de mate waarin de werkelijke schoolpraktijk werd onderzocht.

Ruim 60% van de gevonden artikelen bleken theoretische beschouwingen te zijn (*n*=146). Juist de praktijkgerichte artikelen waren voor het merendeel een theoretische beschouwing (73,2%). Dit is een opvallend resultaat, aangezien toetsing een bezigheid is die dagelijks plaatsvindt in de scholen. Toch blijkt uit Tabel 3 dat de mate waarin artikelen onderzoek hebben gedaan in de praktijk op hoofdcategorieniveau niet afwijkend was (*p*=.149, *p*=.418 en *p*=.813). Op het niveau van de subcategorieën verschilden alleen artikelen waarbij indirect onderzoek in de praktijk is gedaan (*p*=.009). De subcategorie *aanvaardbaarheid* werd hierbij vaker genoemd, terwijl er minder vaak over *validiteit* werd gesproken. Hoewel er tevens een significant verschil was op het niveau van onderdelen bij indirect onderzoek (*p*=.001), kunnen hier in verband met de kleine hoeveelheid data geen sterke uitspraken over gedaan worden.

Tabel 3. *Toetsingsgrootheden (X^2) van het totaal aantal artikelen vergeleken met de drie soorten artikelen.*

	Directe onderzoeken			**Indirecte onderzoeken**			**Theoretische beschouwingen**		
	Hoofd-categorie	Sub-categorie	Onderdeel	Hoofd-categorie	Sub-categorie	Onderdeel	Hoofd-categorie	Sub-categorie	Onderdeel
Alle artikelen [a]	6,76	22,70	33,24	3,92	33,71*	58,15*	1,58	8,55	20,73

Noot: [a] basisgroep; * $p \leq .01$

Vervolgens zijn de artikelen waarin direct en indirect onderzoek uitgevoerd werd samengenomen en vergeleken met de theoretische beschouwingen. Het is aannemelijk om te verwachten dat er verschil bestaat tussen (directe en indirecte) onderzoeken enerzijds en theoretische beschouwingen anderzijds wat betreft hun kwaliteitsaspecten. Tabel 4 laat zien dat in theoretische beschouwingen op hoofdcategorieniveau significant andere kwaliteitsaspecten worden genoemd vergeleken met wat wordt verwacht als de (directe en indirecte) onderzoeken als basisgroep werden genomen (X^2=11.69, $p \leq .05$). Zo werd de hoofdcategorie *generaliseer-baarheid* vaker genoemd in theoretische beschouwingen. Als echter de theoretische beschouwingen als basisgroep werden genomen, bleken de onderzoeken niet significant af te wijken (*p*=.084).

Op subcategorieniveau bleken beide groepen significant van elkaar te verschillen ($p \leq .01$). Er werd in de onderzoeken vaker gesproken over *aanvaardbaarheid*, terwijl in theoretische beschouwingen juist vaker werd gesproken over *validiteit*, *theoretische onderbouwing*, *generaliseerbaarheid* en *criteriumvaliditeit*.

Op het niveau van onderdelen verschilden de twee groepen significant van elkaar als de theoretische beschouwingen als basisgroep werd genomen ($p \leq .01$). Zo bleek het kwaliteitsaspect *meerdere beoordelaars* vaker in theoretische beschouwingen voor te komen dan verwacht. Uit de klankbordgesprekken bleek dat dit in de praktijk wel belangrijk werd gevonden, maar dat het vaak niet haalbaar was. Meerdere beoordelaars inzetten voor de scoring van toetsprestaties is niet efficiënt en vaak te duur. Hetzelfde geldt voor de inzet van *verschillende toetsvormen*.

Verder bleek in beide soorten artikelen veel aandacht te zijn voor *betrouwbaarheid* en aspecten die daarmee samenhangen. Het kwaliteitsaspect *validiteit* werd daarentegen vooral in theoretische beschouwingen vaak genoemd (o.a. Birenbaum, 2007; Newton, 2012; Shephard, 2009) en kwam minder vaak als subcategorie in de directe en indirecte onderzoeken voor. Mogelijk wordt dit aspect door onderzoekers minder relevant gevonden of is dit aspect moeilijker meetbaar.

Tabel 4. *Toetsingsgrootheden (X^2) van direct/indirect onderzoek vergeleken met theoretische beschouwingen.*

	Onderzoek (direct/indirect)			Theoretische beschouwing		
	Hoofd-categorie	Sub-categorie	Onderdeel	Hoofd-categorie	Sub-categorie	Onderdeel
Onderzoek (direct/indirect) [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; * $p \leq .05$; ** $p \leq .01$.

In Tabel 5 is zichtbaar hoe de verdeling van de wetenschappelijke en praktijkgerichte artikelen was ten opzichte van de drie soorten artikelen. Hieruit blijkt opnieuw dat zowel wetenschappelijke als praktijkgerichte artikelen merendeels niet in de praktijk zijn uitgevoerd (respectievelijk 53,7% en 73%). Daarnaast laat de tabel zien dat de meeste artikelen waarin direct of indirect onderzoek in de praktijk is gedaan, wetenschappelijke artikelen zijn (23,8% en 22,5%). De praktijkgerichte artikelen, zoals artikelen uit vaktijdschriften, staan weliswaar dichter bij de praktijk, maar gaven relatief toch vaker een theoretische beschouwing.

De significante verschillen tussen de artikelen waarin direct, indirect of niet in een praktijksituatie onderzoek is uitgevoerd, worden door de wetenschappelijke artikelen veroorzaakt. Hier is hetzelfde patroon van significante effecten zichtbaar als van de gehele groep artikelen. Binnen de groep praktijkgerichte artikelen was daarentegen geen significant verschil aanwezig.

Tabel 5. *Frequenties van het soort onderzoek in de wetenschappelijke en praktijkgerichte artikelen.*

	Direct onderzoek		Indirect onderzoek		Theoretische beschouwing		Totaal	
	N	%	N	%	N	%	N	%
Wetenschappelijk artikel	38	23,8%	36	22,5%	86	53,7%	160	100%
Praktijkgericht artikel	13	16,0%	9	11,0%	60	73,0%	82	100%

Artikelen waarin (in)direct in de praktijk onderzoek is uitgevoerd, noemden andere kwaliteitsaspecten dan theoretische beschouwingen. Dit verschil werd veroorzaakt door de wetenschappelijke artikelen. Betrouwbaarheid is in beide categorieën veel genoemd, maar bleek binnen de directe en indirecte onderzoeken het meest onderzocht. Het kwaliteitsaspect validiteit werd vooral in theoretische beschouwingen vaak genoemd.

5. CONCLUSIE EN DISCUSSIE

Het doel van deze reviewstudie was om meer inzicht te krijgen in de huidige kennis over kwaliteit van toetsing. De volgende vraag stond centraal: *wat beschouwt men op dit moment als kwaliteit van toetsen in het onderwijs?* In internationale databases en Nederlandse literatuur is op systematische wijze gezocht naar bruikbare literatuur om deze vraag te beantwoorden. Met behulp van experts is vervolgens een begrippenkader opgesteld dat de basis vormde voor de codering van de artikelen. De onderzoeksresultaten zijn tot slot geanalyseerd en ter validering voorgelegd aan verschillende klankbordgroepen.

Het begrippenkader heeft in kaart gebracht welke aspecten met betrekking tot toetskwaliteit in de literatuur worden genoemd. Hierbij is onderscheid gemaakt in hoofdcategorieën, subcategorieën en onderdelen. Uit de analyses is gebleken dat de hoofdcategorie *betrouwbaarheid* het meest frequent voorkomt in de artikelen. Vooral de subcategorie *objectiviteit* en de onderdelen *beoordelingsvoorschrift* en *deskundigheid van de beoordelaar* worden hierbij frequent genoemd. Ook bleek dat de frequentie van de kwaliteitsaspecten verschilt afhankelijk van twee factoren: het toetsdoel en de fase in de toetscyclus. Bij een summatief toetsdoel kwamen de kwaliteitsaspecten *validiteit* en *normering en cesuur* meer voor, terwijl bij formatieve toetsen vooral *consequentiële validiteit* werd genoemd. In de fase afname en beoordeling bleek *betrouwbaarheid* meer frequent genoemd te zijn dan in de andere fasen van de toetscyclus. Tussen de onderwijssectoren werd weliswaar geen verschil gevonden in kwaliteitsaspecten, maar wel in de mate van aandacht voor toetskwaliteit. Ook bleken in alle onderwijssectoren weinig alternatieve beoordelingsmethodieken te worden ingezet. Tot slot kan geconcludeerd worden dat deze resultaten verschillen afhankelijk van de mate waarin een onderzoek in de praktijk is uitgevoerd.

Zoals bij ieder onderzoek zijn er ook bij deze reviewstudie kanttekeningen te plaatsen. Allereerst is het bijna onwaarschijnlijk dat werkelijk alle artikelen die betrekking hebben op toetskwaliteit zijn gevonden. Daarnaast zijn er relatief veel artikelen gevonden op toetsniveau en weinig op het niveau van toetsprogramma of van opgaven binnen een toets, hoewel er op verschillende manieren is gezocht naar artikelen in zowel internationale databases als in de Nederlandse toetspraktijk. Mogelijk is de focus van het onderzoek en daarmee de keuze van de zoektermen bepalend geweest voor dit resultaat. Tevens kan het gemaakte onderscheid binnen toetsdoel, toetscyclus en soort beoordelaar deels bepalend zijn geweest voor de gevonden resultaten. In dit onderzoek is gekozen voor de meest gangbare indelingen in de geselecteerde literatuur, bijvoorbeeld het onderscheid tussen een formatief en summatief toetsdoel. Andere indelingen zouden mogelijk tot andere resultaten kunnen leiden. Tot slot laat deze reviewstudie zien naar welke kwaliteits-aspecten veel 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 van dit onderzoek geven aanleiding voor vervolgonderzoek. Allereerst kwam naar voren dat de fase in de toetscyclus bepalend is voor welke kwaliteitsaspecten van belang zijn. Een aanbeveling op grond van dit resultaat is 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 toetsproces beter wordt ondersteund met als uiteindelijk resultaat dat de toetskwaliteit wordt verbeterd?

Daarnaast bleek dat het toetsdoel een beïnvloedende factor is wat betreft de kwaliteitsaspecten. Het is voor docenten van belang dat zij eerst het toetsdoel vaststellen en aan de hand daarvan nagaan met welke kwaliteitsaspecten zij vooral rekening moeten houden.

Hoewel er geen opvallende verschillen zijn gevonden tussen de onderwijssectoren, bleek er wel verschil in de mate van aandacht voor toetskwaliteit. Dit suggereert dat er sprake is van verschillende behoeften aan informatie over toetskwaliteit. In toekomstig onderzoek en praktijkgerichte interventies is het van belang om hiermee rekening te houden. Zo zou het binnen 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.

Tevens bleek dat in empirisch onderzoek en theoretische beschouwingen veel aandacht is voor betrouwbaarheid en aspecten die daarmee samenhangen, zoals objectiviteit en deskundigheid van de beoordelaar. Het kwaliteitsaspect validiteit wordt echter vrijwel alleen in theoretische beschouwingen genoemd. Een mogelijk onderzoeksthema is daarom validiteitsaspecten in de praktijk. Dit zou zich kunnen richten op het identificeren van bedreigingen voor de validiteit van toetsscores bij het meten van specifieke vaardigheden. Het zou zich ook kunnen richten op het adresseren van specifieke vragen uit de praktijk ten aanzien van de validiteit van toetsen, zoals het stellen van eisen ten aanzien van de inhoud van een toets of 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.

Tot slot is er vooral onderzoek gevonden op het niveau van de toets. Er kan echter ook meer gedetailleerd worden ingezoomd op de opgaven (itemniveau) of juist worden uitgezoomd naar het toetsprogrammaniveau. Het komt namelijk in alle sectoren van het onderwijs voor dat toetsresultaten of metingen van een vaardigheid van studenten met elkaar worden gecombineerd om te komen tot een beoordeling. De combinatie van metingen moet zo efficiënt mogelijk gebeuren, bijvoorbeeld door het inrichten van een kwalitatief goed toetsprogramma. Daarnaast geldt dat toetsen verweven kunnen worden in het leerproces door ze op een meer formatieve manier in te zetten. 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.

Concluderend geeft dit onderzoek inzicht in de aspecten die een rol spelen bij het bepalen van de kwaliteit van een toets en laat het zien waar hiaten in onderzoek binnen dit thema liggen. Op basis van deze kennis en inzichten kan vervolgonderzoek worden opgezet en kan de werkwijze in de praktijk worden verbeterd, zodat de toetsen op grond waarvan belangrijke beslissingen over studenten tot stand komen van goede kwaliteit zijn.

6. REFERENTIES

Referenties met een asterisk (*) zijn referenties waar in deze rapportage naar verwezen wordt.

Adema, J. J. (2010). Voorwaardelijk toetsen in het mbo: Borging van het niveau en besparing op de kosten. *Examens, 2*, 21-24.

Alem, J., & Boudreau-Larivière, C. (2012). Evaluation of an internship assessment grid for francophone physical and health education student interns. *The Canadian Journal for the Scholarship of Teaching and Learning, 3*(1), article 5. doi:10.5206/cjsotl-rcacea/vol3/issl/5

* 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.

Andrade, J., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education, 32*(2), 159-181. doi:10.1080/02602930600801928

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

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*(1), 125-141. doi:10.1177/0265532212452396

Baartman, L. K. J., Bastiaens, T. J., & Kirschner, P. A. (2005). Examens als graadmeter van het onderwijspeil: Kwaliteitscriteria voor Competentie Assessment Programma's. *Examens, 2*, 13-22.

* Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Vleuten, C. P. M. van der (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., Bastiaens, T. J., Kirschner, P. A., & Vleuten, C. P. M. van der (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review, 2*, 114-129. doi:10.1016/j.edurev.2007.06.001

* Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Vleuten, C. P. M. van der (2007). Teachers' opinions on quality criteria for Competency Assessment Programs. *Teaching and Teacher Education, 23*(6), 857-867. doi:10.1016/j.tate.2006.04.043

Baartman, L. K. J., & Gulikers, J. T. M. (2014). Beoordelen als fundament van goed opleiden in het beroepsonderwijs: een analyse van toetsprogramma's in het mbo en hbo. *Pedagogische Studiën, 91*(1), 54-68. Verkregen van: http://www.vorsite.nl/content/bestanden/ped_studie_1_samenvatting-kopie-4.pdf

Baartman, L. K.J., Gulikers, J. T. M., & Dijkstra, A. (2013). Factors influencing assessment quality in higher vocational education. *Assessment & Evaluation in Higher Education, 38*(8), 978-997. doi:10.1080/02602938.2013.771133

Baartman, L. K. J., & Kloppenburg, R. (2013). *KIT: KwaliteitsInstrument Toetsprogramma's in beroepsgericht onderwijs: Zelfevaluatie-instrument voor docenten*. Verkregen van: www.kwaliteit-toetsprogramma.nl

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.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Vleuten, C. P. M. van der (2007). Determining the quality of Competence Assessment Programs: A self-evaluation procedure. *Studies in Educational Evaluation, 33*, 258-281. doi:10.1016/j.stueduc.2007.07.004

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Vleuten, C. P. M. van der (2007). Kwaliteitsmeting van Competentie Assessment Programma's via zelfevaluatie. *OnderwijsInnovatie, 1*, 17-26.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Vleuten, C. P. M. van der (2011). *Self-evaluation of assessment programs: A cross-case analysis. Evaluation and Program Planning, 34*, 206-216. doi:10.1016/j.evalprogplan.2011.03.001

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice, 21*(3), 5-18. doi:10.1111/j.1745-3992.2002.tb00095.x

Bakker, M. E. J., Sanders, P., Beijaard, D., Roelofs, E., Tigelaar, D., & Verloop, N. (2008). De betrouwbaarheid en generaliseerbaarheid van competentiebeoordelingen op basis van een videodossier. *Pedagogische Studiën, 85*, 240-260.

Barneveld, S. (2008). Mening leraar is meer waard dan Cito-toets. *Didactief, 1-2*, 22-23.

Baume, D., Yorke, M., & Coffey, M. (2004). What is happening when we assess, and how can we use our understanding of this to improve assessment? *Assessment & Evaluation in Higher Education, 29*(4), 451-477. doi:10.1080/02602930310001689037

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

Beek, A. ter, & Rijken, R. (1998). *Meetinstrument Kwaliteit Schooltoetsen*. 's-Hertogenbosch: KPC Onderwijs Innovatie Centrum.

Berkel, A. van (2012). Kritische reflectie op competentietoetsen in het hbo. *OnderwijsInnovatie, 2*, 17-26.

Berkel, A. van, Dinther, M., Oudkerk Pool, I., & Speetjes, J. (2008). Certificeren van assessoren in het hbo. Waarom en hoe? *OnderwijsInnovatie, 2*, 17-25.

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

Berkel, H. J. M. van (2007). Terecht of niet? *Examens, 1*, 20-22.

Berkel, H. J. M. van (2011). Meten is weten; vergeet het maar! Over het zoeken naar ware score. *Examens, 3*, 10-14.

Berkel, H. van, & Bax, A. (2006). Toetsen: toetssteen of dobbelsteen. In: Van Berkel, H., & Bax, A. (red.), *Toetsen in het hoger onderwijs* (pp. 41-56). Houten: Bohn Stafleu van Loghum.

Berkel, H. J. M. van, & Bax, A. E. (2008). Dat heb ik niet gezegd: Over het afleggen van mondelinge examens. *Examens, 3*, 5-8.

Berkel, H. J. M. van, & Draaijer, S. (2011). Gids voor toetsontwikkeling. *Examens, 1*, 1-8.

Beukers, G., Bokhoven, M. van, Gerritsen, K., Hees, C. van, Wildering, B., & Woord, R. ter (2013). *Toetsbeleidskader Saxion: Kaders en richtlijnen voor het toetsbeleid en toetsplan*. Deventer/Apeldoorn/ Enschede: Saxion.

Binsbergen, M., Verstappen, J., & Van Selm, M. (2014). *Toetsing in het hoger onderwijs, percepties van studenten over toetsing in het hoger onderwijs.* Culemborg: Bureau ICE. Utrecht: Onderzoeksbureau LSVb.

* Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation, 33*(1), 29-49. doi:10.1016/j.stueduc.2007.01.004

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., Wiesemes (Ed), R., et al. (2006). A learning integrated assessment system. *Educational Research Review, 1*, 61-67. doi:10.1016/j.edurev. 2006.01.001

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

Black, P., & Wiliam, D. (2001). Inside the black box: Raising standards through classroom assessment. *PhiDelta Kappan, 80*(2), 139-148. Verkregen van: http://weaeducation.typepad.co.uk/files/blackbox-1.pdf

Black, P., & Wiliam, D. (2007). Large-scale assessment systems: Design principles drawn from international comparisons. *Measurement: Interdisciplinary Research and Perspectives, 5*(1), 1-53. doi:10.1080/ 15366360701293386

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: the role of assessment criteria in UK higher education grading practices. *Studies in Higher Education, 36*(6), 655-670. doi:10.1080/03075071003777716

Bloxham, S., & West, A. (2004). Understanding the rules of the game: Marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment & Evaluation in Higher Education, 29*(6), 721-733. doi:10.1080/0260293042000227254

Boers-Visker, E. M. (2007). *Onderzoek naar de validiteit en betrouwbaarheid van een toets taalbeheersing Nederlandse Gebarentaal.* Unpublished master's thesis, Hogeschool Utrecht, Utrecht, The Netherlands.

Bonesrønning, H. (2004). Do the teachers' grading practices affect student achievement? *Education Economics, 12*(2), 151-167. doi:10.1080/0964529042000239168

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151-167. doi:10.1080/713695728

Boud, D. and Associates (2010). *Assessment 2020: Seven propositions for assessment reform in higher education*. Sydney: Australian Learning and Teaching Council.

Boxel, P. van, Reumer, C., Os, W. van, & Boter, J. (2008). De inzet van online peer assessment als formatief en summatief beoordelingsinstrument. *Tijdschrift voor Hoger Onderwijs, 26*(4), 229-246.

Bramley, T., & Gill, T. (2010). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education, 25*(3), 293-317. doi:10.1080/02671522.2010.498147

Brink, M. van der, & Heuves, T. (2005). De kwaliteitscommissie aan het werk: Hoe goed is een goede proeve van bekwaamheid? *Examens, 2*, 22-24.

Bronkhorst, L. H., Baartman, L. K. J., & Stokking, K. M. (2012). The explication of quality standards in self-evaluation. *Assessment in Education: Principles, Policy & Practice, 19*(3), 357-378. doi:10.1080/0969594X.2011.570731

Brookhart, S. M. (2005).The quality of local district assessment used in Nebraska's school-based teacher-led assessment and reporting system (STARS). *Educational Measurement: Issues and Practice, 24*(2), 14-21. doi:10.1111/j.1745-3992.2005.00007.x

Brookhart, S. M. (2013). The use of teacher judgment for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice, 20*(1), 69-90. doi:10.1080/0969594X.2012.703170

Brown, G. (2001). *Assessment: A guide for lecturers*. York: Learning and Teaching Support Network (LTSN) Generic Centre.

Brown, S. (2004). Assessment for learning. *Learning and Teaching in Higher Education, 1,* 81-89. Verkregen van: http://www2.glos.ac.uk/offload/tli/lets/lathe/issue1/articles/brown.pdf

* Calhoon, M. B., Greenberg, D., & Hunter, C. V. (2010). A comparison of standardized spelling assessments: Do they measure similar orthographic qualities? *Learning Disability Quarterly, 33*, 159-170.

Carjuzaa, J., & Ruff, W. G. (2010). When western epistemology and an indigenous worldview meet: Culturally responsive assessment in practice. *Journal of the Scholarship of Teaching and Learning, 10*(1), 68-79. Verkregen van: www.iupui.edu/~josotl

Chappuis, S., Chappuis, J., & Stiggins, R. (2009). The quest for quality. *Multiple Measures, 67*(3), 14-19. Verkregen van: http://www.ascd.org/publications/educational-leadership/nov09/vol67/num03/The-Quest-for-Quality.aspx

* 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

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing form instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891-901. doi:10.1037/0022-0663.98.4.891

Cizek, G. J. (2009). Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts. *Theory Into Practice, 48*, 63-71. doi:10.1080/00405840802577627

Clarkeburn, H., & Kettula, K. (2012). Fairness and using reflective journals in assessment. *Teaching in Higher Education, 17*(4), 439-452. doi:10.1080/13562517.2011.641000

Cluitmans, J., & Klarus, R. (2005). Competentiebeoordeling: Een pleidooi voor congruentie. *Tijdschrift voor Hoger Onderwijs 23*(4), 221-238.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice, 3*(3), 265-286. doi:10.1080/0969594960030302

* 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
Davies, A., & LeMahieu, P. (2003). Assessment for learning: Reconsidering portfolios and research evidence. In: *Optimistic new modes of assessment: In search of qualities and standards* (pp. 141-169). Dordrecht/Boston/ London: Kluwer Academic Publishers.
Davis, N. T., Kumtepe, E. G., & Aydeniz, M. (2007). Fostering continuous improvement and learning through peer assessment: Part of an integral model of assessment. *Educational Assessment, 12*(2), 113-135. doi:10.1080/10627190701232720
* 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
Dierick, S., & Dochy, F. (2001). New lines in edumetrics: New forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27*(4), 307-329. doi:10.1016/S0191-491X(01)00032-3
Dierick, S., Dochy, F., & Watering, G. van de (2001). Assessment in het hoger onderwijs: Over de implicaties van nieuwe toetsvormen voor de edumetrie. *Tijdschrift voor Hoger Onderwijs, 19*(1), 2-17.
Dierick, S., Watering, G. van de, & Muijtjens, A. (2002). De actuele kwaliteit van assessment: Ontwikkelingen in de edumetrie. In: F. Dochy, L. Heylen, & H. van de Mosselaer, *Assessment in onderwijs. Nieuwe toetsvormen en examinering in studentgericht onderwijs en competentiegericht onderwijs* (pp. 91-122). Utrecht: Uitgeverij Lemma BV.
Dijkstra, A., & Baartman, L. (2011). Zelfevaluatie van de kwaliteit van assessment. *OnderwijsInnovatie, 1*, 17-26.
Dijkstra, I. (2007). Kleutertoets deugt niet. *Didactief, 37*(6), 21.
Dijkstra, J., Galbraith, R., Hodges, B. D., McAvoy, P. A., McCrorie, P., Southgate, L. J., Vleuten, C. P. M. van der, et al. (2012). Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Medical Education, 12*(1), 20-37. doi:10.1186/1472-6920-12-20
Dijkstra, J., Vleuten, C. P. M. van der, & Schuwirth, L. W. T. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education, 15*, 379-393. doi:10.1007/s10459-009-9205-z
Dijkstra, M. (2010). *Toetssteen der wijzen: Op weg naar betere toetsen in het schoolexamen Nederlands*. Unpublished master's thesis, Universiteit Utrecht, Utrecht, The Netherlands.
Dochy, F. (2009). The edumetric quality of new modes of assessment: Some issues and prospects. In: G. Joughin (ed.), *Assessment, learning and judgement in higher education* (pp. 85-114). Springer Science + Business Media B.V.
Dochy, F., Admiraal, W., & Pilot, A. (2003). Peer- en co-assessment als instrument voor diepgaand leren: bevindingen en richtlijnen. *Tijdschrift voor Hoger Onderwijs, 21*(4), 220-229.
* 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
Dolkar, D. (2009). *Studying school-based summative assessments in high-stakes examinations in Bhutan: A question of trust?* Unpublished master's thesis, Universiteit Twente, Enschede, The Netherlands.
Driessen, E. W., Tartwijk, J. van, Govaerts, M., Teunissen, P., & Vleuten, C. P. M. van der (2012). The use of programmatic assessment in the clinical workplace: A Maastricht case report. *Programmatic Assessment in the Clinical Workplace, 34*, 226-231. doi:10.3109/0142159X.2012.652242
Driessen, E., Vleuten, C. P. M. van der, Schuwirth, L., Tartwijk, J. van, & Vermunt, J. (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: A case study. *Medical Education, 39*, 214-220. doi:10.1111/j.1365-2929.2004.02059.x
Dysthe, O., Engelsen, K. S., & Lima, I. (2007). Variations in portfolio assessment in higher education: Discussion of quality issues based on a Norwegian survey across institutions and disciplines. *Assessing Writing, 12*(2), 129-148. doi:10.1016/j.asw.2007.10.002
* 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.

Elen, J., Sasanguie, D., Coens, J., Clarebout, G., Noortgate, W. van den, Vandenabeele, J., & Fraine, B. de (2009). Ontkoppelen van begeleiden en summatief beoordelen in het hoger onderwijs: een aanzet tot discussie. *Tijdschrift voor Hoger Onderwijs 27*(3), 157-170.

Elsen, M., Wolters, L., Pilot, A., & Nedermeijer J. (2001). Geschiktheid van toetsvormen voor het meten van academische competenties. *Tijdschrift voor Hoger Onderwijs, 19*(1), 33-43.

Erkens, T., Kuhlemeier, H., & Weerden, J. van (2007). Op weg naar betere schoolexamens: de kwaliteitsmonitor schoolexamens van Cito. *Examens, 3*, 5-9.

Evers, A. (2001). The revised Dutch rating system for test quality. *International Journal of Testing, 1*(2), 155-182. doi:10.1207/S15327574IJT0102_4

Evers, A. (2004). Hoe verantwoord meten we? Het COTAN-systeem voor de beoordeling van de kwaliteit en toepasbaarheid van tests. *Examens, 1*(1), 31-34.

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

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing, 10*, 295-317. doi:10.1080/15305058.2010.518325

Fallows, S., & Chandramohan, B. (2001). Multiple approaches to assessment: Reflections on use of tutor, peer and self-assessment. *Teaching in Higher Education, 6*(2), 229-246. doi:10.1080/13562510120045212

Freriksen, J. R., & Collins, A. (1990). A systems approach to educational testing. *Educational Researcher*, *18*(9), 27-32. Verkregen van: http://www.jstor.org/stable/1176716

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39-53. doi:10.1111/j.1745-3992.2009.00154.x

Fuchs, L. S., & Fuchs, D. (2001). Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research & Practice, 16*(3), 174-181. doi:10.1111/0938-8982.00018

Geelen, A., & Heij, K. (2014). Het belang van een goed toetsmodel: Waarom heb ik een onvoldoende en hij niet? *Toets!, 2*, 32-35.

Gerritsen-van Leeuwenkamp, K. J., & Joosten-ten Brinke, D. (2013). De perceptie van hbo-studenten op de kwaliteit van toetsing: het belang van vijftig kwaliteitskenmerken van toetsing. *Examens, 1*, 11-16.

* Gielen, S., Dochy, F., Onghena, P., Struyven, K., & Smeets, S. (2011). Goals of peer assessment and their associated quality concepts. *Studies in Higher Education, 36*(6), 719-735. doi:10.1080/03075071003759037

* 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

Glerum, J. (2010). *Kwaliteit van assessment in het competentiegericht middelbaar beroepsonderwijs: Een case studie*. Unpublished master's thesis, Open Universiteit, Heerlen, The Netherlands.

Goedendorp, I. (2008). Vaststellen van toetsen nader bekeken. *Examens, 2*, 18-19.

Goldin, I. M., & Ashley, K. D. (2012). Eliciting formative assessment in peer review. *Journal of Writing Research, 4*(2), 203-237. Verkregen van: http://www.jowr.org/Ccount/click.php?id=56

Grainger, P., Purnell, K., & Zipf, R. (2008). Judging quality through substantive conversations between markers. *Assessment & Evaluation in Higher Education, 33*(2), 133-142. doi:10.1080/02602930601125681

Gulikers, J. T. M. (2007) Authentiek beoordelen in een curriculum. *OnderwijsInnovatie, 2*, 11-14.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*(3), 67-86. doi:10.1007/BF02504676

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2006). Authentieke toetsing: de beroepspraktijk in het vizier. *OnderwijsInnovatie, 2*, 17-24.

Gulikers, J., Biemans, H., & Mulder, M. (2009). Developer, teacher, student and employer evaluations of competence based assessment quality. *Studies in Educational Evaluation, 35*, 110-119. doi:10.1016/ j.stueduc.2009.05.002

Guskey, T. R. (2011). Stability and change in high school grades. *NASSP Bulletin, 95*(2), 85-98. doi:10.1177/0192636511409924

Hamers, P. (2007). Onbedoeld oneerlijk toetsen. *Didactief, 37*(7), 20-21.

Harlen, W. (2005). Teachers' summative practices and assessment for learning - tensions and synergies. *The Curriculum Journal, 16*(2), 207-223. doi:10.1080/09585170500136093

Harlen, W. (2007). Criteria for evaluating systems for student assessment. *Studies in Educational Evaluation, 33*, 15-28. doi:10.1016/j.stueduc.2007.01.003

HBO-raad, vereniging van hogescholen (2012). *Vreemde ogen dwingen: Eindrapport Commissie externe validering examenkwaliteit hoger beroepsonderwijs.* Den Haag.

Hees, C. K. van (2013). Beoordelen van de kwaliteit van toetsing: Uitvoering van interne kwaliteitsmeting bij Saxion. *Examens, 3,* 11-16.

Helderman, A., & Wind, T. (2006). KCE-audit: deugdelijk op basis van zelfevaluatie van de instellingen. *Examens, 2*, 10-13.

Hendriks, P., & Schoonman, W. (red.) (2006). *Handboek assessment deel 1, gedragsproeven: Ontwikkeling, implementatie en evaluatie*. Assen: Van Gorcum.

Hensley, L. G., Smith, S. L., & Thompson, R. W. (2003). Assessing competencies of counselors-in-training: Complexities in evaluating personal and professional development. *Counselor Education & Supervision, 42*, 219-230. doi:10.1002/j.1556-6978.2003.tb01813.x

Herman, J. L., & Baker, E. L. (2005). Making benchmark testing work. *Assessment to Promote Learning, 63*(3), 48-54. Verkregen van: http://www.ascd.org/ASCD/pdf/journals/ed_lead/el200511_herman.pdf

Higgins, R., Hartley, P., & Skelton, A. (2002). The conscientious consumer: Reconsidering the role of assessment feedback in student learning. *Studies in Higher Education, 27*(1), 53-64. doi:10.1080/03075070120099368

Hyvärinen, M., Tanskanen, P., Katajavuori, N., & Isotalus, P. (2012). Evaluating the use of criteria for assessing profession-specific communication skills in pharmacy. *Studies in Higher Education, 37*(3), 291-308. doi:10.1080/03075079.2010.510183

* 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

Jacobs, J. W. J., & Knecht-van Eekelen, A., de (2007). Persoonlijke vakbekwaamheidscertificaten uitgereikt: een extra borg op kwaliteit in het mbo. *Examens, 4*, 15-16.

Jamieson, J., Chapelle, C. A., & Preiss, S. (2004). Putting principles into practice. *ReCALL, 16*, 396-415. doi:10.1017/S0958344004001028

Johnson, R. L., Fisher, S. Willeke, M. J., & McDaniel, F. (2003). Portfolio assessment in a collaborative program evaluation: The reliability and validity of a family literacy portfolio. *Evaluation and Program Planning, 26*, 367-377. doi:10.1016/S0149-7189(03)00053-3

Johnston, B. (2004). Summative assessment on portfolios: an examination of different approaches to agreement over outcomes. *Studies in Higher Education, 29*(3), 395-412. doi:10.1080/03075070410001682646

Johnston, P., & Costello, P. (2005). Theory and research into practice: Principles for literacy assessment. *Reading Research Quarterly, 40*(2), 256-267. doi:10.1598/RRQ.40.2.6

Jonsson, A. (2010). The use of transparency in the 'Interactive examination' for student teachers. *Assessment in Education: Principles, Policy & Practice, 17*(2), 183-197. doi:10.1080/09695941003694441

Jonsson, A., Baartman, L. K. J., & Lennung S. A. (2009). Estimating the quality of performance assessments: The case of an 'interactive examination' for teacher competencies. *Learning Environments Research, 12*(3), 225-241. doi:10.1007/s10984-009-9061-z

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144. doi:10.1016/j.edurev.2007.05.002

Joosten-ten Brinke, D. (2011). *Eigentijds toetsen en beoordelen*. Lectorale Rede. Tilburg: Fontys Lerarenopleiding Tilburg.

* 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

Joosten-ten Brinke, D., Sluijsmans, D. M. A., & Jochems, W. M. G. (2010). Assessors' approaches to portfolio assessment in assessment of prior learning procedures. *Assessment & Evaluation in Higher Education, 35(1)*, 55-70. doi:10.1080/02602930802563086

Jung, L. A., & Guskey, T. R. (2007). Standards-based grading and reporting: A model for special education. *Teaching Exceptional Children, 40*(2), 48-53.

Kame'enui, E. J., Fuchs, L., Francis, D. J., Good III, R., O'Connor, R. E., Simmons, D. C., Tindal, G. T., et al. (2006). The adequacy of tools for assessing reading competence: A framework and review. *Educational Researcher, 35*(4), 3-11. doi:10.3102/0013189X035004003

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*(3), 135-170. doi:10.1207/s15366359mea0203_1

Kappe, F. R. (2008). Hoe betrouwbaar is peer-assessment? *Tijdschrift voor Hoger Onderwijs, 26*(2), 93-102.

Kloppenburg, R. T. H. M. (2011). *Bekwaam beoordeeld: Inhoud, functie en kwaliteit van competentiegerichte assessments in social work opleidingen*. Unpublished doctoral dissertation, Universiteit Utrecht, Utrecht, the Netherlands.

Knevel, R. (2013). Taxonomieën zijn hot … en handig. *Toets!, 1*, 1-10.

Kok, B., & Bax, A. (2007). IBC als ontwerpmethode: Kwaliteitsimpuls voor praktijkbeoordelingen. *Examens, 2*, 20-23.

Laarhoven, F. van (2008). *Taaltoetsen testen: De kwaliteit van toetsen binnen een methode voor moderne vreemde talen*. Unpublished master's thesis, Universiteit Utrecht, Utrecht, The Netherlands.

Lans, W. & Verkroost, M. (2004). Bottom-up: teachers' contribution in developing uniform criteria to assess design products. *European Journal of Engineering Education, 29*(2), 275-282. doi:10.1080/0304379032000157222

Leigh, I. W., Bebeau, M. J., Nelson, P. D., Rubin, N. J., Smith, I. L., Lichtenberg, J. W., Portnoy, S., et al. (2007). Competency assessment models. *Professional Psychology: Research and Practice, 38*(5), 463-473. doi:10.1037/0735-7028.38.5.463

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practices, 26*(2), 3-16. doi:10.1111/j.1745-3992.2007.00090.x

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance-based assessment: expectations and validation criteria. *Educational Researcher, 20*(8), 15-21. doi:10.3102/0013189X020008015

Lizzio, A. & Wilson, K. (2008). Feedback on assessment: Students' perceptions of quality and effectiveness. *Assessment & Evaluation in Higher Education, 33*(3), 263-275. doi:10.1080/02602930701292548

Loopik, D. van (2012). *Criteria voor het valide examineren van mondelinge examens: Een delphi studie*. Unpublished master's thesis, Universiteit Utrecht, Utrecht, The Netherlands.

Lucas, L. (2012). Write more, grade less: Five practices for effectively grading writing. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 85*(4), 136-140. doi:10.1080/00098655.2012.659772

Maclellan, E. (2004). How convincing is alternative assessment for use in higher education? *Assessment and Evaluation in Higher Education, 29*(3), 311-321. doi:10.1080/0260293042000188267

* 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

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education, 37*(8), 919-931. doi:10.1080/02602938.2011.586991

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75-100. doi:10.1177/0265532208097337

McMillan, J. H., & Lawson, S. R. (2001). Secondary science teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*, 20-32.

Meeus, W., Petegem, P. van, & Engels, N. (2009). Validity and reliability of portfolio assessment in pre-service teacher education. *Assessment & Evaluation in Higher Education, 34*(4), 401-413. doi:10.1080/02602930802062659

Miller, D. M., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24*, 367-378. doi:10.1177/01466210022031813

Morley, L., Leonard, D., & David, M. (2002). Variations in vivas: Quality and equality in British PhD assessments. *Studies in Higher Education, 27*(3), 263-273. doi:10.1080/03075070220000653

Moss, C. M. (2013). Research on classroom summative assessment. In: J. H. McMillan (Ed.) (2013). *SAGE Handbook of Research on Classroom Assessment* (pp. 235-255). United States of America: SAGE Publications, Inc.

National Joint Committee on Learning Disabilities (2010). Comprehensive assessment and evaluation of students with learning disabilities. *Learning Disability Quarterly, 34*(1), 3-16. doi:10.1177/073194871103400101

Newell, J. A., Dahm, K. D., & Newell, H. L. (2002). Rubric development and inter-rater reliability issues in assessing learning outcomes. *Chemical Engineering Education*, *36*(3), 212-215. Verkregen van: http://www.engr.uky.edu/~aseeched/papers/2002/0323.pdf

* 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.

Nitko, A. J., & Brookhart, S. M. (2007). Reliability of Assessment Results. In: Nitko, A. J., & Brookhart, S. M., *Educational Assessment of Students* (pp. 66-83). New Jersey: Pearson Education, Inc.

Nitko, A. J., & Brookhart, S. M. (2007). Validity of Assessment Results. In: Nitko, A. J., & Brookhart, S. M., *Educational Assessment of Students* (pp. 37-65). New Jersey: Pearson Education, Inc.

O'Donovan, B., Price, M. & Rust, C. (2001). The student experience of criterion-referenced assessment (through the introduction of a common criteria assessment grid). *Innovations in Education and Teaching International, 38*(1), 74-85. doi:10.1080/14703290130000287 3

O'Donovan, B., Price, M., & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education, 9*(3), 325-335. doi:10.1080/1356251042000216642

Omelicheva, M. Y. (2005). Self and peer evaluation in undergraduate education: Structuring conditions that maximize its promises and minimize the perils. *Journal of Political Science Education, 1*(2), 191-205. doi:10.1080/15512160590961784

* Onderwijsraad (2006). *Advies Examinering: draagvlaken 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.

Orsmond, P., Merry, S., & Callaghan, A. (2004). Implementation of a formative assessment model incorporating peer and self-assessment. *Innovations in Education and Teaching International, 41*(3), 273-290. doi:10.1080/14703290410001733294

Orsmond, P., Merry, S. & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 27*(4), 309-323. doi:10.1080/0260293022000001337

Os, W. van, & Beek, M. van (2012). Toetskwaliteit in objectieve zin en volgens het oordeel van studenten: Een casestudy. *Tijdschrift voor Hoger Onderwijs 30*(4), 259-269.

Oudkerk-Pool, I. (2013). *Expertiseontwikkeling en professionalisering van de assessor: Ontwikkeling van een rubriek voor assessorenkwaliteit criteriumgericht beoordelen*. Amsterdam: Hogeschool van Amsterdam, Kenniscentrum Onderwijs en Opvoeding.

Pelgrum, M. B. Th. A. (2009). Competenties ontwikkelen – competenties beoordelen: De Stichting Consortium Beroepsonderwijs - zorg en welzijn. *Examens, 2*, 21-23.

Perie, M., Marin, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*(3), 5-13.

Plake, B. S. (2002). Evaluating the technical quality of educational tests used for high-stakes decisions. *Measurement and Evaluation in Counseling and Development, 35*(3), 144-152. Verkregen van: http://eric.ed.gov/?id=EJ657132

Plake, B. S., Impara, J. C., & Buckendahl, C. W. (2004). Technical quality criteria for evaluating district assessment portfolios used in the Nebraska STARS. *Educational Measurement: Issues and Practice, 23*(2), 12-16. doi:10.1111/j.1745-3992.2004.tb00154.x

* 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

Poldner, E., Simons, P. R. J., Wijngaards, G., & Schaaf, M. F. van der (2012). Quantitative content analysis of procedures to analyse students' reflective essays: A methodological review of psychometric and edumetric aspects. *Educational Research Review, 7*(1), 19-37. doi:10.1016/j.edurev.2011.11.002

Quellmalz, E. S. (1991). Developing criteria for performance assessments: The missing link. *Applied Measurement in Education, 4*(4), 319-331. doi:10.1207/s15324818ame0404_5

Rademacher, J. A. (2000). Involving students in assignment evaluation. *Intervention in School and Clinic, 35*(3), 151-156. doi:10.1177/105345120003500304

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*, 18-39. doi:10.1016/j.asw.2010.01.003

Riessen, M., van (2005). Digitale toets effectief. *Didactief, 35*(9), 10-11.

Roelofs, E. C. (2006). Een procesmodel voor de beoordeling van competent handelen. *Tijdschrift voor Hoger Onderwijs, 24*(3), 152-167.

Rom, M. C. (2011). Grading more accurately. *Journal of Political Science Education, 7*(2), 208-223. doi:10.1080/15512169.2011.564916

Ryan, G. J., Marshall, L. L., Porter, K., & Jia, H. (2007). Peer, professional and self-evaluation of class participation. *Active learning in higher education, 8*(1), 49-61. doi:10.1177/1469787407074049

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30*(2), 175-194. doi:10.1080/0260293042000264262

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159-179. doi:10.1080/02602930801956059

Sadler, D. R. (2009). Transforming holistic assessment and grading into a vehicle for complex learning. In: G. Joughin (Eds.), *Assessment, learning and judgement in higher education* (pp. 45-63). Springer Science + Business Media B.V.

Sadler, D. R. (2010). Fidelity as a precondition for integrity in grading academic achievement. *Assessment & Evaluation in Higher Education, 35*(6), 727-743. doi:10.1080/02602930902977756

Sambel, K., McDowell, L., & Brown, S. (1997). But is it fair? An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation, 23*(4), 349-371. Verkregen van: http://eder603.wikispaces.com/file/view/but+is+it+fair.pdf

Sanders, P. (2013). De kwaliteit van de staatsexamens voortgezet onderwijs: Sterke en zwakke aspecten van staatsexamens. *Examens, 4*, 25-27.

* 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.

Schaaf, M. F. van der, & Stokking, K. M. (2008). Developing and validating a design for teacher portfolio assessment. *Assessment & Evaluation in Higher Education, 33*(3), 245-262. doi:10.1080/02602930701292522

Schuwirth, L. W. T., & Vleuten, C. P. M., van der (2004). Changing education, changing assessment, changing research? *Medical Education, 38*(8), 805-812. doi:10.1111/j.1365-2929.2004.01851.x

* Segers, M. (2004). Assessment en leren als twee-eenheid: Onderzoek naar de impact van assessment op leren. *Tijdschrift voor Hoger Onderwijs, 22*(4), 188-219.

Segers, M., Dierick, S., & Dochy, F. (2001). Quality standards for new modes of assessment. An exploratory study of consequential validity of the OverAll Test. *European Journal of Psychology of Education, 16*(4), 569-588.

Segers, M., & Dochy, F. (2001). New assessment forms in problem-based learning: The value-added of the students' perspective. *Studies in Higher Education, 26*(3), 327-343. doi:10.1080/0307507012007629 1

Shalem, Y., & Slonimsky, L. (2010). Seeing epistemic order: Construction and transmission of evaluative criteria. *British Journal of Sociology of Education, 31*(6), 755-778. doi:10.1080/01425692.2010.515106

* 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

* Sijtsma, K. (2011). Studietoetsen: zinvolle inhoud of betrouwbare score? De standaardmeetfout is belangrijker dan de betrouwbaarheid. *Examens, 4*, 5-8.

Silva, M., Munk, D. D., & Bursuck, W. D. (2005). Grading adaptations for students with disabilities. *Intervention in School and Clinic, 41*(2), 87-98. doi:10.1177/10534512050410020901

Sinke, G. P. J. (2006). Deel IV Kwaliteitsbewaking. In: Sinke, G. P. J., *Aan de slag met assessment: Toetsen en beoordelen in competentiegerichte leeromgeving* (pp. 181-207). Nuenen: Onderwijsadviesbureau drs. M.A.F. Dekkers B.V.

Sireci, S. G., Hauger, J. B., Wells, C. S., Shea, C., & Zenisky A. L. (2009). Evaluation of the standard setting on the 2005 grade 12 national assessment of educational progress mathematics test. *Applied Measurement in Education, 22*, 339-358. doi:10.1080/08957340903221659

Sluijsmans, D. M. A. (2002). *Student involvement in assessment: The training of peer assessment skills.* Unpublished doctoral dissertation, Open Universiteit Nederland, Heerlen, the Netherlands.

Sluijsmans, D. M. A. (2003). Peerassessment en de ontwikkeling van reflectievaardigheden in de lerarenopleiding. *Tijdschrift voor Hoger Onderwijs, 21*(4), 230-249.

Sluijsmans, D. M. A. (2008). *Betrokken bij beoordelen*. Lectorale rede. Nijmegen: Hogeschool Arnhem en Nijmegen.

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

Sluijsmans, D., & Behrend, D. (red.) (2013). *Verantwoord toetsen en beslissen in het hoger beroepsonderwijs: Een voorstel voor een programma van eisen voor een basis- en seniorkwalificatie examinering (BKE/SKE).* Den Haag: Vereniging Hogescholen.

Sluijsmans, D. M. A., Brand-Gruwel, S., Merriënboer, J. J. G. van, & Martens, R. L. (2004). Training teachers in peer-assessment skills: effects on performance and perceptions. *Innovations in Education and Teaching International, 41*(1), 59-78. doi:10.1080/1470329032000172720

Sluijsmans, D. M. A., Eldik, S. van, Joosten-ten Brinke, D., & Jakobs, L. (2013). Bewust en bekwaam toetsen: Een eerste invulling van een kennisbasis toetsen voor lerarenopleiders. *Tijdschrift voor Lerarenopleiders, 34*(3), 27-42.

Sluijsmans, D., Martens, R., & Verheijen, H. (2000). Peer-assessment en onderwijsontwerp. *OnderwijsInnovatie, 2*, 17-24.

Sluijsmans, D., Peeters, A., Jakobs, L., & Weijzen, S. (2012). De kwaliteit van toetsing onder de loep. *OnderwijsInnovatie, 4*, 17-25.

Sluijsmans, D. M. A., Prins, F. J., & Martens, R. L. (2006). The design of competency-based performance assessment in e-learning. *Learning Environments Research, 9*(1), 45-66. doi:10.1007/s10984-005-9003-3

Sluijsmans, D. M. A., Martens, R. L., & Verheijen, H. (2000). Peer-assessment en onderwijsontwerp. *OnderwijsInnovatie, 2*, 17-24.

Sluijsmans, D. M. A., Straetmans, G. J. J. M., & Van Merriënboer, J. J. G. (2008). Integrating authentic assessment with competency based learning: The protocol portfolio scoring. *Journal of Vocational Education and Training, 60*(2), 157-172. Verkregen van: http://www.tandf.co.uk/journals/titles/13636820.asp

Smith, K., & Tillema, H. (2007). Use of criteria in assessing teaching portfolios: Judgemental practices in summative evaluation. *Scandinavian Journal of Educational Research, 51*(1), 103-117. doi:10.1080/00313830601078696

Sol, Y. (2012). De kunst van het leerzaam beoordelen. *Didactief, 42*(8), 24-25.
Stam, H. van der (2006). En we noemen het assessment…: Over nut en noodzaak van gedragsproeven. *Examens, 3*, 11-13.
Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice, 6*(3), 33-42. doi:10.1111/j.1745-3992.1987.tb00507.x
Stokking, K., Schaaf, M. van der, Jaspers, J., & Erkens, G. (2004). Teachers' assessment of students' research skills. *British Educational Research Journal, 30*(1), 93-115. doi:10/1080.01411920310001629983
Straetmans, G. J. J. M. (2004). Protocol portfolio scoring. *OnderwijsInnovatie, 2*, 17-27.
* Straetmans, G. J. J. M. (2006). *Bekwaam beoordelen en beslissen*. Lectorale rede. Deventer: Saxion Hogescholen.
Straetmans, G. J. J. M. (2014). *Beoordelen van competentie*. Manuscript in preparation.
Straetmans, G. J. J. M., & Eggen, T. J. H. M. (2011). WISCAT-pabo: Ontwerp, kwaliteit en resultaten van een geruchtmakende toets. In: Schramade, P. (Red.), *Handboek Effectief Opleiden* (pp. 55- 63). Den Haag: Reed Business Information.
Straetmans, G. J. J. M., Roelofs, E., & Peters, M. (2011). Vaststellen van didactische bekwaamheid tijdens de LIO-stage. *Tijdschrift voor lerarenopleiders, 32*(1), 4-11.
Straetmans, G. J. J. M., Sluijsmans, D., Bolhuis, B. G., & Merrienboer, J. J. G., van (2003). Integratie van instructie en assessment in competentiegericht onderwijs. *Tijdschrift voor Hoger Onderwijs, 21*(3), 170-197.
Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education, 30*(4), 325-341. doi:10.1080/02602930500099102
Strijbos, J. (2013). Een kwaliteitsborgingssysteem van toetsing: Van beleid naar praktijk. *Examens, 3*, 22-26.
Suskie, L. (2006). *Five dimensions of good assessment*. Presentation given at the 7th Annual Texas A&M Assessment Conference, February, 2007.
* 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
Tierney, R. D., Simon, M., & Charland, J. (2011). Being fair: Teachers' interpretations of principles for standards-based grading. *The Educational Forum, 75*(3), 210-227. doi:10.1080/00131725.2011.577669
Tigelaar, D. E. H., Dolmans, D. H. J. M., Wolfhagen, I. H. A. P., & Vleuten, C. P. M. van der (2005). Quality issues in judging portfolio's: Implications for organizing teaching portfolio assessment procedures. *Studies in Higher Education, 30*(5), 595-610. doi:10.1080/03075070500249302
Tillema, H. (2008). Wat is er mis met beoordelen? *Pedagogische Studiën, 85*, 294-304.
* 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
Tillema, H., & Smith, K. (2007). Portfolio appraisal: In search of quality criteria. *Teaching and Teacher Education, 23*, 442-456. doi:10.1016/j.tate.2006.12.005
Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In: *Optimistic new modes of assessment: In search of qualities and standards* (pp. 55-87). Dordrecht/Boston/London: Kluwer Academic Publishers.
Vleuten, C. P. M. van der, & Driessen, E. W. (2000). *Toetsing in probleemgestuurd onderwijs*. Groningen: Wolters-Noordhoff.
Vleuten, C. P. M. van der, & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309-317. doi:10.1111/j.1365-2929.2005.02094.x
Vleuten, C.P.M. van der, Schuwirth, L.W.T., Driessen, E.W., Dijkstra, J., Tigelaar, D., Baartman, L.K.J., & Tartwijk, J. van (2012). A model for programmatic assessment fit for purpose. *Medical Teacher, 34*, 205-214. doi:10.3109/0142159X.2012.652239

Vleuten, C. P. M. van der, Schuwirth, L. W. T., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practice & Research Clinical Obstetrics and Gynaecology, 24*, 703-719. doi:10.1016/j.bpobgyn.2010.04.001

Welther, L. (2010). Een wirwar van toetsen. *Didactief, 40*(9), 18-19.

* White, R. (2007). Balance in assessment. *Measurement: Interdisciplinary Research and Perspectives, 5(*1), 65-67. doi:10.1080/15366360701293634

Williamson, D. M., Bennett, R. E., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., Rubin, D. P. et al. (2010). *Automated scoring for the assessment of common core standards.*

Woolf, H. (2004). Assessment criteria: Reflections on current practices. *Assessment & Evaluation in Higher Education, 29*(4), 479-493. doi:10.1080/02602930310001689046

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

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice, 29*(4), 15-27. doi:10.1111/j.1745-3992.2010.00190.x

Wubs, W. L. (2013). *Opvattingen over toetsing van jonge kinderen: Een kwalitatief onderzoek naar meningen van leerkrachten, leidsters, ouders en kinderen over toetsing van taal en rekenen bij kleuters en peuters*. Unpublished master's thesis, Universiteit Twente, Enschede, The Netherlands.

Yao, Y., Thomas, M., Nickens, N., Downing, J. A., Burkett, R. S., & Lamson, S. (2008). Validity evidence of an electronic portfolio for preservice teachers. *Educational Measurement: Issues and Practice, 27*(1), 10-24. doi:10.1111/j.1745-3992.2008.00111.x

Yin, A. C., & Volkwein, J. F. (2009). Assessing general education outcomes. In J. F. Volkwein (ed.), *Assessing Student Outcomes: Why, who, what, how? New Directions for Institutional Research Assessment Supplement* (pp. 79-100). San Francisco: Jossey-Bass.

Yorke, M. (2011). Summative assessment: Dealing with the 'measurement fallacy'. *Studies in Higher Education, 36*(3), 251-273. doi:10.1080/03075070903545082

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

Yun, W. (2003). How raters' and writers' perceptions of a topic affect the scoring of compositions. *College Teaching, 51*(3), 115-118. doi:10.1080/87567550309596424

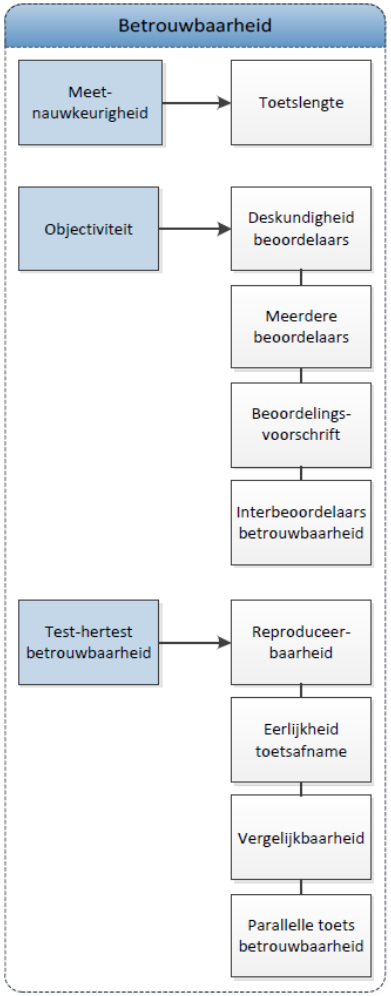
Zandsteeg, G. A. B., & Schaaf, M. F., van der (2014). Valideren van examens: De 'argument based approach' voor valideren. *Examens, 1*, 17-21.

Zoeckler, L. G. (2005). *Moral dimensions of grading in high school English*. Unpublished doctoral disstertation, University of Indiana. UMI No. AAT3183500.

Zutven, G. van, Polderdijk, M., & Voider, M. de (2004). *Toetsplanontwikkeling in competentiegericht onderwijs: Beleid voor verantwoord plannen van toetsing en examinering in het hoger onderwijs*. Utrecht: Stichting Digitale Universiteit.

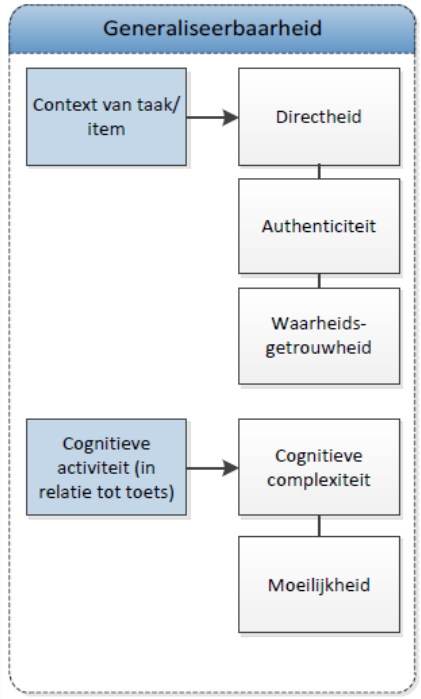
APPENDIX A. DEFINITIES KWALITEITSASPECTEN BEGRIPPENKADER

Betrouwbaarheid = De betrouwbaarheid is de mate waarin men staat kan maken op meetresultaten: de mate waarin de scores consistent, nauwkeurig en reproduceerbaar zijn. Kortom: de meetresultaten zijn vrij van meetfouten.



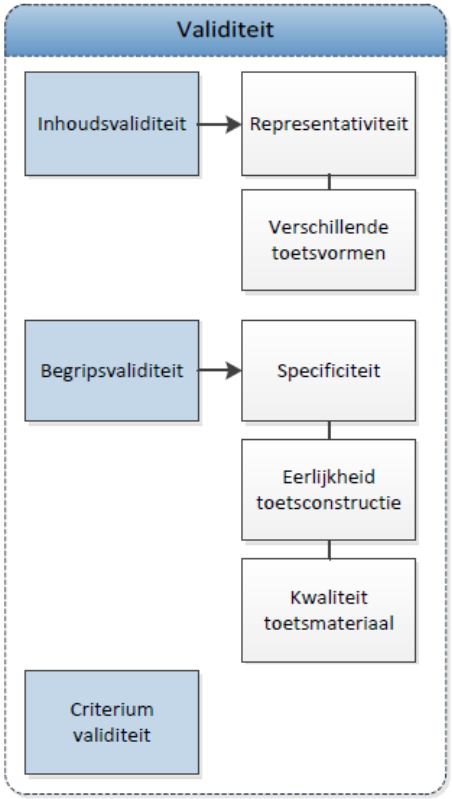
Meetnauwkeurigheid	De mate van precisie van een meting. Hieronder valt ook *accuraatheid* (mate van precisie) en *interne consistentie* (mate waarin de opgaven van een toets onderling samenhangen in statistische zin).
Toetslengte	De hoeveelheid items die een toets bevat.
Objectiviteit	De informatie die we in de toets verzamelen en de conclusies die we eruit trekken kunnen geheel aan de capaciteit van de student worden toegeschreven. De beoordelingen zijn dan niet subjectief: dus niet beïnvloed door de persoon (docent, zelf, klasgenoot) die de beoordeling deed en de conclusies trok. Hieronder valt ook *voorkomen van een dubbelrol beoordelaar versus begeleider*.
Interbeoordelaarsbetrouwbaarheid	Een maat om de objectiviteit van beoordelaars te meten: de mate waarin overeenstemming tussen beoordelaars wordt bereikt.
Deskundigheid beoordelaars	De mate waarin een beoordelaar de expertise heeft om een kwalitatief goede beoordeling te geven van een gemaakte toets. Hieronder valt ook *training beoordelaars, vertrouwen van student in beoordelaar* en *samen bespreken van de criteria met de beoordelaars.*
Meerdere beoordelaars	Het aantal beoordelaars dat een toets beoordeeld.
Beoordelingsvoorschrift	Er zijn beoordelingscriteria waarop een student wordt beoordeeld.
Test-hertest betrouwbaarheid	Het onderzoeken van de betrouwbaarheid van een toets door de toets letterlijk te herhalen, op dezelfde manier en bij dezelfde onderzoekseenheden.
Reproduceerbaarheid	De mate waarin een toets hetzelfde resultaat geeft wanneer deze opnieuw wordt afgenomen. Hieronder valt ook het *afnemen van dezelfde toets op meerdere momenten*.
Eerlijkheid (afname conditie)	Iedereen heeft gelijke kansen bij de afname van een toets onafhankelijk van afkomst, geslacht, leeftijd, SES, etc. Hieronder valt ook *opportunity to learn*.
Vergelijkbaarheid	De toetsing wordt op consistente wijze uitgevoerd. De omgevingsinvloeden zijn constant.
Parallelle toets betrouwbaarheid	Het onderzoeken van de betrouwbaarheid van een toets door verschillende versies van het instrument te construeren en de overeenkomst tussen de resultaten te meten. Hieronder valt ook dat er *verschillende versies* van de toets worden gebruikt met hetzelfde format.

Generaliseerbaarheid (externe validiteit) = De generaliseerbaarheid (externe validiteit) van de conclusies is de mate waarin datgene wat er in het onderzoek wordt gevonden (in deze specifieke omstandigheden en bij deze steekproef), ook opgaat in andere omstandigheden en voor andere individuen. Hieronder valt ook *ecologische validiteit* (de mate waarin de omstandigheden in het onderzoek overeenkomen met omstandigheden in het gewone leven).



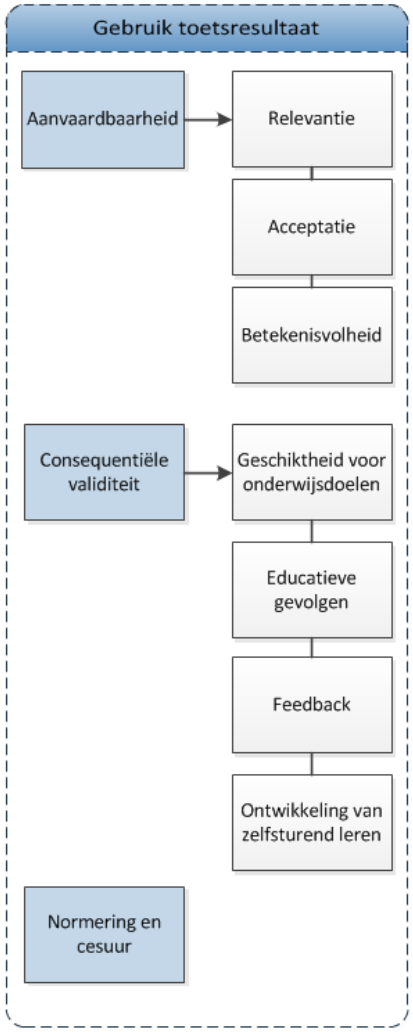
Context van taak/item	
Directheid	De mate waarin docenten/assessoren direct de toetsresultaten kunnen interpreteren, zonder deze te moeten vertalen van theorie naar praktijk.
Authenticiteit	De mate van overeenkomst van een toetsprogramma met het toekomstige beroep. Een toetsprogramma moet overeenkomen met de beroepspraktijk. Hieronder valt ook *extrapoleerbaarheid*. Dit wil zeggen dat de behaalde prestaties op een toets niet alleen betekenis hebben in de context van de toets, maar ook op situaties buiten de toets zoals de beroepspraktijk. De score weerspiegelt het prestatieniveau in de reële werksituatie.
Waarheidsgetrouwheid	De mate waarin elementen van de toets die zorgen voor het cijfer ook werkelijk aan prestatie is toe te schrijven.
Cognitieve activiteit (in relatie tot toets)	De mate waarin een toets een beroep doet op het denkvermogen van de student.
Cognitieve complexiteit	Dit heeft te maken met de overeenkomst van een toetsprogramma met de toekomstige beroepssituatie, maar is meer gericht op de denkprocessen. Een beoordelaar moet kunnen vaststellen of iemand de cognitieve vaardigheden bezit om in een beroep te kunnen functioneren.
Moeilijkheid	De toets moet overeenstemmen met het niveau van de studenten voor wie de toets bestemd is. Een te moeilijke of te makkelijke toets zal een weinig selecterende functie hebben.

Validiteit = De validiteit van een toets is de eigenschap dat de toets meet wat de constructeur bedoeld heeft ermee te meten. Aangezien een toets veel en uiteenlopende bedoelingen kan hebben zijn er evenzoveel validiteiten te onderscheiden, die een toets in verschillende mate kan bezitten.



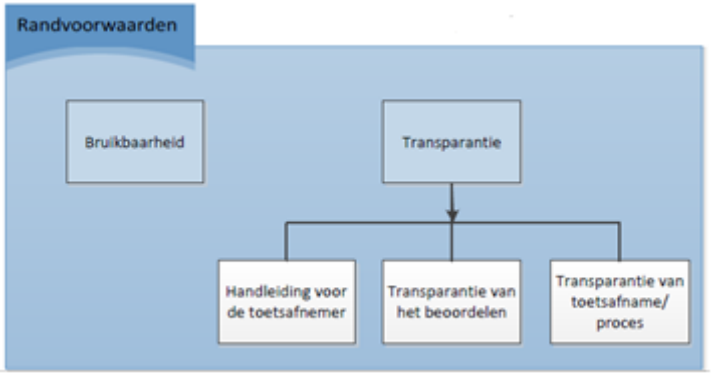
Inhoudsvaliditeit	Inhoudsvaliditeit van een toets is de eigenschap dat de opgaven een representatieve steekproef vormen van opgaven voor de te toetsen kennis of vaardigheid. Meestal wordt gedoeld op de representativiteit qua leerstofgebied.
Representativiteit	De toets bestrijkt voldoende de gehele te bestuderen stof. Het is breder dan de term inhoudsvaliditeit: het vormt de operationalisatie van inhoudsvaliditeit. Hieronder valt ook *toetsmatrijs* en *evenwichtigheid* (het aantal vragen in een toets over een bepaald onderwerp is in verhouding met het belang van dat onderwerp).
Verschillende toetsvormen	Het gebruik van verschillende soorten toetsen of instrumenten om tot een score of oordeel te komen, waarbij het format verschillend is. Een voorbeeld is het gebruik van een observatie en een schriftelijke toets.
Begripsvaliditeit	De eigenschap die een toets heeft als kan worden aangetoond dat de toets het door de constructeur beoogde kenmerk van de student (onderliggende trek, vaardigheid) meet.
Specificiteit	Alleen die studenten die de stof bestudeerd hebben kunnen tot een goede oplossing komen. In dit geval dient er bijvoorbeeld voor gezorgd te worden dat studenten geen aanknopingspunten voor het goede antwoord kunnen vinden in de vraagstelling zelf of andere vragen. Het gaat hier dus om een goed discriminerend of onderscheidend vermogen.
Eerlijkheid toetsconstructie	De mate waarin toetsitems discriminerende aspecten bevatten.
Kwaliteit toetsmateriaal	Alle aspecten die te maken hebben met de kwaliteit van het toetsmateriaal en de (wettelijke) eisen die hieraan gesteld worden. Hieronder valt ook *vorm*, *leesbaarheid*, *taalgebruik*, *gebruiksgemak*, etc.
Criteriumvaliditeit	De term criteriumvaliditeit heeft betrekking op de mate waarin de uitkomst van een instrument samenhangt met een of meer criteriumvariabelen. Criteriumvariabelen zijn de dingen die je eigenlijk had willen meten, maar om een of andere reden niet of moeilijk rechtstreeks kunt vaststellen. Om de criteriumvaliditeit van een instrument te kunnen vaststellen moet je het begrip op een andere wijze meten. Dit kan door een vergelijkings-instrument: de mate waarin de test voorspellend is voor latere test (predictief) of de samenhang met gegevens op dit moment (concurrent).

Gebruik toetsresultaat = Hoe worden de toetsresultaten verwerkt en wat wordt er gedaan met deze scores?



Aanvaardbaarheid	De mate waarin studenten de toets relevant, acceptabel en betekenisvol vinden.
Relevantie	Iets dat van betekenis is in de gegeven situatie. Hieronder valt ook *recent.*
Acceptatie	Alle betrokkenen bij een toetsprogramma kunnen zich kunnen vinden in de taken of opdrachten, de criteria en de uitvoering.
Betekenisvolheid	Betekenisvolheid heeft betrekking op hoe zinvol en nuttig de betrokkenen de toetsing vinden. Goede toetsing is zinvol voor studenten omdat ze er wat van leren of omdat ze inzicht krijgen in hun sterke en zwakke kanten.
Consequentiële validiteit	Consequentiële validiteit van een toets beschrijft de gewenste en ongewenste effecten van een toets op het leren van studenten. De manier van toetsen is bijvoorbeeld een belangrijke sturende factor: er wordt anders gestudeerd voor een multiple-choice toets dan voor een mondeling examen. Hieronder valt *toetsmisbruik* en *langere termijn gebruik van de scores*.
Geschiktheid voor onderwijsdoelen	De afstemming tussen doelen van de toetsing en de doelen van het onderwijs. Hieronder valt ook het *voldoen aan wettelijke eisen* en de *congruentie* tussen toets, instructie en doel van de toets.
Educatieve gevolgen	De bedoelde en onbedoelde gevolgen, positieve en negatieve effecten van een toets op hoe docenten en studenten hun leren en lesgevende activiteiten aanpassen als gevolg van het resultaat van de toets. Dit heeft in vergelijking met consequentiële validiteit betrekking op de directe gevolgen, hoe docenten de feedback gebruiken voor hun onderwijs.
Feedback	De reactie die een student krijgt op de toets. Dit kan het cijfer zijn, maar ook een voortgangsproces wat bijhoudt of een student over een bepaalde vaardigheid beschikt. Daarnaast kunnen de snelheid, hoeveelheid en kwaliteit van de feedback een rol spelen bij het leerproces.
Ontwikkeling van zelfsturend leren	De studenten geven zelf vorm aan het leerproces. Toetsing kan hieraan bijdragen doordat er aandacht wordt besteed aan zelf- en peerbeoordelingen.
Normering en cesuur	De manier waarop de resultaten van de studenten worden omgezet in een cijfer. Hieronder valt ook de mate waarin verschillende onderdelen van een toets bijdragen aan het eindcijfer.

Randvoorwaarden = randvoorwaarden waaraan een goede toets moet voldoen.



Bruikbaarheid	De uitvoerbaarheid en efficiëntie van een toets (de verhouding tussen de inspanning die iemand moet leveren voor een toets en de nuttigheid van de uitkomst van deze toets). Hieronder valt ook *tijd*, *kosten* en *haalbaarheid* van het maken van de toets binnen de gestelde tijd.
Transparantie	Het is voor de student helder wat er van hem/haar wordt verwacht.
Transparantie m.b.t toetsafname/proces	Het is voor aanvang van een toets helder wat er van studenten verwacht wordt qua afname van een toets, zoals tijdstip, inhoud, vorm, toetslengte, beschikbare tijd, etc.
Transparantie m.b.t het beoordelen	Het is voor aanvang van een toets helder wat er van studenten verwacht wordt qua beoordeling van de toets, zoals cesuur, normering, beoordelingscriteria, informatie over beoordelaars, etc.
Handleiding voor de toetsafnemer	Er is een handleiding van de toets beschikbaar voor diegene die de toets afneemt. Deze is begrijpelijk voor de afnemer. Het is voor de afnemer helder hoe de toets moet worden afgenomen en hoe deze gescoord moet worden.

Aspecten die het gehele traject van belang zijn:

Toetsanalyse

Toetsanalyse: Het gebruiken van een methode om voor- of achteraf de kwaliteit van de toets te onderzoeken en te analyseren. Hieronder valt ook *kwaliteitsborging*.

Theoretische onderbouwing

Theoretische onderbouwing: Theoretische onderbouwing van de toets, zoals het vaststellen van het doel, doelgroep, operationalisatie, etc. Hieronder valt ook *argumentatieve validiteit*.

APPENDIX B. SEMIGESTRUCTUREERD INTERVIEW KLANKBORDGROEP

1. Wat kunt u vertellen over toetsing op uw school?
    - Welke vormen toetsen zijn er?
    - Hebben toetsen vooral open vragen of meerkeuze vragen?
    - Is er een nadruk op toetsen om het leerproces bij te stellen en/of op toetsen om te examineren?
    - Worden alle toetsen nagekeken door de docent zelf? Of is er tevens sprake van zelfbeoordelingen, beoordelingen door studiegenoten of een beoordelingen door docent en student samen?
    - Wie maakt de toetsen?

2. Is er aandacht voor kwaliteit van toetsen op uw school?
    - Zo ja, hoe uit zich dat?
    - Zo nee, waarom niet denkt u?

3. Hoe wordt de kwaliteit op uw school geborgd? Wat wordt er aan de kwaliteit van toetsing gedaan?

4. In het begrippenkader worden verschillende aspecten genoemd.
    - Met welke aspecten van toetskwaliteit wordt op uw school rekening gehouden?

5. Zijn er problemen qua kwaliteit van toetsing?
    - Zo ja, welke problemen ervaart u ten aanzien van kwalitatief goed toetsen?

6. Heeft u behoefte om meer te weten over de kwaliteit van toetsen?
    - Zo ja, over welke onderwerpen?
        - Op welke manier zou u meer willen weten?

Kwaliteit van toetsen binnen handbereik

Een reviewstudie van onderzoek en onderzoeksresultaten naar de kwaliteit van toetsen