



Er zijn meerdere
interpretaties van
kwaliteit van toetsing

Kwaliteit van toetsing in het tertiaire onderwijs

Een conceptualisering op basis van een literatuurstudie

■ Karin J. Gerritsen-van Leeuwenkamp

Dr. K.J. Gerritsen-van Leeuwenkamp is adviseur onderwijs en onderzoek bij de Dienst Onderwijs en Student Support, Bureau Kwaliteitszorg van Saxion. Emailadres: k.j.vanleeuwenkamp@saxion.nl

■ Desirée Joosten-ten Brinke

Dr. D. Joosten-ten Brinke is universitair hoofddocent bij het Welten Instituut van de Open universiteit en lector bij Fontys lerarenopleiding Tilburg.

■ Liesbeth Kester

Prof. dr. L. Kester is hoogleraar Onderwijswetenschappen bij de afdeling Educatie van de faculteit Sociale Wetenschappen, Universiteit Utrecht.

De kwaliteit van toetsing beïnvloedt de geschiktheid, accuraatheid en geloofwaardigheid van informatie die verzameld wordt voor de functies van leren, selecteren en certificeren en verantwoording. Hoewel onderwijsorganisaties verantwoordelijk zijn voor het garanderen van de kwaliteit van toetsing, is het niet altijd duidelijk wat verstaan wordt onder kwaliteit van toetsing. Dit artikel beschrijft een conceptualisering van kwaliteit van toetsing op basis van een literatuurstudie en legt uit wat deze conceptualisering betekent voor het realiseren van kwaliteit van toetsing in de dagelijkse onderwijspraktijk.

Consequenties van onvoldoende kwaliteit van toetsing

Kwaliteit van toetsing in het tertiaire onderwijs¹ richt zich op meer dan alleen de toets. Het richt zich op allerlei aspecten waaronder het toetsbeleid, toetsprogramma, toetsen, toetsvragen en opdrachten (Joosten-ten Brinke, 2011). De gevolgen van onvoldoende kwaliteit van toetsing zijn groot en variëren afhankelijk van de functie van toetsing: leren, selecteren en certificeren en verantwoording (Anderson & Rogan, 2010; Stobart, 2008). Zo kan onvoldoende kwaliteit van toetsing ervoor zorgen dat het onbekend is waar studenten staan in hun leren, waar ze naar toe

werken en op welke wijze ze daar het beste kunnen komen (ARG, 2002). Onvoldoende kwaliteit van toetsing kan ook zorgen voor onjuiste beslissingen over zakken en slagen als basis waarop studenten geselecteerd of gecertificeerd worden (Anderson & Rogan, 2010; Stobart, 2008). Dan slaagt een student bijvoorbeeld voor een rekentoets terwijl hij niet rekenvaardig is. Ook kan onvoldoende kwaliteit van toetsing de behaalde niveaus beïnvloeden van studenten waar onderwijsinstellingen verantwoordelijk voor worden gehouden (Stobart, 2008). De maatschappij stelt immers eisen aan het niveau van afgestudeerden en een diploma moet dit niveau garande-

¹Tertiaire onderwijs is het hoger onderwijs en het middelbaar beroepsonderwijs

ren. Om voorgaande gevolgen te voorkomen, besteden onderwijsinstellingen veel tijd aan diverse maatregelen om kwaliteit van toetsing te evalueren, garanderen en verbeteren.

Uitdagingen bij het realiseren van kwaliteit van toetsing

Het is voor onderwijsinstellingen moeilijk om kwaliteit van toetsing te realiseren omdat er in de literatuur of in de dagelijkse onderwijspraktijk niet één overkoepelende conceptualisering is van kwaliteit van toetsing. Belanghebbenden (waaronder bijvoorbeeld studenten en medewerkers van een onderwijsinstelling) hanteren vanuit hun eigen perspectief verschillende definities van kwaliteit. Dit maakt het moeilijk om eenduidig vast te stellen wat kwaliteit van toetsing is en wat nodig is om kwaliteit van toetsing te bereiken. Daarom biedt dit artikel inzicht in een conceptualisering van kwaliteit van toetsing op basis van een literatuurstudie.

Conceptualisering van kwaliteit van toetsing

Door middel van een literatuurstudie² is nagegaan wat er in peer-reviewed wetenschappelijke artikelen wordt verstaan onder kwaliteit van toetsing (Gerritsen-van Leeuwenkamp et al., 2017). Er zijn 78 artikelen geselecteerd en geanalyseerd. Tekstfragmenten zijn aan de hand van een vooraf opgesteld framework gecodeerd. Daarnaast zijn er conceptuele en relationele tekstanalyses met de software Leximancer uitgevoerd (Leximancer Pty Ltd, 2008). Dit betekent dat de aanwezigheid en frequentie van concepten in de gecodeerde tekstfragmenten en de relaties tussen de concepten objectief zijn vastgesteld. Hoe belangrijk een concept is wordt bepaald door de mate waarin een concept relaties heeft met andere concepten en hoe vaak het concept wordt genoemd in de tekstfragmenten. Concepten die samenhangen worden automatisch geclusterd in thema's³. Dit is door de software visueel gemaakt in zogenoemde 'concept

maps'. De geautomatiseerde tekstanalyses resulteerden in een conceptualisering van kwaliteit van toetsing bestaande uit: 1) drie thema's met kwaliteitscriteria; 2) vier thema's met invloeden; 3) twee thema's gericht op de evaluatie van de kwaliteitscriteria; en 4) vier thema's met de perspectieven van belanghebbenden die meegenomen zouden moeten worden in de evaluatie van de kwaliteitscriteria. Deze thema's worden in de volgende paragrafen bondig toegelicht aan de hand van voorbeelden. Bij elke paragraaf is een globale 'concept map' als figuur opgenomen. Een volledig, gedetailleerd en interactief overzicht van alle thema's is te vinden op de website www.assessmentquality.com. Zie voor een screenshot Figuur 1.

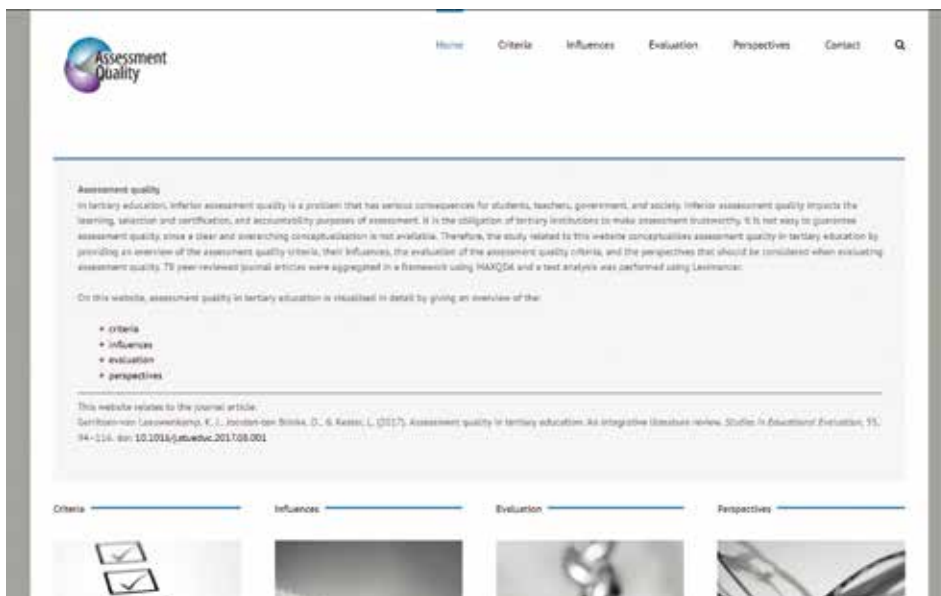
Drie thema's met kwaliteitscriteria

Een kwaliteitscriterium geeft een beschrijving van de kwaliteit van toetsing. Op basis van de tekstanalyse zijn 98 kwaliteitscriteria van toetsing gegroepeerd in drie thema's met kwaliteitscriteria: validiteit, transparantie en betrouwbaarheid (zie Figuur 2).

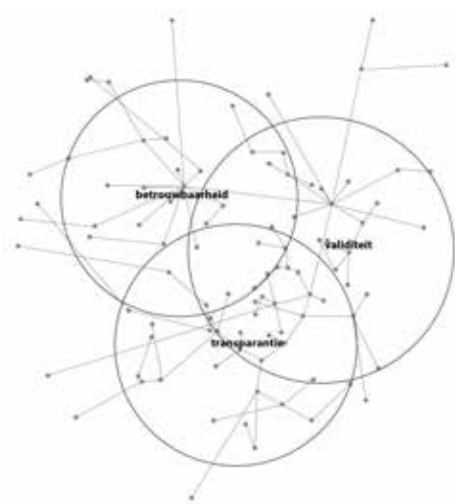
Het eerste thema validiteit wordt 446 keer genoemd (n=446) in de tekstfragmenten van de geselecteerde artikelen. Er zijn verschillende definities gegeven van het concept validiteit. Validiteit wordt door verschillende auteurs gezien als een kenmerk van de toets dat vaststelt of een toets meet wat hij zou moeten meten (bijvoorbeeld Ebel, 1983). Bijvoorbeeld 'meet de rekentoets rekenvaardigheid?'. Soms wordt deze definitie ook wel inhoudsvaliditeit genoemd en worden er daarnaast nog andere typen van validiteit onderscheiden zoals 'face validity'. Bijvoorbeeld 'meet de rekentoets op het eerste gezicht rekenvaardigheid?'. Messick (1995) doet het voorstel om constructvaliditeit te gebruiken als overkoepelende term. Hij geeft aan dat het onderscheiden van verschillende typen van validiteit voor een onvolledig beeld kan zorgen. Validiteit gaat dan over het

²Raadpleeg voor de gehele literatuurstudie inclusief een gedetailleerde beschrijving van de methode Gerritsen-van Leeuwenkamp, Joosten-ten Brinke, en Kester (2017).

³Een thema is een cluster van concepten (Leximancer, 2008).



Figuur 1. Een screenshot van de website www.assessmentquality.com.



Figuur 2. De thema's van kwaliteitscriteria. De cirkels geven de thema's weer. De grijze stipjes geven de kwaliteitscriteria weer die gegroepeerd zijn in het thema. De lijnen geven de relaties weer.

op een zinvolle en verantwoorde wijze interpreteren en gebruiken van toetsscores (Kane, 2001; Messick, 1995). Bijvoorbeeld 'zegt de score op een rekentoets daadwerkelijk iets

over de rekenvaardigheid van een student?'. Dit betekent dat in plaats van de vraag te beantwoorden: "meet de toets wat hij moet meten?" (Shepard, 1993, pp. 409–410) de centrale vraag wordt "doet de toets wat hij zou moeten doen?" (Shepard, 1993, p. 444). In het gebruik van assessments en toetsprogramma's roepen sommige onderzoekers op om extra kwaliteitscriteria te onderscheiden. Met als belangrijkste redenen het voorkomen van containerconcepten en het bruikbaar maken van het concept validiteit voor de dagelijkse praktijk (bijvoorbeeld Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2007). Denk bijvoorbeeld aan het kwaliteitscriterium 'onderwijsconsequenties' (Linn, Bakker, & Dunbar, 1991). Maar ook in deze kwaliteitscriteria is de definitie van constructvaliditeit nog steeds herkenbaar aanwezig (Baartman et al., 2007).

Het tweede thema is transparantie ($n = 74$). Transparantie richt zich op de begrijpelijkheid van toetsing voor belanghebbenden, het voorkomt onduidelijkheden over wat er wordt getoetst en op welke manier dat wordt

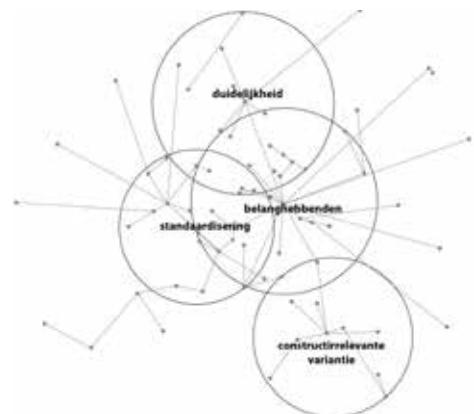
gedaan (Ploegh, Tillema, & Segers, 2009). Transparantie is een kenmerk van vrijwel alle elementen van toetsing, zoals de items, toetsprocedures, toetsen, beoordelingen en toetsprogramma's. Het is vereist dat belanghebbenden voldoende begrijpen waarop beoordeeld wordt, door wie beoordeeld wordt, hoe er beoordeeld wordt en wat het doel van de toetsing is (Baartman et al., 2007). Het derde en laatste thema is betrouwbaarheid ($n = 187$). Hoewel het vaker voorkomt in de literatuur dan transparantie, is het minder verbonden met de andere kwaliteitscriteria. Betrouwbaarheid is een kenmerk van de toetsscore (Downing, 2004). Bij gestandaardiseerde toetsen (bijvoorbeeld meerkeuze-toetsen) betekent het dat de toetsscore consistent, precies en stabiel is (Ebel, 1983). Wanneer meer gevarieerde assessments gebruikt worden (bijvoorbeeld performance assessments, mondelinge examens en probleemtaken), dan wordt de rol van het menselijke oordeel in de beoordelingen groter en de steekproef van situaties kleiner. Ook wisselen toetstaken meer doordat studenten in verschillende contexten de taken uitvoeren. Het is dan moeilijker om stabiliteit te bereiken en dus betrouwbaarheid (Van der Vleuten, 1996). Daarom stellen diverse auteurs voor om niet alleen naar de individuele toets te kijken, maar naar de betrouwbaarheid van toetsprogramma's waarin meerdere toetsen zijn opgenomen en meerdere beoordelaars een oordeel geven over de competentieverwerving van studenten (bijvoorbeeld Baartman et al., 2007). Betrouwbaarheid wordt bij gevarieerde assessments dan ook anders bereikt dan bij gestandaardiseerde toetsen. De focus ligt op de generaliseerbaarheid van beslissingen, waar bewijzen over de competentieontwikkeling van studenten verzameld worden vanuit diverse bronnen en op basis van een adequate steekproef van situaties (Schuwirth & Van der Vleuten, 2004). In het gebruik van assessments en toetsprogramma's stellen verschillende auteurs een range aan kwaliteitscriteria voor. Vergelijkbaar

met validiteit, is de belangrijkste reden dat zij betrouwbaarheid als containerconcept willen voorkomen en betrouwbaarheid bruikbaar willen maken voor de dagelijkse praktijk. Denk bijvoorbeeld aan het kwaliteitscriterium vergelijkbaarheid (Baartman et al., 2007).

Vier thema's met invloeden op de kwaliteitscriteria

Er zijn verschillende invloeden op de mate van validiteit, transparantie en betrouwbaarheid. In totaal kunnen er op basis van de literatuur 73 invloeden onderscheiden worden. Deze invloeden zijn met behulp van tekstanalyse geclusterd in vier thema's: 1) standaardisering, 2) belanghebbenden, 3) duidelijkheid en 4) constructirrelevante variantie (zie Figuur 3).

Het eerste thema is standaardisering ($n = 74$). Het gaat dan om standaardisering door procedures, richtlijnen, checklists, beoordelingsschalen en formats. Standaardisatie beïnvloedt alle drie de kwaliteitscriteria. Zo kunnen bijvoorbeeld criteria ervoor zorgen dat er daadwerkelijk getoetst wordt wat beoogd is (Van der Vleuten, 1996) en het kan studenten inzicht geven waarop ze beoordeeld worden



Figuur 3. De thema's van invloeden op de kwaliteitscriteria. De cirkels geven de thema's weer. De grijze stipjes geven de invloeden weer die gegroepeerd zijn in het thema. De lijnen geven de relaties weer.

(Holmes & Smith, 2003). Het gebruik van een rubric (een beschrijvende schaal) kan het beoordelen consistenten maken (Holmes & Smith, 2003).

Het tweede thema is de groep met belanghebbenden (n = 185). Belanghebbenden zijn onder andere de studenten en medewerkers van de onderwijsorganisatie. Zij beïnvloeden de kwaliteitscriteria door de manier waarop ze toetsen gebruiken, de mate waarin ze toetsing belangrijk vinden en de mate van betrokkenheid in en deskundigheid op het gebied van de verschillende fasen van het toetsproces (Baartman et al., 2007; Van der Vleuten et al., 2012). Bijvoorbeeld de mate van professionalisering van assessoren kan de betrouwbaarheid beïnvloeden.

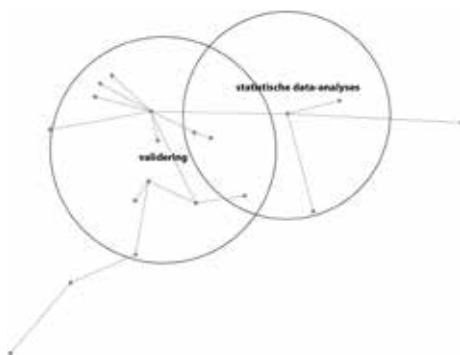
Het derde thema is duidelijkheid (n = 39). Het verwijst naar de duidelijkheid van onder andere procedures, richtlijnen, checklist, beoordelingsschalen en formats. Duidelijkheid beïnvloedt de validiteit, bijvoorbeeld als studenten aan het werk zijn aan een onduidelijke toetstaak geeft de toetsscore geen afspiegeling van hun daadwerkelijke niveau van presteren (MacLellan, 2004). Toetsen worden transparanter en consistenten wanneer belanghebbenden weten wat er beoordeeld wordt (Stowell, 2004).

Het vierde en laatste thema met invloeden is constructirrelevante variantie (n = 24). Hier van is sprake wanneer een toets meer meet dan het zou moeten meten. Het is één van de belangrijkste bedreigingen van het kwaliteitscriterium validiteit (Messick, 1995). Een bekend voorbeeld is wanneer het taalgebruik in een rekentoets te complex voor de doelgroep is. De toets meet dan naast rekenvaardigheid ook leesvaardigheid en wordt daarmee voor sommige studenten onbedoeld moeilijker. Maar een toets kan ook onbedoeld makkelijker worden voor studenten doordat er hints gegeven worden waardoor toetsvragen eenvoudiger te beantwoorden zijn.

Twee thema's gericht op de evaluatie van de kwaliteitscriteria

Ten derde bestaat de conceptualisering van kwaliteit van toetsing uit twee thema's gericht op de evaluatie (n = 22) van de kwaliteitscriteria validiteit en betrouwbaarheid: validering en statistische data-analyses (zie Figuur 4). In de literatuurstudie zijn geen methoden naar voren gekomen om het kwaliteitscriterium transparantie te evalueren.

Validering (n = 180) wordt gebruikt om de mate van validiteit te evalueren. Vanuit het perspectief van de inhoudsvaliditeit wordt nagegaan of de toets het kenmerk van validiteit heeft (Borsboom, Mellenbergh, & Van Heerden, 2004). Valideren bestaat dan bijvoorbeeld uit het aantonen van een oorzaak-gevolg relatie tussen de gemeten attributen en de toetsscore (Borsboom et al., 2004). Vanuit het oogpunt van de constructvaliditeit worden systematisch allerlei bewijzen verzameld die de interpretatie van de toetsscores ondersteunen en deze bewijzen rechtvaardigen de beslissingen die op basis van toetsscores genomen worden (Kane, 2008). Dit richt zich zowel op individuele toetsen als toetsprogramma's. Welke bewijzen nodig zijn, hangt af van de conclusies en aannames die gedaan



Figuur 4. De thema's gericht op de evaluatie van de kwaliteitscriteria. De cirkels geven de thema's weer. De grijze stipjes geven de concepten weer die gegroepeerd zijn in het thema. De lijnen geven de relaties weer.

worden in het interpretatieve argument. Voorbeelden van bewijzen zijn: toetsmatrizen, betrouwbaarheid van toetsscores, inzicht in de cognitieve processen van studenten terwijl zij de toets maken of in de consequenties van toetsing voor het leren van studenten (AERA, APA, & NCME, 2014).

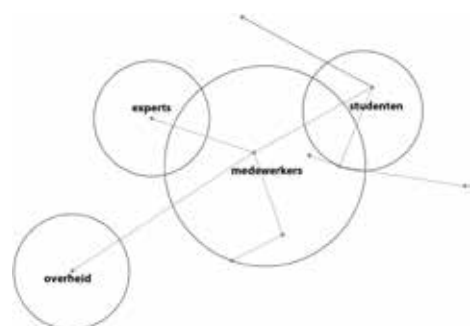
Het thema statistische data-analyses ($n = 146$) omvat allerlei methoden om de kwaliteitscriteria te evalueren. Zo kunnen analyses van toetsvragen in kennistoetsen bruikbare informatie geven over de kwaliteit van toetsvragen. Denk bijvoorbeeld aan vraagmoeilijkheid of itemdiscriminatie (Anderson & Rogan, 2010). Factoranalyses kunnen een bewijs zijn voor de validiteit (Birenbaum, 2007). Ook is een range aan statistische data-analyse methoden voor de betrouwbaarheid beschikbaar. Denk bijvoorbeeld aan test-hertestbetrouwbaarheid, waarbij de correlaties worden berekend tussen de eerste afname van de toets en de herafname van de toets (Barman, 2011). Aangezien deze betrouwbaarheidsmaat lastig is uit te voeren, omdat het kennisniveau van studenten niet constant is en het inefficiënt is, worden er alternatieven voorgesteld. Bijvoorbeeld de split-half methode, waarin de toets in tweeën wordt gedeeld en de correlaties tussen de twee helften worden berekend (Barman, 2011). Daarnaast zijn er nog methoden om de overeenstemming tussen beoordelaars te berekenen, zoals het percentage overeenstemming of de Kappa-coëfficiënt, waarbij rekening wordt gehouden met de overeenstemming tussen beoordelaars op basis van toeval (Downing, 2004).

Vier thema's met de perspectieven van belanghebbenden

Tot slot bestaat de conceptualisering van kwaliteit van toetsing uit vier thema's met de perspectieven van belanghebbenden die meegenomen worden in de evaluatie van de kwaliteitscriteria (zie Figuur 5). De perspectieven van belanghebbenden zijn relevant omdat zij de toetsing gebruiken in de dagelijkse

onderwijspraktijk en de consequenties ervan ervaren (Baartman et al., 2007). Op basis van verschillen in hun rollen en ervaring met toetsing hebben ze elk een eigen kijk op de kwaliteit van toetsing (Baartman, Prins, Kirschner, & Van der Vleuten, 2011). In de literatuur zijn 10 perspectieven van belanghebbenden onderscheiden. Deze konden worden geclusterd in vier thema's: studenten, medewerkers, overheid en experts.

Allereerst wordt het studentenperspectief ($n = 130$) genoemd in de literatuur. De perspectieven van studenten zijn relevant omdat zij de toetsing gebruiken en de directe consequenties ervaren. In de literatuur komen vooral de perspectieven van studenten op authenticiteit (behorende bij het thema validiteit), eerlijkheid (behorende bij het thema transparantie) en vraagmoeilijkheid naar voren. Bijvoorbeeld onderzoek van Gulikers, Kester, Kirschner, en Bastiaens (2008) liet zien dat het leren van studenten belemmerd kan worden wanneer zij een toetstaak niet authentiek vinden. Ten aanzien van eerlijkheid beschrijft Sambell, McDowell, en Brown (1997) dat eerlijkheid door studenten gebruikt wordt om hun waarde van de toetsing uit te drukken en dit hangt nauw samen met definitie van validiteit van toetsex-



Figuur 5. De thema's gericht op de perspectieven van belanghebbenden die meegenomen worden in de evaluatie van de kwaliteitscriteria. De cirkels geven de thema's weer. De grijze stipjes geven de perspectieven weer die gegroepeerd zijn in het thema. De lijnen geven de relaties weer.

perts. Ten aanzien van vraagmoeilijkheid liet Van de Watering en Van de Rijt (2006) zien dat studenten deze overschatten, terwijl docenten deze onderschatten.

Het tweede thema is het medewerkersperspectief ($n = 77$). Hiertoe behoren de managers, docenten, examencommissieleden en administratief medewerkers. Zij zijn veelal direct betrokken bij de constructie, evaluatie en goedkeuring van toetsing en hun perspectieven verschillen op basis van hun rollen en verantwoordelijkheden. Bijvoorbeeld uit onderzoek van Meyer et al. (2010) is naar voren gekomen dat managers zich vooral richten op het voorkomen van fouten en dat zij zich minder richten op een gemeenschappelijke visie ten aanzien van evidence-based toetsing. Ook blijkt uit hetzelfde onderzoek dat zowel managers als docenten bezorgd zijn over de beperkingen ten aanzien van tijd of werkbelasting die de kwaliteit van toetsing kunnen beïnvloeden (Meyer et al., 2010).

Het derde thema is het perspectief van de overheid ($n = 14$). Het belang van het perspectief van de overheid komt vooral voort uit de wetten en regels die zijn voorgeschreven. Deze wetten en regels kunnen de kwaliteit beïnvloeden (Meyer et al., 2010). Denk bijvoorbeeld aan beleid omtrent de financiering van het hoger onderwijs of politieke ontwikkelingen die verbeteringen initiëren (Dijkstra, Van der Vleuten, & Schuwirth, 2010).

Het vierde thema is het perspectief van de experts ($n = 10$). Dit zijn de vakinhoudelijke experts of experts van accreditatieorganen die met name betrokken zijn bij het construeren, reviewen of goedkeuren van de toetsing (Dijkstra et al., 2010; Linn et al., 1991). Externe experts zijn vaak betrokken bij de accreditatie van opleidingen (Dijkstra et al., 2010).

Realiseren van kwaliteit van toetsing

Dit artikel heeft kwaliteit van toetsing in het tertiaire onderwijs met behulp van een litera-



De drie kwaliteitscriteria zijn validiteit, transparantie en betrouwbaarheid

tuurstudie geconceptualiseerd. Deze conceptualisering heeft gevolgen voor het realiseren van de kwaliteit van toetsing in de dagelijkse onderwijspraktijk.

Er blijken, zowel in de literatuur als in de praktijk, diverse interpretaties van het concept kwaliteit van toetsing te bestaan. Dit komt doordat het concept kwaliteit van toetsing zich geleidelijk, door de jaren heen, heeft ontwikkeld door verschillende economische en sociaal-culturele ontwikkelingen (Gerritsen-van Leeuwenkamp, 2019). Daarom zijn er in deze literatuurstudie veel verschillende kwaliteitscriteria gevonden ($n = 98$) en er zijn meerdere definities gevonden voor identieke concepten (bijvoorbeeld voor validiteit) (Gerritsen-van Leeuwenkamp, 2019). Daarnaast blijkt ook dat het concept kwaliteit van toetsing niet stabiel is. Het hangt af van de manier waarop belanghebbenden het begrip definiëren. Zo lijkt bijvoorbeeld de definitie van eerlijkheid

die gehanteerd wordt door studenten op de definitie van validiteit van experts (Sambell et al., 1997). Om de perspectieven van studenten ten aanzien van kwaliteit van toetsing in kaart te brengen kan er gebruik gemaakt worden van een vragenlijst (Gerritsen-van Leeuwenkamp, Joosten-ten Brinke, & Kester, 2018). Op basis van de tekstanalyses zijn alle kwaliteitscriteria geclusterd in drie thema's met kwaliteitscriteria: validiteit, transparantie en betrouwbaarheid. Voor de dagelijkse praktijk betekent dit dat er naast deze drie kwaliteitscriteria geen aanvullende criteria onderscheiden hoeven te worden. Deze drie criteria dienen echter wél gedefinieerd te worden in relatie tot de beoogde functie en het gebruik van toetsing. Daarvoor is het nodig dat de verschillen in conceptualisering en perspectieven tussen groepen belanghebbenden inzichtelijk worden gemaakt. Dit vraagt een expliciete dialoog, discussie en onderhandeling met de vier groepen belanghebbenden



Neem de perspectieven van belanghebbenden mee

(i.e. studenten, medewerkers, overheid en experts) om tot een gemeenschappelijke conceptualisering van kwaliteit van toetsing in de onderwijspraktijk te komen. De conceptualisering van kwaliteit van toetsing gegeven in dit artikel kan de input zijn voor de dialoog. Deze dialoog leidt tot de ontwikkeling van beleid en richtlijnen om kwaliteit van toetsing in een onderwijsorganisatie te optimaliseren. Dit alles om de kwaliteit van toetsing in het tertiaire onderwijs te verbeteren en te garanderen.

Literatuurlijst

- AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Anderson, T., & Rogan, J. M. (2010). Bridging the educational researchteaching practice gap. Tools for evaluating the quality of assessment instruments. *Biochemistry and Molecular Biology Education*, 38, 51–57. doi:10.1002/bmb.20362
- ARG. (2002). *Assessment for learning: 10 principles*. Retrieved from <https://www.aaia.org.uk/content/uploads/2010/06/Assessment-for-Learning-10-principles.pdf>
- Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (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., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (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
- Barman, A. (2011). Feasibility of applying classical test theory in testing reliability of student assessment. *International Medical Journal*, 18, 110–113.
- Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational*

Evaluation, 33, 29–49. doi:10.1016/j.stueduc.2007.01.004

- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071. doi:10.1037/0033-295X.111.4.1061
- Dijkstra, J., Van der Vleuten, C. P. M., & 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
- Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *The Metric of Medical Education*, 38, 1006–1012. doi:10.1111/j.1365-2929.2004.01932.x
- Ebel, R. L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice*, 2(2), 7–10. doi:10.1111/j.1745-3992.1983.tb00688.x
- Gerritsen-van Leeuwenkamp, K. J. (2019). Assessment quality in higher education: Students' perspectives in relation to their learning. (Doctoral thesis). Open Universiteit Nederland, Heerlen.
- Gerritsen-van Leeuwenkamp, K. J., Joosten-ten Brinke, D., & Kester, L. (2017). Assessment quality in tertiary education: An integrative literature review. *Studies in Educational Evaluation*, 55, 94–116. doi:10.1016/j.stueduc.2017.08.001
- Gerritsen-van Leeuwenkamp, K. J., Joosten-ten Brinke, D., & Kester, L. (2018). Developing questionnaires to measure students' expectations and perceptions of assessment quality. *Cogent Education*, 5(1464425). doi:10.1080/2331186X.2018.1464425
- Gulikers, J., Kester, L., Kirschner, P. A., & Bastiaens, T. J. (2008). The effect of practical experience on perceptions of assessment authenticity, study approach, and learning outcomes. *Learning and Instruction*, 18, 172–186. doi:10.1016/j.learninstruc.2007.02.012
- Holmes, L. E., & Smith, L. J. (2003). Student evaluations of faculty grading methods. *Journal of Education for Business*, 78, 318–323. doi:10.1080/08832320309598620
- Joosten-ten Brinke, D. (2011). Eigentijds toetsen en beoordelen. Tilburg, Nederland: Fontys Hogescholen.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x
- Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher*, 37, 76–82. doi:10.3102/0013189X08315390
- Leximancer. (2008). *Leximancer the why, not just the what. Leximancer manual*. Brisbane, Australia: Leximancer Pty Ltd.
- Leximancer Pty Ltd. (2008). Leximancer 3 [Computer software]. Brisbane, Australia: Author.
- Linn, R. L., Bakker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15–21. doi:10.3102/0013189X020008015
- Maclellan, E. (2004). How convincing is alternative assessment for use in higher education? *Assessment & Evaluation in Higher Education*, 29, 311–321. doi:10.1080/0260293042000188267
- Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749. doi:10.1037/0003-066X.50.9.741
- Meyer, L. H., Davidson, S., McKenzie, L., Rees, M., Anderson, H., Fletcher, R., & Johnston, P. M. (2010). An investigation of tertiary assessment policy and practice: Alignment and contradictions. *Higher Education Quarterly*, 64, 331–350. doi:10.1111/j.1468-2273.2010.00459.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
- Sambell, 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, 349–371. doi:10.1016/S0191-491X(97)86215-3
 - Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (2004). Different written assessment methods: What can be said about their strengths and weaknesses? *Medical Education*, 38, 974–979. doi:10.1111/j.1365-2929.2004.01916.x
 - Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405–450. doi:10.3102/0091732X019001405
 - Stobart, G. (2008). *Testing times. The uses and abuses of assessment*. Oxon, England: Routledge.
 - Stowell, M. (2004). Equity, justice and standards: Assessment decision making in higher education. *Assessment & Evaluation in Higher Education*, 29, 495–510. doi:10.1080/02602930310001689055
 - Van de Watering, G., & Van de Rijt, J. (2006). Teachers' and students' perceptions of assessment: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review*, 1, 133–147. doi:10.1016/j.edurev.2006.05.001
 - Van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education*, 1, 41–67. doi:10.1007/BF00596229
 - Van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & Van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, 34, 205–214. doi:10.3109/0142159X.2012.652239

Oproep

Werp eens een blik op de Toetskaart van Nederland

Enkele jaren geleden nam de redactie van EXAMENS het initiatief voor de 'Toetskaart van Nederland', een overzicht van veel op toetsgebied, dat door aanbieders zelf wordt ingevuld. Een blik op de toetskaart laat zien dat ook veel te vinden is. Biedt u iets aan op toetsgebied, plaats het op de toetskaart. Het gebruik is eenvoudig. Nieuwsgierig naar wat de Toetskaart al aan informatie biedt? Zie <http://e-xamens.nl/nl/toetskaart-van-nederland/>