ToetsSpecial

Beoordelen van de kwaliteit van toetsen en examens

Deel 2: De praktijk

Cor Sluijter, Bas Hemker & Theo Eggen (Cito)

ToetsWijzer

Inhoud

Inleiding

In de eerste *ToetsSpecial* over de kwaliteit van toetsen en examens hebben we laten zien welke richtlijnen, standaarden en systemen er zijn om uitspraken te kunnen doen over de kwaliteit van deze instrumenten. In dit tweede deel staan we stil bij de beoordeling van toetsen en examens in de praktijk en maken we duidelijk waarom dat lastig is.

1. Wat maakt het beoordelen van de kwaliteit van toetsen lastig?

Er zijn uiteenlopende factoren die de discussie over toetskwaliteit bemoeilijken. Om te beginnen kunnen toetsen uiteenlopende functies en doelen hebben. Dat heeft tot gevolg dat bij het beoordelen van de kwaliteit van verschillende toetsen ook deels uiteenlopende criteria kunnen meewegen. De mate van belang van verschillende criteria kan hierbij variëren. Verder is er in discussies over toetskwaliteit regelmatig sprake van begripsverwarring en wel op twee manieren. Niet alleen worden verschillende termen gehanteerd voor overeenkomstige begrippen. Maar ook kernbegrippen als betrouwbaarheid en validiteit, worden geregeld op verschillende manieren opgevat, waarover later meer.

Bij de betrouwbaarheid van een toets gaat het om de mate waarin staat te maken valt op de resultaten op het instrument, dat wil zeggen de mate waarin toetsscores consistent, nauwkeurig en reproduceerbaar zijn, kortom vrij van meetfouten. Een toetsscore is betrouwbaarder naarmate deze minder beïnvloed wordt door storende factoren die samenhangen met de toets, de kandidaat of een eventuele beoordelaar. Denk hierbij aan zaken als het tijdstip waarop getoetst wordt, de specifieke vormgeving van het instrument en het hebben van pech of geluk bij het beantwoorden van de opgaven. De mate waarin een toetsscore ongevoelig is voor deze en andere verstorende factoren, kan worden geschat door berekening van een betrouwbaarheidscoëfficiënt. De

waarde daarvan ligt tussen 0 en 1, waarbij geldt hoe hoger de betrouwbaarheidscoëfficiënt, des te vrijer de toetsscore is van meetfouten. Bij validiteit gaat het om de vraag of een toets daadwerkelijk de informatie biedt die past bij de beoogde functie. In eerste instantie ligt hier een taak voor de ontwikkelaar van de toets. Deze moet duidelijk aangeven wat de bedoelde interpretatie van de toetsscores is en wat het exacte gebruiksdoel is. De gebruiker heeft óók een verantwoordelijkheid. De gebruiker moet namelijk op basis van de informatie die de ontwikkelaar levert, bepalen of de toets de informatie oplevert die nodig is. En dus of de toets geschikt is voor het door de gebruiker beoogde doel. Validiteit is geen intrinsieke eigenschap van een toets, maar komt pas tot uiting in het gebruik ervan.

Bij het beoordelen van toetsen met verschillende doelen kunnen dus andere normen gelden. En bovendien zijn relevante kwaliteitscriteria eigenlijk niet goed strikt toe te passen. Tot slot is binnen verschillende onderwijssectoren ook nog eens sprake van een andere toetscultuur. De verschillen in toetscultuur hangen deels samen met de uiteenlopende functies van toetsen in die sectoren. Maar ook met het feit dat er andere vaardigheden gemeten worden, dat de wijze van toetsen deels verschilt en dat de rol van onderwijsgevenden kan variëren bij de toetsing. In de volgende paragrafen gaan we nader in op deze vier punten.

1.1 Uiteenlopende doelen en functies

Een gebruikelijke indeling van soorten toetsen is de volgende (Sanders, 2013):

- Toetsen voor het beoordelen van personen, groepen en scholen
- Toetsen voor het beoordelen van schoolsystemen
- Toetsen voor het ondersteunen van het onderwijsleerproces

Iets strikter gesproken, gaat het om het gebruik van de resultaten op de toetsen. Bij de eerste twee categorieën toetsen is sprake van summatief gebruik van toetsresultaten. Op basis van de resultaten valt te bepalen wat de impact van het voorafgaande onderwijs was op de leerling, dan wel student, de onderwijsinstelling of het onderwijssysteem. Bij de derde minstens zo belangrijke categorie gaat het om formatief gebruik; het inzetten van toetsresultaten voor het ondersteunen van het onderwijsleerproces.

Het gebruik van toetsresultaten voor het beoordelen van personen, groepen en scholen

Toetsen voor het beoordelen van personen worden in het onderwijs ingezet bij beslissingen rond selectie, plaatsing, classificatie en certificering. Voor een verdere uitleg van deze verschillende soorten beslissingen verwijzen wij naar Sanders (2013). Het is gebruikelijk om de resultaten die individuele leerlingen of studenten binnen een bepaalde groep of onderwijsinstelling behaald hebben ook te gebruiken om informatie over de groep of de school als totaal te geven. Er worden daarom zelden of nooit specifiek toetsen ontwikkeld voor het beoordelen van groepen leerlingen of scholen.

Het gebruik van toetsresultaten voor het beoordelen van schoolsystemen

Van geheel andere aard is de toetsing in het kader van het beoordelen en internationaal vergelijken van schoolsystemen. In Nederland vindt bijvoorbeeld al sinds 1987 onderzoek naar het niveau van het basisonderwijs in Nederland plaats om inzicht te krijgen in de leeropbrengsten van het basisonderwijs en het periodieke verloop ervan. Omdat dit onderzoek tot doel heeft in detail uitspraken te doen over het landelijk niveau en niet over individuele leerlingen, heeft de toetsing een heel ander karakter. De onderzoeksopzet is zodanig dat deelnemende leerlingen slechts een beperkt deel voorgelegd krijgen van alle opgaven die ontwikkeld worden om het landelijk niveau van het onderwijs voor een bepaald domein in kaart te brengen.

Het gebruik van toetsresultaten voor het ondersteunen van het onderwijsleerproces

Er zijn drie verschillende hoofdsoorten van formatief toetsgebruik . Allereerst kan er sprak zijn van diagnostische gebruik van toetsresultaten. Die resultaten moeten dan zo gedetailleerd zijn dat het mogelijk is om stappen in de ontwikkeling of misvattingen bij het leren te kunnen identificeren. Ook kunnen

toetsresultaten gebruikt worden om leerkrachten en leerlingen te ondersteunen bij het inrichten van leerprocessen. Toetsen binnen leerlingvolgsystemen en voortgangstoetsen zijn hier voorbeelden van. En tot slot zijn er toetsen die alleen maar bedoeld zijn om van te leren en waarbij de geleverde prestatie weinig van belang is. Dergelijke toetsen maken bijvoorbeeld onderdeel uit van computergestuurde oefen- en leersystemen (Van der Kleij et al., 2015).

Het feit dat toetsen uiteenlopende doelen en functies hebben, leidt ertoe dat er andere accenten gelegd worden op verschillende kwaliteitscriteria. Zo draait het bij summatief gebruik van toetsresultaten vooral om het voldoende nauwkeurig meten op het niveau van individuele leerlingen, maar gaat het bij formatief gebruik van toetsresultaten eerder om de vraag of de toets in positieve zin bijdraagt aan het leerproces. Bij toetsing op systeemniveau is het niet van belang om de prestaties van individuele leerlingen nauwkeurig te meten, maar om voldoende gedetailleerd informatie te krijgen over de landelijke mate van beheersing van een concreet omschreven domein. We zien dus dat het binnen de discussie over kwaliteit ook van groot belang is te weten welk doel een specifieke toets heeft.

1.2 Begripsverwarring

Hierboven zijn we al ingegaan op de twee belangrijkste begrippen die verband houden met toetskwaliteit: betrouwbaarheid en validiteit. Voor het begrip betrouwbaarheid bestaat een reeks verschillende maten die dit construct allemaal op een andere manier schatten. Met betrouwbaarheid kan verwezen worden naar (Evers, Lucassen, Meijer, & Sijtsma, 2010, p.32-35):

- Paralleltestbetrouwbaarheid
- Betrouwbaarheid op basis van inter-itemrelaties
- Test-hertestbetrouwbaarheid
- Interbeoordelaarsbetrouwbaarheid
- Betrouwbaarheid gebaseerd op methoden uit de itemresponstheorie
- Betrouwbaarheid gebaseerd op methoden uit de generaliseerbaarheidstheorie of structurele vergelijkingsmodellen.

Welke van deze maten relevant zijn, is afhankelijk van wat een toets meet en de specifieke toepassing ervan.

Een concreet en eenduidig antwoord op de vraag wat nu exact validiteit is, is eveneens niet goed te geven. Op dit vlak woedt eigenlijk nog steeds een discussie tussen verschillende kampen (zie voor meer informatie hierover Newton & Shaw, 2014). De oorspronkelijke definitie van validiteit was “de mate waarin een test meet wat men ermee beoogt te meten” (Garret, 1937, p.324, geciteerd in Angoff, 1988). Tegenwoordig is de definitie genuanceerder en ziet men validiteit als “de mate waarin scores op een toets te interpreteren zijn en een toets te gebruiken is zoals bedoeld” (Kane, 2013). De bewijsvoering is ook veel genuanceerder dan oorspronkelijk. Tot omstreeks 1950 was het gebruikelijk de samenhang tussen de score op een toets en een of ander criterium te rapporteren als indicatie voor de validiteit ervan (Angoff, 1988; Shepard, 1993). Maar deze simpele praktische invulling werd verdrongen door een meer theoretisch georiënteerde zienswijze, waarin verschillende soorten validiteit een plaats innamen (Messick, 1989). Uiteindelijk kwam men tot de conclusie dat de gegevens die men verzamelde om de validiteit van een toets te bepalen alle bijdragen aan een juiste interpretatie en een correct gebruik van de scores op deze test. De huidige opvatting is dat sprake moet zijn van een argumentgerichte benadering: de beoogde interpretatie en het beoogde gebruik van toetsscores bepalen welke uiteenlopende criteria in welke mate relevant zijn (Wools, 2015).

Het bovenstaande laat zien dat dé betrouwbaarheid van een toets dus niet bestaat. Welke maat gebruikt kan of moet worden, hangt van een aantal factoren af. Iets soortgelijks geldt voor validiteit: welke bewijzen men aandraagt is afhankelijk van de beoogde interpretatie van de toetsscores en het beoogde gebruiksdoel van de toets. Kortom: onder de termen

betrouwbaarheid en validiteit gaat een complexe wereld schuil die een eenvoudig te voeren discussie over de kwaliteit van toetsen bemoeilijkt.

1.3 Wanneer is toetskwaliteit voldoende?

Toetskwaliteit heeft te maken met de mate waarin zo'n instrument geschikt is voor het ermee beoogde doel. Dat brengt natuurlijk een probleem met zich mee, want wanneer is een toets dan voldoende geschikt voor een bepaald doel? Dit probleem valt goed te illustreren aan de hand van de vraag wanneer een toets voldoende betrouwbaar is. Het is een gegeven dat een toets een feilbaar instrument is. Het is een middel om onder min of meer gecontroleerde omstandigheden informatie te verzamelen over de kennis, vaardigheden, competenties of prestaties van een persoon. De informatie is echter wel beperkt waardoor het generaliseren altijd gepaard gaat met meetfouten. Maar de vraag hoe groot een meetfout mag zijn, is niet in absolute zin te beantwoorden.

Het mechanistisch toepassen van een beoordelingssysteem lost dit probleem natuurlijk op. De COTAN stelt bijvoorbeeld dat voor toetsen die gebruikt worden bij belangrijke beslissingen op individueel niveau, zoals personeelsselectie of verwijzingen naar speciaal onderwijs, een betrouwbaarheidscoëfficiënt lager dan 0,80 onvoldoende is. Maar welbeschouwd leidt dit alleen maar tot een verplaatsing van het probleem, want wanneer is sprake van een 'belangrijke beslissing'?

1.4 Toetsculturen

De verschillende sectoren van het Nederlands onderwijs onderscheiden zich op allerlei manieren van elkaar. Niet alleen wat de leeftijd van de leerlingen en studenten betreft, maar ook in de onderwerpen, kennis en vaardigheden die onderwezen worden. Illustratief in dit kader is dat de reeks Toetsen op School die Cito uitbrengt en onderhoudt, naast een algemeen deel, aparte delen heeft voor het primair en het voortgezet onderwijs, het middelbaar beroepsonderwijs en het hoger onderwijs. De genoemde verschillen brengen met zich mee dat ook sprake is van andere, van elkaar verschillende, toetsculturen waarin aandacht is voor uiteenlopende aspecten van toetsing en examinering.

2. Conclusies

We hebben in het voorafgaande duidelijk gemaakt waarom de discussie over toetskwaliteit in de praktijk gecompliceerd kan verlopen. De oorzaken daarvan zijn deels weg te nemen. Als toetsdoelen en functies voldoende concreet beschreven zijn, kunnen daar geen misverstanden over bestaan.

De begripsverwarring die soms ontstaat valt op te lossen door bij discussies gebruik te maken van een welomschreven reeks van eisen zoals de eisen van het RCEC die we in de eerste *ToetsSpecial* van deze reeks van twee hebben beschreven. Toetscultuur kan ook geen rol meer spelen als doel en functie en kwaliteitscriteria concreet zijn omschreven.

Waar werkelijk over te twisten blijft, is de vraag: hoe goed is goed genoeg? Een grote stap in de goede richting zou te zetten zijn, wanneer toets- en examenontwikkelaars voldoende transparant zijn bij het schrijven van hun verantwoording bij het instrument. Deze verantwoording kan dan de basis vormen voor een gerichte discussie over de kwaliteit van het betreffende instrument op basis van door professionals ontwikkelde beoordelingssystemen die door alle belanghebbenden geaccepteerd worden. Gelet op de belangrijke rol van de instrumenten, zou er bij álle toetsen en examens sprake moeten zijn van een volledige onderbouwing van álle relevante kwaliteitsaspecten.

Bijlage: Literatuur

Angoff, W.H. (1988). Validity: an evolving concept. In: H. Wainer, & H.I. Braun (Eds.). *Test validity* (pp. 19-32). Hillsdale: Lawrence Erlbaum.

Haertel, E.H. (2006). Reliability. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education and Praeger Publishers.

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

Messick, S. (1989). Validity. In: R.L. Linn (Ed.). *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment*. London: Sage.

Sanders, P.F. (2013) Het doel van toetsen. In: P.F. Sanders (Red.), *Toetsen op School. Herziene versie*. Arnhem: Cito. [http://www.cito.nl/Onderzoek%20en%20wetenschap/achtergrondinformatie/toetsen_op_school.aspx/]

Shepard, L.A. (1993). Evaluating test validity. In: L. Darling-Hammond (Ed.). *Review of research in education: Vol. 19* (pp.405-450). Washington, DC: American Educational Research Association.

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

Wools, S. (2012). Towards a Comprehensive Evaluation System for the Quality of Tests and Assessments. In: T.J.H.M. Eggen & B.P. Veldkamp, (Eds*.),* *Psychometrics in Practice at RCEC* (pp.95-106). Enschede: RCEC. [http://www.rcec.nl/e-books/psychometrics-in-practice.pdf]

Wools, S. (2015). All about Validity. An evaluation system for the quality of educational assessment. Proefschrift. Enschede: Universiteit Twente. [https://www.researchgate.net/publication/281434844_All_About_Validity_%20An_evaluation_system_for_the_quality_of_educational_assessment]

ToetsSpecial

Over deze ToetsSpecial | Resultaten op toetsen en examens beïnvloeden in sterke mate het verloop van onderwijsloopbanen en in de nodige gevallen ook carrières daarna. Toetsen en examens moeten daarom van onbesproken kwaliteit zijn. Het bepalen van de kwaliteit van toetsen en examens is echter lastig. *Deze special is de tweede in een reeks die gaat over het beoordelen van de kwaliteit van toetsen en examens.*

Dit tweede deel maakt duidelijk waarom het niet eenvoudig is absolute uitspraken te doen over de kwaliteit van deze instrumenten. De notitie is bedoeld voor toetsdeskundigen en onderwijsgevenden in alle sectoren, maar kan ook inzichtelijk zijn voor iedereen die zich een eigen oordeel wil vormen over de mate waarin de huidige kritiek op de kwaliteit van toetsen en examens terecht is.

Colofon

Uitgave: ToetsWijzer | Stichting Cito Instituut voor Toetsontwikkeling
Auteur(s): Cor Sluijter, Bas Hemker, Theo Eggen (Cito)
Datum: mei 2018
e info@toetswijzer.nl | **i** www.toetswijzer.nl

ToetsWijzer