

Een model voor programmatische toetsing

□ Cees van der Vleuten en Lambert Schuwirth

Dit artikel gaat over een onderwijscontext waarin toetsing ondersteunend is aan het onderwijs. In dit onderwijs heeft zelfsturing een belangrijke rol en de onderwijskundige ambities gaan het klassieke beheersingsleren te boven. Een toetsprogramma in deze context moet het leren maximaal bevorderen en tegelijkertijd informatie opleveren om robuuste beslissingen over studievoortgang te kunnen nemen. Aan de Universiteit Maastricht is een model voor de inrichting van een dergelijk toetsprogramma ontworpen. De auteurs hopen dat ervaringen in het hbo en wo met deze wijze van toetsing worden opgeschreven en gedeeld, zodat ze kunnen leiden tot vernieuwingen in het model.

Van toets naar toetsprogramma

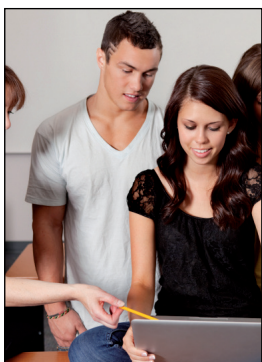
Een toets moet voldoen aan allerlei kwaliteiten. Betrouwbaarheid en validiteit zijn de bekendste. Invloed op leren, acceptabiliteit en kosten van toetsing zijn daaraan toe te voegen (van der Vleuten, 1996), maar er zijn meer kwaliteiten te onderscheiden. Van belang is dat elke toetssituatie een compromis zal zijn tussen deze kwaliteiten. Er bestaat geen oneindig budget, een perfecte betrouwbaarheid valt niet te bereiken in een gegeven toetsvorm, et cetera. Het concrete compromis zal erg afhangen van de specifieke context. In een certificerende toets waarmee een professio-

nele bevoegdheid wordt verkregen, zal een ander compromis worden gesloten dan in een tussentijdse toets op een bepaald moment in een onderwijsstraject. Compromissen worden ook anders gesloten als toetsen een plaats en een functie hebben binnen een geheel van een onderwijsprogramma (van der Vleuten & Schuwirth, 2005). Niet de enkele toets moet maximaal voldoen aan alle toetskwaliteiten, maar het toetsprogramma als geheel. Over een goed toetsprogramma is nagedacht als een arrangement en er wordt regie over gevoerd: de som is meer dan het geheel der delen. Net zoals een goed onderwijsprogramma met verstand in elkaar wordt gestoken, zo is het verstandig om een toetsprogramma te ontwikkelen vanuit een plan. Vergelijkbaar aan een onderwijsprogramma wordt er voortdurend geëvalueerd en bijgesteld in het toetsprogramma. Vergelijkbaar aan een onderwijsprogramma is het na een aantal jaren verouderd of versleten. Een complete revisie zal dan nodig zijn.

Het is interessant om te zien dat in de literatuur bijna uitsluitend aandacht wordt besteed aan de kwaliteitskenmerken van afzonderlijke toetsinstrumenten, terwijl er weinig tot geen aandacht bestaat voor programma's van toetsen. In het proefschrift van Baartman zijn de eerste verkenningen gedaan met betrekking tot de kwaliteiten van toetsprogramma's (Baartman, Bastiaens & Kirschner, 2005). Meer recent is dat gebeurd voor het ontwerpen van toetsprogramma's (Dijkstra, van der Vleuten & Schuwirth, 2009), om uiteindelijk te komen tot een set richtlijnen voor ontwerp. In dit artikel wordt een model beschreven dat op een generieke manier aangeeft hoe een toetsprogramma idealiter in elkaar zou moeten zitten om het leren te bevorderen en tegelijkertijd rijke informatie op te leveren. Cumulatie van die informatie levert vervolgens een solide basis voor het nemen van selectie- of bevorderingsbeslissingen.

Uitgangspunten

Er zijn twee belangrijke uitgangspunten voor een



Het model is bewust generiek gehouden

programmatische benadering van toetsen:

1° naarmate een beslissing belangrijkere gevolgen heeft, moet ze op basis van meer informatie genomen worden;

2° een toets moet maar zeer beperkt een compromis sluiten op informatie- of leerwaarde van de toetsing. Eén enkele toets is in dit denken slechts één datapunt. Een enkel datapunt zal altijd zijn beperkingen hebben, flinke compromissen moeten worden gesloten ten aanzien van de kwaliteitskenmerken. Taakvariabiliteit vormt een van de grootste bronnen van ruis in vrijwel elke toets. Brede steekproeftrekking uit het taakdomein en lange toetsen zijn nodig voor het bereiken van minimale betrouwbaarheid, veel langer dan wat haalbaar is in de meeste opleidingen.

In veel opleidingen is het volstrekt normaal dat toetsvragen geheim worden gehouden na toetsafname. Dat voorkomt dat telkens nieuwe vragen moeten worden gemaakt. Er wordt blijkbaar een gigantisch compromis gesloten op de leerwaarde van de toets: de enige feedback van zo'n toets is of je geslaagd bent of niet, misschien gekoppeld aan een cijfer (een van de zwakste vormen van feedback).

De discussie over summatief/formatief toetsen wordt bij programmatische toetsing vervangen door te spreken over wat voor gevolg er op het spel staat (*stakes* in het Engels). Wat er op het spel staat is een continuüm, geen dichotoom gegeven zoals in de formatief/summatief discussie. Bij een enkel datapunt staat er weinig op het spel. Dramatische beslissingen nemen op basis van een enkel datapunt is onverstandig. Een enkel datapunt staat volledig in het teken van het stimuleren van goed leergedrag. Dat betekent in essentie dat de toets gericht is op het geven van feedback en gewenste leergedragingen ontlokt. In de praktijk beïnvloeden veel toetsen het leren negatief en ontlokken ze maximalisatie-strategieën die ongewenste effecten hebben (memorisatie, minimale voorbereiding, uitstelgedrag, triviale beheersing van vaardigheden, om er maar een paar te noemen). Staat er veel op het spel, dan is het logisch dat veel datapunten nodig zijn. Voor belangrijke beslissingen gaat men niet over één nacht ijs.

Model voor programmatisch toetsen

In het model worden leeractiviteiten, toetsactiviteiten en begeleidingsactiviteiten onderscheiden. Aan het einde van de onderwijsperiode vindt een tussentijdse beoordeling plaats op grond van alle voorliggende geaggregeerde toetsinformatie tegen een achterliggende beoordelingsstandaard. De onderwijsperiodes en tussentijdse beoordelingen herhalen zich. Het aantal zal afhankelijk zijn van de curriculumindeling en de gemaakte keuzes in bekostiging. Ook hier geldt het adagium: liever minder frequent en goed dan het omgekeerde. Er hoeft niet aan elke onderwijsperiode een tussenbeoordeling verbonden te zijn, maar tenminste één tussenbeoordeling voor de finale beoordeling is nodig om het ontwikkelingsproces zijn gang te kunnen laten gaan. Ook is het systeem van finale beoordeling geloofwaardiger als het tot zo min mogelijk verrassingen leidt voor de lerende. Tussentijdse beoordelingen zijn daarvoor noodzakelijk.

Leeractiviteiten

Een onderwijsperiode bestaat uit allerlei leertaken. Er wordt van uitgegaan dat deze leertaken op een onderwijskundig verantwoorde manier in een arrangement zijn geplaatst, bijvoorbeeld gebaseerd op ontwerpregels voor goed onderwijs (Janssen & van Merriënboer, 2002). Sommige leertaken kunnen een product opleveren (bijvoorbeeld een project, een presentatie, een verkoopproduct, een lijst van gedane taken). In diezelfde periode vinden toetsactiviteiten plaats. Betreft het onderwijs een praktijkleerperiode, dan zijn de praktijktaken gelijk aan de leertaken.

Toetsactiviteiten

Een toetsactiviteit varieert van cognitief (feitgericht, toepassingsgericht) tot gedragsmatig (in vitro, in vivo) (Vleuten, Driessen & Tartwijk, 2006). Het areaal aan toetsmethoden is de laatste jaren sterk uitgebreid. Daarbij geldt dat geen enkele methode perfect is of een wondermiddel is. Het succes van veel methoden (m.n. de observatiegerichte) wordt vaak zelfs meer bepaald door de gebruikers van de methode (m.n. hoe serieus ze invulling geven aan het gebruik van de methode), dan van de methode zelf. Kenmerkend voor de toetsactiviteiten in het model is dat ze maximaal aansluiten op het onderwijs in vorm en

Het model gaat uit van verbondenheid tussen onderwijs en toetsing



frequentie, en maximaal betekenis hebben voor het leren. Dat betekent dat niet wordt volstaan met een beoordeling alleen, maar dat voor het leren nuttige feedback gegeven wordt. Dat kan in kwantitatieve vorm of in narratieve vorm en volgens de spelregels van feedback. Kwalitatieve, beschrijvende informatie is zeker voor complexe en gedragmatige competenties van onschatbare waarde. Cijfers vormen een schrale vorm van feedback die leidt tot een cijfercultuur. Het is soms ontmoedigend te zien hoeveel toetsprogramma's in de praktijk van het onderwijs grotendeels of geheel feedbackloos zijn. Toetsing moet volledig ingebed zijn in het onderwijs en rijke informatie opleveren die gerichte betekenis heeft voor het leren. Uiteraard zullen compromissen moeten worden gesloten. Zo zullen de kosten binnen de marges moeten blijven, maar binnen het haalbare wordt de leerwaarde van de toetsing gemaximeerd volgens het principe 'less is more': liever minder frequent en goed dan vaak en ritualistisch.

Sommige leertaken zullen tegelijkertijd toetstaken kunnen vormen waardoor er de minste kans op incongruentie tussen onderwijs- en toetsdoelen (*constructive alignment*) bestaat. Voor elk individueel datapunt van toetsing staat weinig, maar niet niks op het spel. De informatie uit het datapunt kan gebruikt worden als de informatie wordt geaggregeerd. Voor sommige toetstaken staat iets meer op het spel. Sommige taken moeten worden beheerst. De tafeltjes van vermenigvuldiging vormen een typische beheersingstaak. Een student geneeskunde moet bijvoorbeeld kunnen reanimeren. Er wordt net zo lang geoefend tot de taak beheerst wordt. In de toetstaak wordt de bekwaamheidsverklaring voor de beheersingstaak afgegeven.

Begeleidingsactiviteiten

De informatie van de leer- en toetstaken wordt door de lerende gebruikt om het leren te sturen. Reflectie en planning en zelfsturing vinden plaats. Nu is bekend dat zelfsturing geen eenvoudig proces is en niet zo maar gebeurt, en dat sociale interactie rondom reflectie van groot belang is. Eén van de paradoxen in het onderwijs is dat veel sturing moet worden geboden om zelfsturend leren plaats te laten vinden, des

te meer naarmate de lerende jonger is en het brein onvolgroeid. In het model is daarom een vorm van sociale interactie rondom zelfsturing opgenomen. Deze zou kunnen plaatsvinden door middel van coaching, supervisie of intervisie. De ontwikkeling wordt gevolgd, bijsturing vindt plaats. Wellicht dat in de loop der tijd de begeleidingsactiviteit kan afnemen of, afhankelijk van de lerende, meer of minder gedoseerd kan plaatsvinden.

Tussentijdse beoordeling

Aggregatie van toetsinformatie zal geen eenvoudige somming of andere kwantitatieve operatie zijn. Het gaat om rijke informatie, deels kwalitatief, waarbij een kwantitatieve aggregatie snel onmogelijk is. De beoordeling zal door een 'expert' moeten gebeuren, liefst een commissie, die onafhankelijk is van de betrokkenen in de begeleidingsactiviteiten. Het besluit dat genomen wordt is diagnostiserend, remediërend en prospectief. Er zal eerst worden vastgesteld of er voldoende informatie is, vervolgens wat er aan de hand is, hoe eventueel dit geremedieerd moet worden, en er wordt voorspeld wat de uitkomst zal zijn indien dat (niet) gebeurt. Dit is een volkomen andere benadering dan in een klassiek zak/slaag besluit gekoppeld aan een herkansing. Herhaling is typisch iets dat past in het klassieke beheersingsgerichte model. Op grond van gelimiteerde informatie, namelijk zakken of slagen, wordt één vorm van remediëring geboden, namelijk herhaling. Dat is tamelijk informatie-arm, weinig efficiënt en niet ontwikkelingsgericht. Het hier beschreven model is gericht op remediëring, op longitudinale ontwikkeling en opvolging. Herhaling wordt in dit model niet uitgesloten, maar het is één van de opties in een geïndividualiseerd remediëringadvies. De echte beheersingstaken vormen opnieuw een uitzondering. Herhaling en herkansing vormen hier een adequaat remediëringadvies.

Voor de meeste lerenden zal een uitgebreid tussentijds remediëringadvies niet noodzakelijk zijn. Hierin zit in een efficiëntieslag van het werk dat gemoeid is met deze vorm van beoordelen en remediëren. Een andere efficiëntieslag zit in de mate waarin het oordeel mede gevormd wordt door degene die het beste bekend is met het werk van de lerende, de begeleider, coach of mentor. Dit typeren wij als een



Veel toetsen beïnvloeden het leren negatief

‘firewall-dilemma’: de rol van de begeleider komt in een ander daglicht te staan als deze ook een beoordelaar wordt, maar tegelijkertijd is de informatie van de begeleider uniek en kan veel tijd besparen. Tussentijdse oplossingen zijn overigens ook mogelijk, bijvoorbeeld een situatie waarin zowel de lerende als de begeleider een adviserende rol hebben, maar de commissie het uiteindelijke besluit neemt. Beoordeeld wordt aan de hand van een standaard die beschrijft wat iemand moeten kunnen/kennen op een aantal inhoudelijke domeinen voor een bepaalde fase van de opleiding. De resulterende beoordeling is belangrijk, er staat veel meer op het spel, maar nog steeds staat niet alles op het spel. Dat komt later in de finale beoordeling.

Finale beoordeling

Bij de finale beoordeling staat alles op het spel. Hier wordt een selectie-, doorstroom-, of certificerende beslissing genomen waar grote gevolgen aan vast kunnen zitten. In het model vindt die beslissing plaats op basis van veel informatie en op veel datapunten, de beslissing moet geloofwaardig zijn. Bij toetsing is men gewend te denken in objectivering en standaardisering, maar in dit model kan niet alles worden geobjectiveerd, kan objectivering zelfs tot trivialisering leiden en staat het menselijk oordeel noodzakelijkerwijs centraal. Andere maatregelen zijn nodig om bias in de besluitvorming te voorkomen. De oplossing wordt gevonden in procedurele maatregelen (geïnspireerd op kwalitatieve onderzoeksmethodieken (van der Vleuten, Schuwirth, Scheele, Driessen & Hodges, 2010) die de besluitvorming robuust maken. De term geloofwaardigheid van besluitvorming is bewust gekozen. Verdedigbare antwoorden moeten beschikbaar zijn als een buitenstaander vervelende vragen zou stellen over het besluit (een rechter?). Enkele voorbeelden die de geloofwaardigheid van de besluitvorming zullen verhogen:

- De grootte en de expertise van de beoordelingscommissie
- De beschikbaarheid en wijze van totstandkoming van de beoordelingsstandaarden
- De mate van training van beoordelaars
- De mate waarin van moeilijke gevallen wordt geleerd (bijvoorbeeld in bijstelling van standaar-

den)

- De mate van deliberatie in de commissie bij twijfelgevallen
- De mate van beargumentering van het besluit
- De kans tot bijsturing en remediering gedurende de rit
- De mogelijkheid tot beroep tegen het besluit
- De mate waarin één en ander gedocumenteerd is

Al deze maatregelen maken het besluit geloofwaardiger, maar laten het noodzakelijke menselijk oordeel intact en het besluit is gebaseerd op grond van veel informatie.

Het model is haalbaar

Mocht gedacht worden dat wat hier voorgesteld wordt, in de praktijk niet haalbaar zou zijn, dan wordt gewezen op enkele bestaande praktijken. Los van een van de eigen opleidingen in Maastricht die op dergelijke wijze is ingericht, bestaat er in Cleveland een opleiding geneeskunde die al een aantal jaren op deze wijze werkt (Dannefer & Henson, 2007). Het gehele toetsprogramma is feedbackgericht. Elke week worden open vragen gemaakt door de studenten waarop feedback wordt gegeven (punten worden wel gegeven, maar die gaan naar de docenten!). Periodiek is er een uitgebreide zelf- en peerevaluatie aan de hand van standaarden. Er is een systeem van begeleiding en een commissie van beoordeling met een absolute firewall tussen beide. De studenten zijn zeer tevreden met deze aanpak en zijn nu de grootste ambassadeurs ervan.

Discussie

Sommigen zullen een systeem van portfoliolenen en beoordelen herkennen. Een portfolio zou voor het bovenstaande programma van toetsen een nuttig instrument kunnen zijn. Het model is bewust generiek gehouden en spreekt geen enkele voorkeur uit voor welk concreet instrument dan ook. In programma-

De vormende en de besluitvormende functie van toetsing worden geoptimaliseerd



tisch denken is de waarde van een instrument erg afhankelijk van de functie in het systeem als geheel en geen enkel instrument wordt daarbij op voorhand in- of uitgesloten (en al beslist niet de klassieke vormen van toetsen). Het model is algemeen bedoeld. Het is bedoeld als een vorm van denken over toetsing, waarbij de vormende en de besluitvormende functie van toetsing worden geoptimaliseerd. Door programmatisch te denken kunnen beide functies hand in hand gaan.

Risico's

Natuurlijk zijn er risico's aan het model verbonden. Geen onbelangrijk risico om een dergelijk model te realiseren zijn de kosten en de haalbaarheid. Het model gaat uit van verbondenheid tussen onderwijs en toetsing, de grens daartussen vervaagt. Dus gaat het om integrale kosten van onderwijs. Compromissen over wat wel en wat niet te doen is dan geen keuze tussen het één (onderwijs) of het ander (toetsing). Conform het 'less is more' principe is 'minder maar goed' beter dan 'veel maar slecht'. Als beoordelingsinstrumenten een doel op zichzelf gaan worden, als lijstjes moeten worden afgewerkt, als puntjes moeten worden verzameld, alles spelletjes gespeeld moeten gaan worden, dan trivialiseert het leerproces (de wet van Goodhart!). Om dat te voorkomen moeten de actoren in dit hele proces (lerenden, docenten) goed beseffen wat hun rol is en waarom ze deze vervullen. Dat is niet makkelijk te realiseren. Voorbeeld: de taak van een docent betrokken bij één datapunt van toetsing is vooral het geven van feedback. Misschien moet het daartoe wel beperkt blijven en moet er geen zak/slaag beslissing of cijfer worden gegeven. Docenten vinden dat lastig, zij ervaren zich in hun 'bevoegdheden' beknot. Maar het zou bevrijdend kunnen zijn, de rol van rechter hoeft niet te worden gespeeld en gecombineerd met de rol als hulpgever, als facilitator, als docent.

Tenslotte kan het huidige juridische kader een bedreiging vormen. De wet en het hele stelsel van studiepunten zijn sterk gebaseerd op het klassieke beheersingsleren-gedachtengoed. Toets gehaald, punten binnen.

Literatuur

- Baartman, L., Bastiaens, T. & Kirschner, P. (2005). Kwaliteitscriteria voor competentie assessment programma's. *Examens*, 2(2), 13-15.
- Dannefer, E.F. & Henson, L.C. (2007). The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. *Academic Medicine*, 82(5), 493-502.
- Dijkstra, J., van der Vleuten, C.P. & Schuwirth, L.W. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education: Theory and Practice*, 15(3), 379-93. Epub 2009 Oct 10.
- Janssen, A.M.B. & van Merrienboer, J.J.G. (2002). *Innovatief Onderwijs Ontwerpen: Van leertaken naar complexe vaardigheden*. Groningen/Houten: Wolters-Noordhoff.
- van der Vleuten, C.P., Schuwirth, L.W., Scheele, F., Driessen, E.W. & Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 24, 703-719.
- van der Vleuten, C.P.M. (1996). The Assessment of Professional Competence: Developments, Research and Practical Implications. *Advances in Health Sciences Education*, 1(1), 41-67.
- van der Vleuten, C.P.M. & Schuwirth, L.W.T. (2005). Assessment of professional competence: from methods to programmes. *Medical Education*, 39, 309-317.
- van der Vleuten, C., Driessen, E., & Tartwijk, J. van. (2006). Toetsprogramma's. In H. van Hout, G. ten Dam, M. Mirande, C. Terlouw & J. Willems, *Vernieuwing in het hoger onderwijs* (pp. 119-136). Assen: Koninklijke Van Gorcum.

□ De heer prof. dr. C.P.M. van der Vleuten is als hoogleraar verbonden aan de vakgroep Onderwijsontwikkeling en Onderwijsresearch van de Universiteit Maastricht, en Wetenschappelijk directeur van de UM Graduate School of Health Professions Education; de heer prof. dr. L.W. Schuwirth is als hoogleraar medisch onderwijs aan de bovengenoemde vakgroep verbonden. E-mail: c.vandervleuten@maastrichtuniversity.nl.

¹Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes" (Goodhart 1975). De wet van Goodhart kan op vele verschillende wijze geformuleerd worden, maar komt er kort gezegd op neer dat zodra een sociale of economische parameter tot doel gemaakt is om sociale of economische politieke beslissingen te nemen hij de informatieve inhoud verliest die noodzakelijk is om deze doelstelling te kunnen realiseren. De wet is genoemd naar Charles Goodhart, een voormalige adviseur van de Bank of England en Emeritus Professor aan de London School of Economics.