Universiteit
Leiden
The Netherlands

Assessment for growth: fostering student learning through assessment innovations in medical education

Wijk, E.V. van

Citation
Wijk, E. V. van. (2025, November 19). *Assessment for growth: fostering student learning through assessment innovations in medical education*. Retrieved from https://hdl.handle.net/1887/4283162

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4283162

Note: To cite this publication please use the final published version (if applicable).

Appendices

Nederlandse samenvatting
List of scientific contributions
Curriculum Vitae
Dankwoord

Nederlandse samenvatting

Toetsing speelt een cruciale rol in het medisch onderwijs; het dient niet alleen om kennis te meten, maar het kan ook leergedrag van studenten sturen. Schriftelijke tentamens worden veelvuldig toegepast binnen medische curricula om op een betrouwbare manier kennisniveaus te beoordelen. Deze toetsen bestaan voornamelijk uit meerkeuzevragen en worden vooral gebruikt om studentprestaties te meten aan de hand van cijfers. Er is echter een verschuiving gaande in de perceptie van toetsing: de focus verschuift van toetsing als beoordelingsinstrument (summatieve toetsen) naar toetsing als leermiddel (formatieve toetsen). De huidige technologische ontwikkelingen bieden nieuwe mogelijkheden om bestaande toetsmethoden te verbeteren en nieuwe innovatieve methoden te ontwikkelen die het leren stimuleren. De implementatie van deze innovatieve toetsmethoden in het medisch onderwijs en de complexe relatie tussen toetsing en leren zijn echter nog onvoldoende onderzocht.

Dit proefschrift richt zich op de mogelijkheden en uitdagingen van verschillende innovaties binnen toetsing met als overkoepelend doel het optimaliseren van toetsing en leren van medisch studenten, zodat zij zich tot competente dokters kunnen ontwikkelen. Allereerst worden de mogelijkheden van een nieuw vraagformat, de *very short answer question* (VSAQ), onderzocht. Vervolgens bestuderen we de implementatie en voordelen van de computer adaptieve voortgangstoets, een toets die vier keer per jaar plaatsvindt en studenten inzicht geeft in hun kennisontwikkeling gedurende de opleiding. Tot slot onderzoeken we de verschillen in feedbackgebruik van studenten tussen formatieve en summatieve toetsen en de factoren die dit feedbackgedrag beïnvloeden. We hebben de effectiviteit van de verschillende innovaties onderzocht aan de hand van vastgestelde criteria voor kwalitatief goede toetsen: aanvaardbaarheid, authenticiteit, cueing effect, educatief effect, haalbaarheid, katalyserend effect, testeffect en validiteit.

Deel I: Very short answer question

Meerkeuzevragen worden in het medisch onderwijs veelvuldig gebruikt vanwege hun betrouwbaarheid en de efficiëntie waarmee ze nagekeken kunnen worden. Echter, ze hebben ook beperkingen, omdat studenten bij het leren vooral op herkenning vertrouwen in plaats van op begrip van de leerstof. Daarnaast kunnen studenten gebruik maken van aanwijzingen (*cues*) in de antwoordopties om bij het juiste antwoord te komen of kunnen ze gokken. VSAQ's kunnen deze nadelen omzeilen. Dit vraagformat vraagt om zeer korte, open antwoorden van slechts één tot vier woorden, wat studenten stimuleert om actief kennis op te halen en zelf een antwoord te generen op basis van hun kennis. Het voordeel van VSAQ's ten opzichte van uitgebreide open vragen is dat ze makkelijker toepasbaar en sneller objectief na te kijken zijn, wat met name van belang is in grote groepen studenten. Ondanks deze voordelen van VSAQ's is er nog weinig onderzoek gedaan naar de betrouwbaarheid en toepasbaarheid van VSAQ's in het medisch onderwijs, vooral in contexten waarin studenten en docenten nog weinig ervaring hebben met dit vraagformat.

In **hoofdstuk 2** beschrijven we een studie waarin we eerdere bevindingen over de betrouwbaarheid (de mate waarin de vraag steeds opnieuw hetzelfde, nauwkeurige resultaat oplevert), het onderscheidend vermogen (de mate waarin de vraag een onderscheid maakt tussen studenten met veel en weinig kennis over de leerstof) en de acceptatie van VSAQ's hebben onderzocht bij geneeskundestudenten en docenten die niet eerder met dit vraagformat werkten.

In deze studie werden studenten verdeeld in twee groepen voor een formatieve toets: de ene groep kreeg eerst VSAQ's en daarna meerkeuzevragen, de andere groep andersom. De resultaten tonen aan dat VSAQ's een hogere betrouwbaarheid en beter onderscheidend vermogen hebben dan meerkeuzevragen. Bovendien was de tijd die nodig was om de VSAQ's na te kijken acceptabel, namelijk gemiddeld twee minuten per vraag voor alle 375 studenten van een geheel cohort. Daarnaast kwam er ook minder '*cueing*' (i.e., het beantwoorden van de vraag door gebruik te maken van cues in de vraag of antwoordopties) voor in VSAQ's dan in meerkeuzevragen. Uit de vragenlijsten bleek ook dat studenten bij VSAQ's minder zeker waren van hun antwoorden. Daarnaast bereidde ongeveer de helft van de studenten zich anders voor op dit format en vond dat VSAQ's een betere afspiegeling van de klinische praktijk gaven. Hiermee bevestigen we dat VSAQ's ook buiten een ervaren onderzoeksgroep goed toepasbaar, betrouwbaar en authentiek zijn.

In **hoofdstuk 3** onderzochten we het onderscheidend vermogen van beide vraagformats verder, waarbij we ons hebben gericht op het onderscheidend vermogen over meerdere toetsen heen in plaats van binnen de toets. Uit eerder onderzoek blijkt dat een vraagformat met een sterk onderscheidend vermogen binnen een toets niet vanzelfsprekend ook een sterk onderscheidend vermogen over meerdere toetsen heen. We bestudeerden welk vraagformat, VSAQ of meerkeuzevraag, beter in staat is om studenten te onderscheiden op basis van hun academische prestatieniveau, gemeten als cijfergemiddelde. In deze retrospectieve studie werden de scores geanalyseerd van twee summative toetsten uit het eerste en tweede jaar van de medische bachelor, die bestonden uit zowel VSAQ's als meerkeuzevragen. De analyses werden uitgevoerd op drie studentcohorten, terwijl binnen elk cohort twee studentpopulaties werden onderscheiden: de eerste populatie bestond uit eerstejaarsstudenten die het eerstejaarstentamen hadden gemaakt, en de tweede populatie bestond uit tweedejaarsstudenten die zowel het eerste- als het tweedejaarstentamen in opeenvolgende jaren hadden afgelegd. De VSAQ's hadden een sterkere positieve associatie met het cijfergemiddelde dan meerkeuzevragen in alle cohorten. Daarnaast waren ze over het algemeen ook beter in staat om slecht en excellent presterende studenten te onderscheiden op basis van hun cijfergemiddelde. Hiermee tonen we aan dat VSAQ's beter in staat zijn om studenten met verschillende academische prestatie niveaus te identificeren. Het gebruik van VSAQ's in toetsen kan daarmee het vroegtijdig identificeren van zowel slecht als excellent presterende studenten verbeteren, waardoor er gepaste begeleiding of extra uitdaging kan worden aangeboden.

In **hoofdstuk 4** richten we ons op een andere waardevolle eigenschap van toetsen, namelijk hun gunstige effect op het vasthouden van kennis op de lange termijn (i.e., kennisretentie). Het ophalen van kennis uit het geheugen, bekend als '*retrieval practice*' of het '*testeffect*', is een bewezen effectieve strategie om leren te verbeteren en kennisretentie te vergroten. Deze strategie wordt vaak toegepast via oefentoetsen, maar het is nog onduidelijk welk vraagformat hierbij het meest effectief is. In een gerandomiseerde studie onderzochten we het effect van oefentoetsen met VSAQ's en meerkeuzevragen op de kennisretentie van studenten. Studenten in een extracurriculaire cursus over leefstijl beantwoordden zowel VSAQ's als meerkeuzevragen, zonder dat ze na afloop feedback kregen. Drie weken na de laatste oefentoets werd hun kennis getest met identieke vragen in beide formats in een eindtoets. Hoewel de VSAQ's moeilijker bleken en studenten hier lagere scores op haalden, wat doorgaans een positief effect heeft op kennisretentie, werd er geen verschil in retentie tussen de vraagformats gevonden. Dit betekent dat we geen bewijs hebben gevonden dat één van de vraagformats effectiever is in het bevorderen van kennisbehoud door middel van *retrieval practice*.

Een mogelijke verklaring voor het ontbreken van een verschil in effectiviteit is dat studenten bij de oefentoetsen lagere scores behaalden op de VSAQ's. Dit kan hebben geleid tot minder succesvol ophalen van kennis, waardoor de retentie op de eindtoets mogelijk ook lager uitviel. Om de effectiviteit van VSAQ's in *retrieval practice* te verbeteren en de kennisretentie te vergroten, kan het helpen om de slagingskans (het succes) bij het oefenen te verhogen. Dit kan bijvoorbeeld door directe feedback te geven en herhaalde oefenmomenten aan te bieden.

We sluiten deel I af met een opinie stuk in **hoofdstuk 5** over de rol van toetsing en vraagformats in het medisch onderwijs. We stellen hierbij de rol van de meerkeuzevraag als standaard vraagformat ter discussie. VSAQ's bieden namelijk een veelbelovend alternatief met meerdere voordelen: ze zijn betrouwbaarder, stimuleren actief leren en bereiden studenten beter voor op de klinische praktijk, waar patiënten zich ook niet met antwoordopties presenteren. Hoewel VSAQ's meer nakijktijd kosten dan meerkeuzevragen, blijkt dit relatief snel en efficiënt te kunnen met de huidige digitale toetsomgevingen. Daarnaast hoeven er voor dit vraagformat geen plausibele foutieve antwoordopties bedacht te worden zoals bij meerkeuzevragen, wat tijd bespaart in de vraagontwikkeling. Om de implementatie van VSAQ's in het medisch onderwijs soepel te laten verlopen, bevelen we onder andere workshops voor docenten en oefentoetsen voor studenten aan. Deze benadering helpt docenten ervaring op te doen met het ontwikkelen en nakijken van VSAQ's, terwijl studenten wennen aan het vraagformat en tegelijkertijd hun kennisniveau kunnen testen.

Deel II: Computer adaptieve voortgangstoets

De voortgangstoets (VGT) is een toets die vier keer per jaar door alle geneeskundestudenten wordt gemaakt om te volgen hoe hun kennis zich tijdens hun opleiding ontwikkelt. De toets bevat vragen uit alle vakgebieden van de opleiding en wordt steeds opnieuw aangeboden, waardoor studenten en docenten inzicht krijgen in wat goed beheerst wordt en wat nog extra aandacht vereist. Oorspronkelijk maakten alle studenten precies dezelfde toets, ongeacht hun studiejaar. Dit zorgde vooral bij jongerejaars studenten voor problemen, omdat zij regelmatig vragen kregen die te moeilijk waren voor hun kennisniveau, waardoor de toets minder betrouwbaar werd. *Computerized Adaptive Testing* (CAT) biedt hier mogelijk een oplossing, waarbij de moeilijkheid van de vragen wordt afgestemd op het individuele niveau van de student tijdens het maken van de toets. In **hoofdstuk 6** beschrijven we een cross-over studie waarin we de conventionele VGT vergeleken met een CAT-variant. Geneeskundestudenten van drie medische faculteiten maakten zowel de conventionele VGT als de CAT, waarna we de correlatie tussen de resultaten en de uitvoerbaarheid van de CAT onderzochten. De scores op de CAT bleken sterk te correleren met de scores op de conventionele VGT en de CAT werd zonder problemen afgenomen. De gemiddelde afnametijd werd met de CAT verkort tot 83 minuten, wat leidt tot een efficiëntere toets zonder daarbij aan betrouwbaarheid in te leveren. Deze positieve bevindingen hebben het invoeren van een landelijke CAT in het medisch curriculum ondersteund, waarbij een goede voorbereiding en voorlichting van studenten belangrijk is, aangezien studenten de CAT als uitdagender ervaren.

In de conventionele VGT werden punten afgetrokken voor incorrecte antwoorden. Studenten konden echter ook een vraagtekenoptie kiezen als ze het antwoord niet wisten, zonder dat dit tot puntenaftrek leidde. Deze scoringsmethode, bekend als '*formula scoring*', is bedoeld om gokken te ontmoedigen. Uit onderzoek blijkt echter dat het gebruik van de vraagtekenoptie sterk verschilt tussen studenten

en mogelijk samenhangt met risicogedrag en andere persoonlijke factoren. Met de introductie van de CAT ontstond een unieke situatie: studenten maakten zowel een VGT mét (conventioneel) als zonder (CAT) vraagtekenoptie. Hierdoor konden we op grote schaal onderzoeken hoe deze vraagtekenoptie de prestaties van studenten beïnvloedt. In **hoofdstuk 7** is de relatie onderzocht tussen het gebruik van de vraagtekenoptie in de conventionele VGT en de prestaties op de CAT. De resultaten laten zien dat er grote verschillen zijn in het gebruik van de vraagtekenoptie en het effect hiervan op de CAT-scores, zowel binnen als tussen studiejaren. Zo zagen we bijvoorbeeld dat in de eerste jaren van de studie het gebruik van meer vraagtekens, dus meer vragen onbeantwoord laten, samenhing met hogere scores op de CAT, terwijl dit effect in latere studiejaren minder werd en in het laatste jaar zelfs omkeerde waarbij meer vraagtekengebruik samenhing met lagere scores op de CAT. In jaar vier observeerden we de meeste variatie in de effecten van vraagtekengebruik op CAT scores tussen verschillende groepen studenten. Dit suggereert dat *formula scoring* niet alleen kennis meet, maar ook andere constructen, zoals invulgedrag en metacognitieve vaardigheden. Daarom is het belangrijk om de scoringmethode zorgvuldig te kiezen en af te stemmen op de toetsdoelen, zodat de resultaten eerlijk en betrouwbaar blijven. Als het doel van de toets puur het meten van het kennisniveau is, lijkt formula scoring hier niet geschikt voor.

Deel III: Feedback

Formatieve toetsen hebben als doel studenten inzicht te geven in hun sterke en zwakke punten door middel van uitgebreide feedback, zodat zij hun leerproces gericht kunnen verbeteren. Ondanks de toenemende implementatie van formatieve toetsen in het medisch onderwijs, blijkt het in de praktijk echter nog moeilijk om studenten optimaal gebruik te laten maken van de feedback. In **hoofdstuk 8** onderzochten we hoe studenten zich voorbereiden op formatieve en summatieve VGT's, hoe zij omgaan met de feedback na de verschillende toetsen en hoe hun motivatie verschilt wanneer een toets wel of niet meetelt voor studiepunten. In deze studie hebben we vragenlijstdata, inloggegevens van het online feedbacksysteem voor de VGT en interviewgegevens geanalyseerd om het feedbackgebruik en de motivatie voor de toetsen tussen de formatieve groep en de summatieve groep te vergelijken. Uit onze studie bleek dat studenten die de summatieve VGT maakten vaker de feedback raadpleegden, maar de voorbereiding en het actief gebruik van de feedback voor het eigen leerproces waren over het algemeen laag en vergelijkbaar in beide groepen. Onze interviewdata toonden aan dat er een onderscheid gemaakt kon worden tussen prestatiegerichte en leergerichte studenten. Prestatiegerichte studenten waardeerden de formatieve VGT minder vanwege het gebrek aan studiepunten en besteedden daarom minder aandacht aan deze toets en de bijbehorende feedback. Leergerichte studenten daarentegen zagen de formatieve VGT wel als een kans om hun kennis te evalueren en verbeteren en waren gemotiveerd om deze toets serieus te maken en de feedback te bekijken. Deze bevindingen suggereren dat studenten over het algemeen minder gemotiveerd zijn om zich in te zetten voor een toets en de bijbehorende feedback wanneer er geen directe (studie)consequenties aan verbonden zijn. Een leeromgeving die het belang van formatieve toetsing duidelijk benadrukt en ondersteunt is daarom essentieel om studenten te motiveren actief gebruik te maken van feedback in hun leerproces.

De studie in **hoofdstuk 9** onderzocht via interviews welke processen en factoren het feedbackgebruik van studenten in de context van de VGT beïnvloeden. Studenten gaven aan dat ze de feedback vaak moeilijk konden interpreteren, dat deze te laat werd gegeven en onvoldoende specifiek was.

Ook misten zij begeleiding in het toepassen van de feedback, waardoor veel studenten alleen actie ondernamen na een onvoldoende. Er werd meer gebruik gemaakt van de feedback in de coschappen, wanneer studenten een serieuzere houding ontwikkelden ten opzichte van hun studie en carrière perspectieven. Om het feedbackgebruik te stimuleren, bevelen we aan om feedback helder en tijdig beschikbaar te maken en om begeleiding aan te bieden, zodat de VGT niet alleen als meetinstrument, maar ook als leermiddel effectiever kan worden ingezet. Bovendien kan een betere integratie van de VGT en de feedback in het curriculum de betrokkenheid van studenten vergroten en hen beter voorbereiden op het toekomstige beroepsleven.

Op basis van de bevindingen in dit proefschrift geven we enkele praktische aanbevelingen voor het onderwijs:

Gebruik VSAQ's in zowel summatieve als formatieve toetsen
VSAQ's bieden een betrouwbaardere manier om kennis te toetsen dan meerkeuzevragen. Ze voorkomen gokken en geven een nauwkeuriger beeld van het werkelijke kennisniveau van studenten. Door het gebruik van VSAQ's kan de kwaliteit van toetsen worden verbeterd. Bovendien kunnen ze door hun goede onderscheidende vermogen docenten helpen bij het vroegtijdig identificeren van studenten die extra ondersteuning nodig hebben. Om studenten te laten wennen aan dit vraagtype, kan het nuttig zijn om VSAQ's al vroeg in het curriculum in formatieve (oefen) toetsen in te zetten. Zo raken studenten vertrouwd met het vraagtype, kunnen ze hun kennishiaten ontdekken en zich beter voorbereiden op summatieve toetsen. Om de kennisretentie te vergroten (het zogenaamde testeffect), is het zinvol om de oefentoetsen te combineren met feedback, zodat studenten direct inzicht krijgen in hun kennisniveau en zwakke punten. Daarnaast is het raadzaam om docenten goed voor te bereiden en te ondersteunen, bijvoorbeeld via interactieve workshops waarin zij leren hoe ze goede VSAQ's opstellen.

Implementeer CAT op grote schaal om studenten in verschillende stadia van hun studie te toetsen
CAT is een innovatieve toetsmethode waarbij de moeilijkheidsgraad van de vragen zich tijdens de toets automatisch aanpast aan het niveau van de student. Uit ons onderzoek blijkt dat CAT betrouwbaar, efficiënt en praktisch haalbaar is voor voortgangstoetsen, omdat studenten op verschillende momenten in hun studie nauwkeurig kunnen worden getoetst. Om CAT succesvol in te voeren, raden we aan om studenten duidelijk voor te bereiden door middel van informatiesessies, oefenmogelijkheden en heldere communicatie over de voordelen van CAT. Hoewel de invoering van CAT een investering vraagt, zijn de voordelen op lange termijn — zoals verbeterde betrouwbaarheid en flexibiliteit — zeer waardevol. Daarnaast is het belangrijk dat verschillende onderwijsinstellingen samenwerken om een uitgebreide, actuele en kwalitatief sterke vragenbank te ontwikkelen en te onderhouden.

Maak feedback een integraal onderdeel van het leerproces
Feedback kan een sterk positief effect hebben op leren en zou daarom een vanzelfsprekend onderdeel moeten zijn van het leerproces van studenten. Hiervoor is het belangrijk dat studenten begeleiding krijgen bij het effectief inzetten van feedback voor hun eigen ontwikkeling. Goede integratie van feedback vraagt duidelijke instructies, goede toegankelijkheid en regelmatige aandacht voor feedback binnen het curriculum. Vooral bij toetsen die minder direct aansluiten bij het actuele onderwijs, zoals voortgangstoetsen, kan het zinvol zijn om feedback actief te bespreken in kleine groepen met een docent.

Op die manier kunnen studenten hun resultaten beter begrijpen, vragen stellen en leerstrategieën ontwikkelen. Het aanbieden van aanvullende oefenvragen die aansluiten bij de inhoud van voortgangstoetsen helpt studenten bovendien om de verbinding te zien tussen feedback en hun onderwijs. Zo wordt feedback een effectief hulpmiddel voor studenten om hun leerproces te verbeteren.

Conclusie

Toetsing is meer dan alleen een meetinstrument; het is een sterke drijfveer voor leren. Door innovatieve methoden zoals VSAQ's, CAT en feedback, wordt toetsing niet alleen betrouwbaarder, maar ook een waardevol leermiddel voor studenten. VSAQ's verbeteren de kwaliteit van toetsen en stimuleren authentiek leren, terwijl CAT zorgt voor een efficiëntere, nauwkeurigere en meer gepersonaliseerde toetsing. Structureel geïntegreerde feedback kan studenten bovendien motiveren om actief met hun eigen leerproces aan de slag te gaan en zich te blijven ontwikkelen. Door deze verschillende methoden te combineren kunnen we toetsing betrouwbaarder en betekenisvoller maken voor studenten, zodat zij optimaal worden voorbereid op hun toekomstige carrière.