Implementing assessment innovations in higher education

Boevé, Anna Jannetje

Samenvatting

Inleiding

In dit proefschrift staat toetsing in het (universitair) hoger onderwijs centraal. De onderzoeken die zijn uitgevoerd in hoofdstuk twee tot en met zes zijn een verzameling van studies waarin de implementatie van verschillende innovaties op het gebied van toetsing in het onderwijs aan de Rijksuniversiteit Groningen zijn onderzocht.

Enkele recente belangrijke ontwikkelingen vormden de aanleiding voor het onderzoek in dit proefschrift. Deze zijn: de digitalisering van de maatschappij die ook in het onderwijs merkbaar is en tot verschillende veranderingen leidt; de groeiende studentaantallen, maar ook de politieke ontwikkelingen zoals het recent ingevoerde model van prestatiebekostiging voor hogeronderwijsinstellingen.

Doordat het aantal studenten in het hoger onderwijs is toegenomen in de afgelopen jaren (Hornsby & Osman, 2014), krijgen docenten te maken met steeds grotere groepen van soms wel honderden studenten. Hierdoor wordt de verhouding docent(en) ten opzichte van het aantal studenten erg klein, waardoor docenten maar zeer beperkte tijd en middelen ter beschikking hebben om de kwaliteit en voortgang van het leerproces van studenten te waarborgen.

De toenemende digitalisering is ook van belang in alle lagen van het onderwijs, dus ook voor het hoger onderwijs. Enerzijds is het niet meer mogelijk om toegang tot het hoger onderwijs te krijgen zonder de beschikking te hebben over digitale middelen, anderzijds blijft de rol van digitale middelen soms zeer beperkt in het onderwijs. Studenten moeten bijvoorbeeld dikwijls nog tentamens op papier maken, welke vervolgens ook met de hand worden nagekeken. En hoewel docenten gestimuleerd worden om digitale middelen in colleges te gebruiken, zien docenten hier soms van af en willen ze het gebruik van digitale middelen juist weer beperken Er ligt dus een uitdaging voor management en docenten om zo goed mogelijk gebruik te maken van digitale middelen op een manier die de kwaliteit van het leerproces ten goede komt.

Een andere aanleiding voor het onderzoek in dit proefschrift was de prestatiebekostiging van het hoger onderwijs in Nederland; in sommige is deze al eerder, op verschillende manieren ingevoerd (De Boer et al., 2015). De prestatiebekostiging in Nederland houdt in dat sinds 2012 afspraken zijn gemaakt tussen hogeronderwijsinstellingen– universitair en HBO – en de overheid. Wanneer instellingen de gestelde doelen halen binnen de afgesproken termijn, dan blijft de financiering van kracht. Wanneer een instelling niet de gestelde doelen behaald, wordt de financiering door de overheid gekort. Duidelijke indicatoren binnen deze afspraken zijn uitvalpercentages en het afstudeerrendement na vier jaar. Door deze prestatiebekostiging zou de kwaliteit van het onderwijs beter gewaarborgd worden. Hoewel minister Bussemaker (2014) heeft erkend dat kwaliteit niet alleen gemeten kan worden door uitval en rendement, bleven de andere kwaliteitsindicatoren vaag. Bussemaker (2014) stelde echter wel: “Nieuwe ontwikkelingen als open-online onderwijs bieden mogelijkheden om de kwaliteit van het hoger onderwijs verder te verbeteren. Het hoger onderwijs moet aan al deze

[2] Zie https://nos.nl/op3/artikel/2133931-laptops-in-de-collegezaal-streng-verboden.html

(bestaande en nieuwe) uitdagingen de juiste aandacht geven." In dit proefschrift wordt de implementatie van enkele van deze nieuwe mogelijkheden onderzocht in de praktijk.

Assessment en toetsen

In de wetenschappelijke literatuur wordt een onderscheid gemaakt tussen de termen 'assessment' en toetsen. Met assessment wordt het proces van informatieverzameling bedoeld waarmee men inzicht krijgt in de kennis en kunde van studenten. Deze informatie kan gebruikt worden voor diverse doeleinden, zoals de richting van vervolginstructie, diagnose van sterke/zwakke kanten van studenten en/of het nemen van beslissingen over studenten. Een toets daarentegen is een verzameling van vragen of opdrachten die tot doel heeft de kennis en/of kunde op een specifiek tijdstip van een student te meten. Een toets kan en is meestal onderdeel van een assessment-programma. In dit proefschrift gaan sommige hoofdstukken specifiek over toetsen (de hoofdstukken 2, 3 en 6), terwijl andere hoofdstukken over assessment gaan (de hoofdstukken 4 en 5).

Toetsen kunnen met verschillende bedoelingen ontworpen en/of ingezet worden. Wanneer het vooral belangrijk is om een beslissing te nemen – of om te selecteren – wordt de functie van een toets omschreven als 'summatief'. Voorbeelden van summatieve toetsen zijn toelatingstoetsen of tentamens aan het einde van een onderwijsperiode. Formatieve toetsen daarentegen zijn bedoeld om het leerproces te bevorderen en/of bij te sturen. Dit onderscheid wordt ook wel omschreven als het 'toetsen van leren' of het 'toetsen voor leren' (Schuwirth & van der Vleuten, 2011). In de praktijk is het onderscheid tussen formatieve en summatieve toetsen niet altijd even duidelijk, zoals duidelijk wordt in hoofdstukken 3 tot en met 5 van dit proefschrift.

Er zijn verschillende tradities van onderzoek naar de kwaliteit van toetsing. Men spreekt in de literatuur van "high-stakes" grootschalige toetsing, wanneer beslissingen op basis van de toetsen van groot belang zijn en voor zeer grote aantallen worden uitgevoerd. De bekendste vorm van grootschalig high-stakes toetsing in Nederland is de eindtoets van de basisschool die mede bepalend is naar welk vervolgonderwijs kinderen gaan. Ook in andere landen zijn er dergelijke "high-stakes" toetsen zoals bijvoorbeeld de SAT in de Verenigde Staten, die gebruikt wordt om studenten te selecteren voor het hoger onderwijs. Deze toetsen zijn regelmatig in het nieuws vanwege de vermeende negatieve effecten op de selectie van minderheden in het hoger onderwijs. Wat minder bekend bij dit soort grootschalige toetsen, is dat deze toetsen – los van het gebruik in de praktijk – aan strenge kwaliteitseisen moeten voldoen, waardoor hier doorgaans veel onderzoek naar gedaan wordt. Dit heeft bijgedragen aan de ontwikkeling van theorie en statistische methoden om de toetsen te beoordelen.

Naast de onderzoekstraditie van het grootschalige toetsen is er recent meer aandacht voor het kleinschaliger "classroom testing" waarbij toetsing vooral in dienst staat van het onderwijsleerproces. Deze toetsen worden doorgaans ontwikkeld door docenten voor relatief kleine groepen leerlingen of studenten. Vaak zijn niet de tijd en middelen beschikbaar om uitgebreid de kwaliteit van de toetsen te onderzoeken zoals dit wel het geval is bij grootschalige "high-stakes" toetsen. In de wetenschappelijke

literatuur is een groeiende interesse om het onderzoek naar grootschalig en kleinschalig toetsen te integreren. In dit proefschrift worden ook methoden die ontwikkeld zijn voor grootschalige toetsing toegepast bij kleinschalig “classroom testing” in het universitair onderwijs.

De context van het onderzoek in dit proefschrift

De studies die in dit proefschrift staan beschreven zijn uitgevoerd aan de Rijksuniversiteit Groningen (RUG) en de meeste daarvan zijn gedaan in het propedeusejaar. De resultaten behaald in het eerste jaar van de bachelor zijn belangrijk omdat er een sterk verband is tussen deze resultaten en de resultaten in het vervolg van de opleiding (Niessen, Meijer & Tendeiro, 2016) en omdat studenten te maken hebben met het bindend studieadvies (BSA). Het BSA is gebaseerd op een minimaal aantal studiepunten dat een student moet behalen – aan de RUG bijvoorbeeld 45 van de 60 ECTS – in het eerste jaar om door te mogen gaan met hun studie. Dus wanneer een student dit aantal niet haalt, mogen ze de opleiding niet meer vervolgen. Hier zijn ook financiële implicaties mee gemoeid: wanneer een student na februari besluit te stoppen – of het BSA niet haalt – dan blijft het collegegeld verschuldigd. Hierdoor zijn alle studieresultaten in het eerste jaar in de beleving van studenten ook wel te omschrijven als “high stakes”.

Samenvatting van de resultaten van het uitgevoerde onderzoek

De studies in dit proefschrift zijn uitgevoerd in samenwerking met docenten die hun onderwijs wilden verbeteren door veranderingen in toetsing door te voeren, waarbij veelal digitale middelen werden ingezet. In hoofdstuk 2 tot en met 6 worden de verschillende studies beschreven, Hoofdstuk 1 bevat een algemene introductie en in hoofdstuk 7 wordt op de resultaten teruggeblikt. Hieronder volgt een korte samenvatting van elk onderzoek:

In hoofdstuk 2 wordt de implementatie van digitale tentamens onderzocht. Digitale tentamens bieden mogelijkheden om de kwaliteit van toetsen te verbeteren en het toetsproces te vergemakkelijken. Daarentegen is het van belang om ervoor te zorgen dat de prestaties op de traditionele en digitale toetsen vergelijkbaar zijn, dat de studenten de toetsen als eerlijk ervaren en dat de stress die deze nieuwe vorm van toetsing met zich meebrengt minimaal is (Whitelock, 2009). De belangrijkste onderzoeksvragen in het onderzoek dat wordt beschreven in hoofdstuk 2 waren: is er een verschil in de prestatie tussen studenten die digitaal en schriftelijke werden getoetst? En: hoe ervaren studenten digitale toetsen? Dit is onderzocht door middel van een zogeheten ‘veldexperiment’, een experiment buiten een laboratorium. Studenten die het vak ‘Biopsychologie’ volgden – in 2013/2014 in de Nederlandstalige psychologieopleiding en in 2014/2015 in de Engelstalige psychologieopleiding – maakten op willekeurige basis ofwel de eerste deeltoets ofwel de tweede deeltoets digitaal en de ander op papier. De toetsresultaten van de groep studenten die de deeltoets op papier had gemaakt waren nagenoeg hetzelfde vergeleken met de groep studenten die de deeltoets digitaal had gemaakt, voor zowel de eerste als tweede deeltoets. Ongeveer een kwart van de studenten bleek een voorkeur te hebben voor de digitale toets, ongeveer de helft een voorkeur voor

de papieren versie en ongeveer een kwart had geen voorkeur (met uitzondering van de tweede deeltoets van de internationale opleiding waarbij de voorkeur voor digitale en papieren tentamens ongeveer gelijk was). Aangezien ook uit de literatuur blijkt dat studenten geen grote voorkeur hebben voor digitale toetsen, is het belangrijk dat instellingen de overgang naar digitaal toetsen zo inrichten dat studenten controle ervaren over het tentamen. Uit aanvullend kwalitatief onderzoek bleek dat dit op verschillende manieren zou kunnen worden gerealiseerd: door studenten bijvoorbeeld de mogelijkheid te bieden om bij digitale toetsen te kunnen onderstrepen, doorhalingen en markeringen te maken of door het flexibeler aanbieden van vragen.

In hoofdstuk 3 wordt het nut van het rapporteren van deelscores op een toets onderzocht. Gegeven de beperkte tijd en middelen van docenten in het hoger onderwijs, zou het wenselijk kunnen zijn om op efficiënte wijze diagnostische feedback aan studenten te geven door middel van het rapporteren van deelscores. Dit gebeurt soms bij “large-scale testing” en is ook steeds vaker een afweging bij “classroom testing”. Er is echter ook wetenschappelijke literatuur die aantoont dat dit maar beperkt zinvol is (bijv. Sinharay, 2010). In hoofdstuk 3 wordt aan de hand van twee verschillende tentamens geïllustreerd hoe onderzocht kan worden of het zinvol is om naast de totaalscore op de hele toets ook deelscores te rapporteren. Voor een van de tentamens werd naast de totaalscore ook deelscores van verschillende soorten kennis berekend. Voor het andere tentamen werd een deelscore berekend voor de open vragen en voor de meerkeuze vragen. In beide gevallen bleek het niet zinvol om de deelscores te rapporteren, dit kwam doordat deelscores relatief onbetrouwbaar waren en hoog correleerden met de totaalscore. Wel was interessant dat een deel van de open vragen sterk bijdroegen aan het vergroten van de meetprecisie van het totale tentamen.

In hoofdstuk 4 en 5 wordt onderzocht wat de effecten waren van het gebruik van digitale leermiddelen om het leerproces en de resultaten te verbeteren. In hoofdstuk 4 is de implementatie van digitale oefentoetsen onderzocht in verschillende vakken. In het eerste deel van deze studie is gekeken naar het gebruik van oefentoetsen door studenten in twee statistiekvakken en een vak over biopsychologie. Voor de statistiekvakken bleek dat het gebruik van oefentoetsen nauwelijks voorspellend was voor het tentamencijfer van de studenten. Voor het vak over biopsychologie daarentegen was er een duidelijk positieve samenhang tussen de mate van gebruik van oefentoetsen en de tentamenscore. Een mogelijke reden voor de verschillende resultaten kan de cursusinrichting zijn – de studenten bij de statistiekvakken hadden naast hoorcolleges en oefentoetsen ook verplichte werkgroepen en huiswerk. Bij biopsychologie waren geen verplichtingen behalve het maken van het tentamen en konden studenten facultatief naar de hoorcolleges. Tevens is het belangrijk om te erkennen dat een positieve samenhang tussen oefentoets gebruik en tentamenscore niet noodzakelijk iets zegt over de effectiviteit van de oefentoetsen. Een alternatieve verklaring kan zijn dat vooral de gemotiveerde studenten gebruik hebben gemaakt van de oefentoetsen en dat deze studenten ook een goed cijfer hadden behaald wanneer er geen oefentoetsen beschikbaar waren. Daarom is ook een tweede deelstudie uitgevoerd voor het vak biopsychologie.

In de tweede studie van hoofdstuk 4 is gekeken naar het verschil in gemiddelde toetsscores van een cohort biopsychologiestudenten die geen beschikking hadden over oefentoetsen en twee cohorten biopsychologiestudenten die gedeeltelijk of geheel de beschikking hadden over oefentoetsen. Er werd gebruik gemaakt van test-equating om de scores van de verschillende cohorten met elkaar te vergelijken. Uit de resultaten bleek dat er nauwelijks verschil was in de prestaties tussen de groepen die wel of geen beschikking hadden over de oefentoetsen. Het onderzoek in dit hoofdstuk liet ook zien dat het in de praktijk soms lastig is om bevindingen uit (experimenteel) wetenschappelijk onderzoek daadwerkelijk in de praktijk te implementeren. In de praktijk is het belangrijkste dilemma bij het invoeren van formatief toetsen de mate waarin het verplicht zou moeten worden. Hierbij is het belangrijk te beseffen dat wanneer deelname aan formatieve toetsing verplicht wordt, de toets automatisch summatiever wordt. Eerder onderzoek heeft aangetoond dat zelfs niet-dwingende maatregelen zoals bonuspunten negatieve gevolgen kunnen hebben voor de formatieve werking van de toets (Kibble, 2007). Aan de andere kant betekent dat het dat bij daadwerkelijk formatieve toetsing studenten de verantwoordelijkheid moeten nemen voor hun eigen leerproces, met het risico dat de docent middelen beschikbaar stelt die niet gebruikt worden.

In hoofdstuk 5 staat onderzoek naar de implementatie van de “flipped classroom” centraal. In de flipped classroom gaan studenten tijdens het hoorcollege actief aan de slag met de leerstof en vind kennisoverdracht – met hulp van bijvoorbeeld videoclips – ook plaats voorafgaand aan de les (Abeysekera & Dawson, 2015; Street, Gilliland, McNeil, & Royal, 2015). De populariteit van de flipped classroom neemt toe in het (universitair) hoger onderwijs. Hoewel er enig onderzoek is gedaan naar de prestaties van groepen studenten die flipped classroom onderwijs volgden, is er nog weinig bekend over het studiegedrag van studenten in de flipped classroom. Dit studiegedrag is belangrijk omdat het centraal staat in het leerproces en prestaties van de flipped classroom. In hoofdstuk 5 is een studie verricht waarbij een groep studenten een statistiekvak volgden in de vorm van de flipped classroom en een groep studenten die een statistiekvak volgden in de traditionele vorm. Twee keer in de week werden studenten gevraagd om in te vullen hoeveel tijd zij hadden besteed aan het vak en welke studieactiviteiten ze hadden ondernomen voor het statistiekvak. Uit de resultaten bleek dat het studiepatroon van de twee groepen gedurende het vak sterk op elkaar leek en tevens dat de gemeten studiegedrag (tijd en activiteiten) ook niet een sterke samenhang vertoonde met de verkregen tentamenresultaten. In een verdere exploratie van de vakevaluaties voor de groep flipped classroom-studenten werd specifiek gekeken naar evaluaties met betrekking tot hun studiegedrag en perceptie van de “flipped classroom”. Sommige studenten vonden inderdaad dat de “flipped classroom” hun leerproces ondersteunde. Andere studenten daarentegen waren om diverse redenen niet bereid om hun studiegedrag te veranderen in de flipped classroom. Dit biedt interessante mogelijkheden voor vervolg onderzoek. Hoewel er voldoende wetenschappelijk theoretische gronden zijn die zouden moeten ondersteunen dat de flipped classroom een goed idee is, is veel onderzoek gericht op het aantonen van verbeterde prestaties en niet op het gedrag van studenten dat ten grondslag ligt aan de theorie en implementatie

van de “flipped classroom”. Het onderzoek in dit hoofdstuk liet zien dat de zelfregulatie van studenten en hun bereidheid om mee te gaan met de gedragsverandering belangrijk is bij de implementatie van de flipped classroom.

Geïnspireerd door het onderzoek in hoofdstukken 2 tot en met 5, heeft het onderzoek in hoofdstuk 6 een meer methodologisch karakter. In onderzoek naar innovaties of veranderingen in het onderwijs is de prestatie van studenten dikwijls de belangrijkste uitkomst. In de meest gangbare type onderzoek worden of bestaande groepen studenten over verschillende jaren met elkaar vergeleken, of worden twee bestaande verschillende groepen in dezelfde periode met elkaar vergeleken. Dit type onderzoek wordt gebruikt omdat het meestal onmogelijk is om willekeurig studenten aan groepen toe te wijzen zoals bijvoorbeeld gebeurt bij gerandomiseerde gecontroleerde trials. Het nadeel van het gebruik van bestaande groepen studenten is dat alternatieve variabelen naast “de treatment” van invloed kunnen zijn. Dus de verschillen in de prestaties van groepen studenten in verschillende condities hoeven niet het resultaat van bijvoorbeeld ingevoerde onderwijsvernieuwingen. De onderzoeksvragen in hoofdstuk 6 was: in welke mate fluctueren de prestaties van studenten in eerstejaarsvakken over tijd en tussen vakken? Hoe kan deze informatie gebruikt worden om onderwijsinnovaties te evalueren? Om deze vraag te beantwoorden zijn de resultaten van eerstejaars vakken over een periode van zes jaar aan de Rijksuniversiteit Groningen geanalyseerd. In totaal kon 17% van de variatie in cijfers toegekend worden aan fluctuatie over tijd en tussen vakken. Verder kon 40% van de variatie in slagingspercentages worden toegekend aan fluctuaties over tijd en tussen vakken. Gebruikmakend van deze informatie wordt in hoofdstuk 6 geïllustreerd wanneer verschillen in gemiddelde cijfers tussen groepen binnen de natuurlijk te verwachten fluctuatie valt en wanneer er sprake is van een betekenisvol verschil.

Beperkingen van dit onderzoek en toekomstig onderzoek

In het onderzoek in dit proefschrift hebben we geprobeerd een bijdrage te leveren aan een antwoord op de verschillende assessment vragen in het hoger onderwijs. Dit onderzoek werd in de praktijk uitgevoerd, hetgeen ook een aantal beperkingen met zich meebracht. Een beperking van het onderzoek in dit proefschrift is dat het plaats vond aan een enkele universiteit in Nederland. Hierdoor is het mogelijk dat de resultaten niet direct te generaliseren zijn naar andere onderwijsprogramma’s, naar het hbo of instellingen in andere landen.

Een andere beperking, zoals hierboven besproken, is dat het vaak niet mogelijk wanneer onderwijsvernieuwingen worden ingevoerd om experimentele designs te gebruiken, waardoor het moeilijk is om het causale effect van een implementatie vast te stellen. Onderzoek op grote schaal, in de vorm van veldexperimenten, zou misschien een goede aanpak kunnen zijn in de toekomst. Hier zouden verschillende opleidingen binnen meerdere instellingen kunnen deelnemen, zodat wellicht cursussen willekeurig kunnen worden toebedeeld aan onderwijsvernieuwingen.

Een andere belangrijke vraag die nader onderzocht kan worden is de vraag naar de relatie tussen onderwijsvernieuwingen en de context van de vakken waarin deze

geïmplementeerd worden. Het was, bijvoorbeeld, opvallend dat in de statistiekvakken die onderzocht zijn, zowel in hoofdstuk 4 als hoofdstuk 5, dat er nauwelijks een relatie werd gevonden tussen studiegedrag, gebruik van oefentoetsen en het eindcijfer. De statistiekvakken werden naast de vrijblijvende hoorcolleges doorgaans gekenmerkt door meerdere werkvormen en verplichtingen zoals werkgroepen en huiswerk. De onderwijsvernieuwingen zoals de "flipped classroom" en het aanbieden van oefentoetsen hadden misschien geen meerwaarde ten opzichte van de bestaande "good practices". Grootschaliger onderzoek is van belang, om zowel contextuele factoren beter in kaart te brengen, als ook omdat er veel kleinschalige studies in specifieke onderwijscontexten worden gepubliceerd. Dit kan voor vertekening zorgen gezien vooral positieve en statistisch significante resultaten worden gepubliceerd.

Tot slot is het belangrijk dat er wordt nagedacht door diverse belanghebbenden over de te verwachten uitkomsten van onderwijsinnovaties. Wanneer onderwijsinnovaties ten doel hebben om het leerproces van studenten te verbeteren, wat is dan precies de verwachtte uitkomst en voor wie is deze belangrijk? Als het doel is het leerproces te verbeteren, dan moet ook daadwerkelijk evidentie worden verzameld op dit gebied. Ook zou kunnen worden gekeken wat onder "verbeterde prestaties" wordt verstaan. Betekent dit een groter slagingspercentage bij het eerste tentamen of na meerdere tentamens? Voor de gehele groep of voor de minder goede studenten? Deze uitkomsten kunnen informatief zijn voor de effectiviteit van innovaties, maar het is belangrijk om van tevoren te definiëren hoe groot een te verwachten verbetering zou mogen zijn en om informatie over het mechanisme dat tot de verandering leidt te verzamelen. Hierbij is het ook belangrijk om analyses te gebruiken die rekening houden met de praktijk, zoals natuurlijke schommelingen in cijfers en verschillen in de moeilijkheid van tentamens. Door verschillende informatiebronnen te gebruiken en niet exclusief op uitkomsten te focussen, kunnen docenten en onderzoekers beter inzicht krijgen in wanneer onderwijsinnovaties – waaronder innovaties gericht op toetsing – effectief zijn in de praktijk. Een samenwerking is nodig tussen onderzoek en onderwijspraktijk om de kwaliteit van het leren en toetsen in het hoger onderwijs te verbeteren.