Een vergelijking van compensatoir en conjunctief toetsen in het hoger onderwijs

N. Smits, H. Kelderman, en J.B. Hoeksma

Samenvatting

In het Nederlandse hoger onderwijs wordt steeds vaker overwogen om over te stappen op compensatoir toetsen. In deze bijdrage wordt beschreven wat compensatoir toetsen inhoudt en worden de argumenten voor dit toetssysteem beschouwd. Hoewel men vaak voor compensatoir toetsen kiest door de lage betrouwbaarheid van studietoetsen, heeft ook dit toetsregime last van onbetrouwbaarheid; het verschil met het bestaande systeem is welk soort beslisfouten de overhand heeft. We maken daarom onderscheid tussen de aard en precisie van beslisregels. Een probleem is dat de onderbouwing van compensatoir toetsen rust op gemiddelden in plaats van individuele scores, en wordt er vanuit besliskundig oogpunt een bedenkelijke aanname gedaan. Een ander probleem is dat compensatoir toetsen ertoe kan leiden dat fout-positieven op de koop toe worden genomen om fout-negatieven te vermijden. De toename van efficiëntie van het onderwijsproces kan dan ten koste gaan van het voldoen aan de eindtermen. De bijdrage wordt afgesloten met suggesties voor vervolgonderzoek.

kernwoorden: compensatoir toetsen, studiestrategieën, testtheorie

1 Inleiding

Aan steeds meer universiteiten en hoge scholen in Nederland wordt overgegaan op, of nagedacht over het implementeren van, compensatoir toetsen (zie bijvoorbeeld Van Lankveld & Draaijer, 2010; Task Force Studiesucces, 2009; Werkgroep Studiesucces, 2009). Tegelijkertijd zijn sinds enige jaren op het Nederlandse voortgezet onderwijs de exameneisen aangescherpt (Ministerie van OCW, 2008). Ook is men recent in het Vlaamse hoger onderwijs overgestapt op een systeem waarbij men voor elk vak moet slagen waardoor compensatie tussen vakken niet langer mogelijk is (Vlaamse Onderwijsraad, 2007, Adriaens, 2010). Er lijkt dus een spanning tussen beide ontwikkelingen te bestaan.

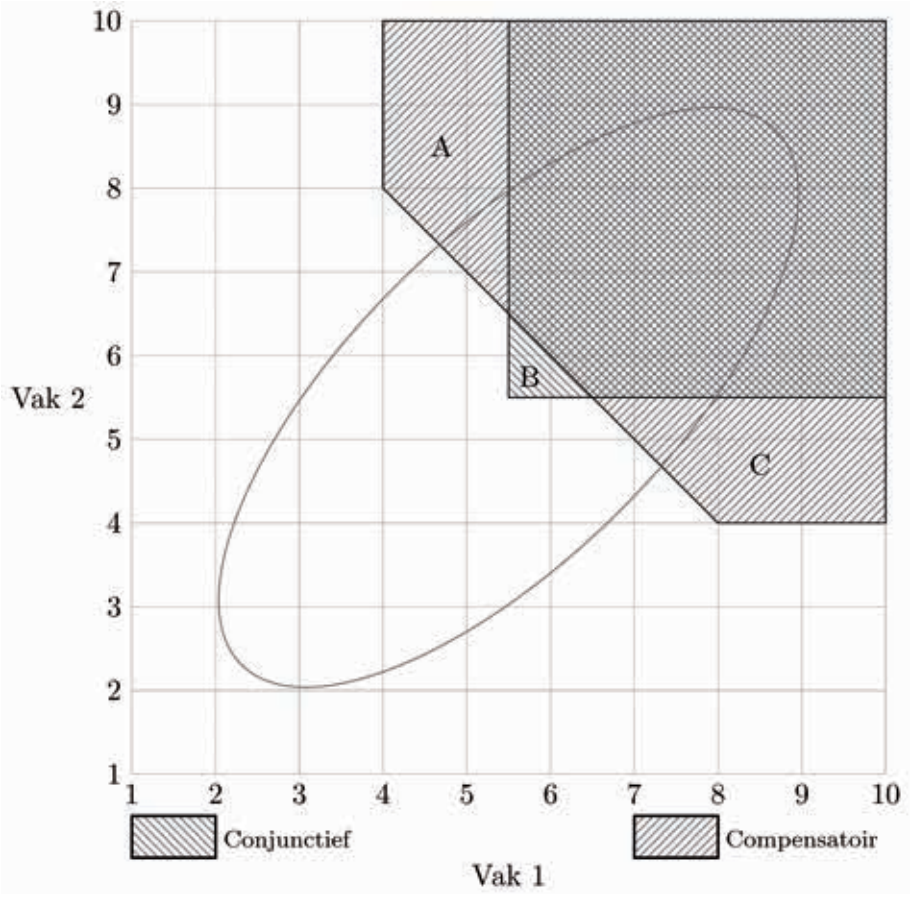
De nieuwe ontwikkelingen in het Nederlands hoger onderwijs hebben vooral aandacht in de media gekregen (zie bijvoorbeeld Bakker, 2012; Peters & Verhoeks, 2012; Arnold & van den Brink, 2012; Bregman, 2013), en in die discussie zijn een aantal uitspraken gedaan die meer nuancering verdienen. Deze bijdrage heeft als doel om te beschrijven wat compensatoir toetsen inhoudt en duidelijk te maken wat de gevolgen kunnen zijn. De nadruk ligt op een wetenschappelijke beschouwing van de argumenten die worden gegeven voor de keuze voor compensatoir toetsen.

2 Wat is compensatoir toetsen?

Compensatoir toetsen houdt in dat men voor een studiejaar of cluster van vakken slaagt indien men gemiddeld over het jaar of het cluster een voldoende cijfer heeft (Rekveld & Starren, 1994). Meestal zijn er bepaalde randvoorwaarden zoals dat geen enkel cijfer lager mag zijn dan een bepaald minimum. Deze manier van toetsen heeft een lange geschiedenis in het Nederlandse voortgezet onderwijs. Bijvoorbeeld, op het eindexamen VWO kunnen één of twee onvoldoendes worden gecompenseerd door de cijfers op andere vakken zolang men gemiddeld een voldoende heeft. Een voorbeeld van een universitaire opleiding waar compensatoir toetsen reeds is ingevoerd is Psychologie aan de Erasmus Universiteit Rotterdam (EUR). In de studiegids voor de Bachelor (Instituut voor Psychologie, FSW, 2013, pp. 11-12) valt te lezen dat in elk bachelorjaar onvoldoendes voor afzonderlijke toetsen binnen een cluster mogen worden gecompenseerd indien het

gemiddelde cijfer minimaal een 6 is en de onvoldoendes niet lager dan een 4 zijn.

De in het hoger onderwijs gebruikelijke manier van toetsen wordt conjunctief toetsen genoemd. Conjunctief toetsen houdt in dat elk tentamen met minimaal een voldoende moet zijn afgerond (Wilbrink, 1979). Vaak representeert een voldoende 55% kennis van de studiestof, en wordt als ondergrens een 5.5 ('afgerond een 6') gehanteerd (De Gruijter, 2008; Mellenbergh, 1993). Om het verschil te illustreren zijn de beslissingen onder conjunctief en compensatoir toetsen in het geval van twee toetsen weergegeven in Figuur 1. In de figuur staan vier gearceerde ruimten afgebeeld. In de geruite ruimte slaagt men onder beide toetssystemen; in de ruimten *A* en *C* slaagt men wel onder het compensatoire systeem, maar niet onder het conjunctieve en in ruimte *B* slaagt men juist wel onder het conjunctieve systeem, maar niet onder het compensatoire. Het verschil in slagingspercentages tussen beide toetssystemen wordt bepaald door hoeveel studenten vallen in ruimten *A* en *C* enerzijds en ruimte *B* anderzijds. Onder welk regime men makkelijker slaagt hangt dus af van verschillende zaken zoals de cesuren die worden gehanteerd onder beide toetsregimes, het gemiddelde cijfer en de standaarddeviatie op elke toets en het verband tussen de toetsen.



Figuur 1. Beslissingen voor twee toetsen bij conjunctief en compensatoir toetsen. De gearceerde gebieden geven de cijfers die bij elk toetssysteem leiden tot slagen. Bij conjunctief toetsen is uitgegaan van minimaal een 5.5 voor voldoende; bij compensatoir toetsen moet er gemiddeld minimaal een 6 worden gehaald, met voor elk vak minimaal een 4. De ellips is een illustratie van hoe de cijfers verdeeld zouden kunnen zijn en vertegenwoordigt een 95%-kansgebied bij een correlatie van 0.7 tussen twee even moeilijke vakken.

3 Is compensatoir toetsen betrouwbaarder?

Een veel gehoorde reden om over te stappen op compensatoir toetsen is de onbetrouwbaarheid van studietoetsen. Bijvoorbeeld, in de studiegids voor de Bachelor Psychologie aan de EUR (Instituut voor Psychologie, FSW, 2013, p. 12) staat: '… de betrouwbaarheid van dergelijke vaktoetsen vaak niet veel groter is dan 70%. Dat betekent dat de uitslag van zo'n toets nooit een 100% nauwkeurige afspiegeling is van het kennisniveau van een individuele student.' De betrouwbaarheid waaraan wordt gerefereerd is die zoals gedefinieerd binnen de klassieke testtheorie (zie bijvoorbeeld, van den Brink & Mellenbergh, 1998): de mate waarin verschillen in ware scores tussen personen kunnen worden geschat met behulp van verschillen in geobserveerde scores. Dit soort betrouwbaarheid is zinvol bij normgeoriënteerd gebruik van studietoetsen, maar veel minder bij criteriumgeoriënteerd gebruik van studietoetsen (van den Brink & Mellenbergh, 1998, pp.19-20). Bij normgeoriënteerd gebruik worden de studieprestaties van studenten onderling vergeleken. Op grond van de studietoets kunnen zij gerangordend worden van de hoogste tot de laagste prestaties. Bij criteriumgeoriënteerd gebruik van een studietoets wordt de prestatie van een individuele student met een onderwijsdoelstelling vergeleken (bijvoorbeeld, tenminste 55% kennis van de studiestof), en zijn de scores van de andere studenten van geen belang. Dat betrouwbaarheidsbepaling in klassieke zin bij criteriumgeoriënteerd gebruik van een studietoets niet altijd zinvol is maakt het volgende voorbeeld duidelijk.[1] Stel, alle studenten hebben voldoende kennis van het vak in termen van de onderwijsdoelstelling, en allen hebben tachtig procent van de vragen van het tentamen goed gemaakt en zijn dus terecht geslaagd. De docent zal zich in dat geval niet druk hoeven te maken om de betrouwbaarheid in klassieke zin, die vrijwel 0 is (zie ook Crocker & Algina, 1986, p. 197). Bij studietoetsen gaat het dus niet om de betrouwbaarheid in klassieke zin, maar om de mate waarin meetfouten de precisie van zak-slaagbeslissingen aantasten. In het onderstaande zal blijken dat beide toetsregimes last hebben van onbetrouwbaarheid, maar dat ze verschillen in wat voor soort beslisfouten de overhand heeft.

Het zij overigens opgemerkt dat meetfouten alleen een effect op de nauwkeurigheid van zak-slaagbeslissing hebben in de buurt van de cesuurscore (zie ook, van Rijn, Béguin & Verstralen, 2012, p. 130). Voor een student die zich niet erg inspant en bijvoorbeeld 55% van studiestof kent is er een redelijk grote kans dat hij ten onrechte zakt voor het tentamen. Voor een student die zich meer inspant en bijvoorbeeld 80% van de studiestof kent is deze kans echter nihil (zie, van den Brink, 1982). De student kan zich met andere woorden wapenen tegen het effect van meetfouten door te zorgen voor meer kennis.

Er zijn vier psychometrische studies die beslisregels, zoals de conjunctieve en compensatoire, voor gecombineerde examens hebben onderzocht. De eerste studie is die van Lord (1962)[2], welke voor Arnold (2011, p. 32) de theoretische basis vormt om over te stappen op compensatoir toetsen. Nauwkeurige lezing van het artikel van Lord laat zien dat hij niet de hierboven beschreven vorm van compensatoir toetsen voorstelt, maar een aangepaste vorm van conjunctief toetsen introduceert waarbij rond de cesuurscore rekening wordt gehouden met de meetprecisie van, en de correlaties tussen de toetsscores. Indien de toetsen hoog positief gecorreleerd zijn is deze methode wat coulanter bij scores net onder de cesuurscore. Toegepast op Figuur 1 betekent dit dat de twee rechte lijnen die het conjunctieve slagingsgebied beschrijven en elkaar linksonder loodrecht ontmoeten, worden vervangen door één hyperbolische curve (schuinliggende U-vorm) waarbinnen ruimte B valt, maar ook delen van ruimten A en C; hoe hoger de correlatie tussen de toetsen is hoe meer er van ruimten A en C in dit gebied valt.

De andere drie studies (Douglas & Mislevy, 2010; Van Rijn, Béguin & Verstralen, 2009; van Rijn et al., 2012) betreffen simulatiestudies waarin de precisie van de regels voor examens met meerdere toetsen werden vergeleken. In deze studies werd voor elke regel de classificatie op basis van de ware, d.w.z. meetfoutloze scores, en

op basis van de geobserveerde, d.w.z. meetfout bevattende scores vergeleken. Hierbij bleek onder andere dat er bij regels met meer compensatie absoluut gezien minder classificatiefouten werden gemaakt, wat een goede reden zou kunnen zijn om voor compensatoir toetsen te kiezen. Echter, de uitkomsten van deze studies zijn moeilijk te interpreteren omdat ze niet voor alle regels hetzelfde, eenduidige, criterium gebruikten. Het criterium betrof de overeenstemming tussen beslissingen over ware en geobserveerde scores binnen regels. Bijvoorbeeld, een student heeft op twee vakken ware cijfers 7 en 4, en geobserveerde cijfers 8 en 5. In dat geval is onder het compensatoire regime de beslissing op basis van de geobserveerde cijfers dezelfde als die op de ware cijfers (namelijk 'slagen'); ook de beslissing onder het conjunctieve regime ('zakken') is voor de geobserveerde en ware cijfers hetzelfde. Omdat de beslissing op dezelfde ware scores niet dezelfde is maakt het de twee regimes onderling onvergelijkbaar; wat is hier de ware beslissing? Het lijkt zinvoller om in simulatieonderzoek juist te kijken naar direct vergelijkbare uitkomsten zoals de verdeling van cijfers van geslaagden onder verschillende toetsregimes; in de Discussie komen we hier op terug.

Een belangrijke aanname bij de vier genoemde studies was dat de cijfers onder de verschillende beslisregels hetzelfde zijn (zie Van Rijn et al., 2009, p. 194). Zoals in een volgende paragraaf zal worden betoogd zullen onder verschillende beslisregels andere studeerstrategieën en daardoor andere verdelingen van cijfers ontstaan, wat uiteraard de validiteit van deze studies ondergraaft.

Er is een fundamenteler probleem met de stelling dat compensatoir toetsen vanuit het oogpunt van betrouwbaarheid te verkiezen is boven conjunctief toetsen. Namelijk, er worden twee zaken door elkaar gebruikt. Aan de ene kant is er de aard van de beslisregel. Een toetsgebruiker heeft op basis van de inhoud en leerdoelen van zijn of haar cursus(sen) een voorkeur voor een bepaalde regel, compensatoir of conjunctief. Aan de andere kant is er de precisie van de gekozen regel. Waarom zou men de aard van de beslisregel aanpassen omdat de precisie in de praktijk tegenvalt (waarom niet zorgen dat de meting preciezer wordt)? Kiezen voor een andere regel leidt weer tot kwalitatief heel andere beslissingen. Een voorbeeld in een andere context maakt dit punt wellicht duidelijker. Stel dat een fabrikant van smartphones hoort dat er fouten worden gemaakt bij de kwaliteitscontrole van het eindproduct: soms wordt een werkende mobiel ten onrechte als defect bestempeld. Door deze onbetrouwbaarheid stapt hij over van een conjunctieve op een compensatoire regel, dat wil zeggen, de mobiel moest eerst op alle eigenschappen voldoende scoren, maar nu moet de mobiel gemiddeld voldoende scoren. Na invoering van de nieuwe regel worden weliswaar minder vaak goede mobieltjes als defect bestempeld, maar wel meer defecte mobieltjes naar de fabriek teruggestuurd. Dit kost veel geld en is bovendien slechte reclame.

4 Is een fout-negatief erger dan een fout-positief ?

Er wordt in de psychometrie onderscheid gemaakt tussen vier typen beslissingssituaties: classificatie, plaatsing, selectie en beheersing (van den Brink & Mellenbergh, 1998). Het beslissen of een student zakt of slaagt voor een studietoets betreft een beheersingssituatie (van den Brink & Mellenbergh, 1998, p. 401). Doordat we te maken hebben met feilbare metingen worden er bij toetsen en tentamens soms verkeerde beslissingen genomen. Een student die de stof eigenlijk wel in voldoende mate beheerst, maar niet weet te slagen is een fout-negatief, en een student die de stof niet beheerst, maar wel weet te slagen is een fout-positief. De discussie of conjunctief dan wel compensatoir toetsen beter is, is eigenlijk een discussie over wat erger is, een fout-positief of een fout-negatief.

In de literatuur waarin compensatoir toetsen wordt gepropageerd worden vooral de fout-negatieven benadrukt. Er wordt gesteld dat de student bij elke toets of tentamen het risico loopt om ten onrechte te zakken en men maakt zich ongerust omdat daardoor mogelijk studievertraging ontstaat (Arnold, 2011, p. 32; Cohen-Schotanus, 1995, pp. 255-256; Instituut voor Psychologie, FSW, 2013, pp. 11-12). In

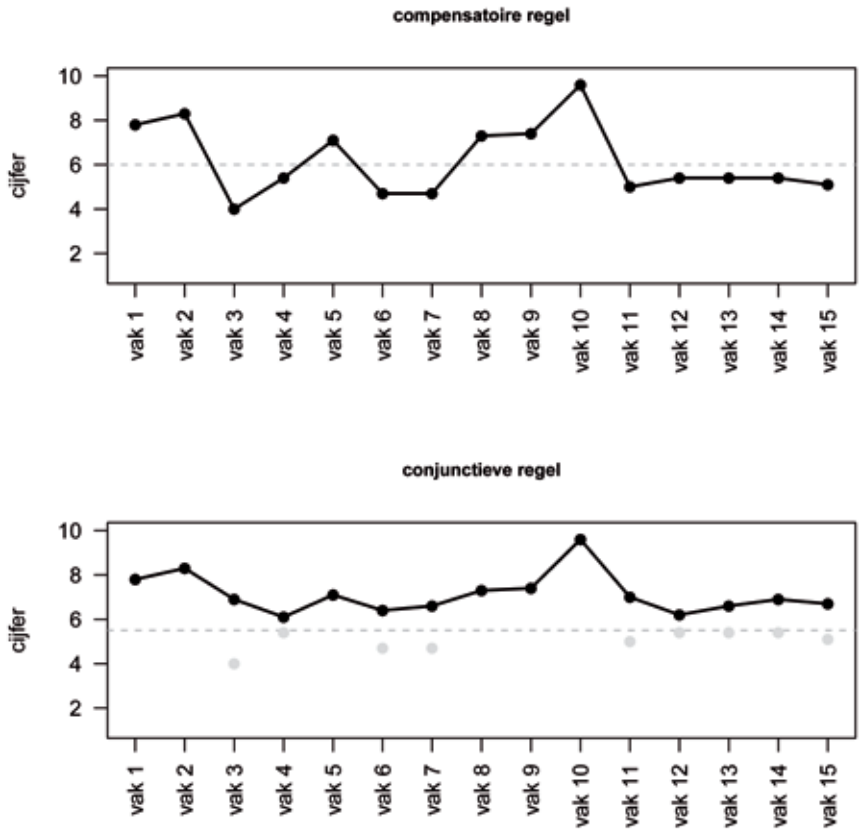
deze literatuur speelt de fout-positief daarentegen slechts een kleine rol en dat is om verschillende redenen opmerkelijk. Ten eerste heeft men in het hoger onderwijs heel duidelijk voor ogen waaraan een afgestudeerde student moet voldoen. Men gebruikt eindtermen die een beschrijving zijn van de minimaal noodzakelijke kwaliteiten van studenten op het gebied van kennis, inzicht, vaardigheden en attitudes die worden geoperationaliseerd in leerdoelen van de verschillende vakken waaruit de opleiding is opgebouwd (Van Berkel, Bax & Joosten-ten Brinke, 2014). Per vak wordt middels toetsen of tentamens bepaald of bij de student leerdoelen zijn bereikt. Vanuit het oogpunt van het hoger onderwijs is het voorkómen van fout-positieven dus heel belangrijk (zie ook, Chester, 2003, p. 40, Douglas & Mislevy, 2010, p. 302 en van Rijn et al., 2012, p. 119). Het hoger onderwijs draagt een maatschappelijke verantwoordelijkheid voor de kwaliteit van het diploma; bijvoorbeeld, toekomstige werkgevers of cliënten moeten ervan op aan kunnen dat de student voldoet aan de eindtermen (Academische Raad KU Leuven, 2009). Je zou van een onderwijsinstelling kunnen verwachten dat in het geval van een niet precieze beslisregel eerder op een strengere dan een minder strenge regel over wordt gestapt. Ten tweede wordt in veel leertheorieën (e.g., Schmidt, Rotgans & Yew, 2011) gesteld dat om goed te kunnen leren er eerst eerder verworven kennis moet worden geactiveerd. Om het leerproces in vervolgonderwijs soepel te laten verlopen is het dus van belang dat er voldoende voorkennis is en dus dat het aantal studenten dat ten onrechte slaagt tot een minimum wordt beperkt.

In de jaren 70 beschreef Van Naerssen (1970, p. 5) de beheersingsbeslissing als een speltheoretische situatie die wordt gekenmerkt door de aanwezigheid van twee groepen met aan de ene zijde de docenten en de andere kant de studenten. De student probeert de studietijd die nodig is om een diploma te halen te minimaliseren en een fout-negatief zal daarom veel erger worden gevonden dan een fout-positief. De docent probeert zo veel mogelijk relevante vaardigheden over te brengen en de leerdoelen te bewaken (van Naerssen, 1970); zij zal een fout-negatief als gelijkwaardig aan of als minder kwalijk dan een fout-positief evalueren (een vaardige student kan op een hertentamen alsnog laten zien dat hij de stof beheerst; een fout-positief kan zij nooit meer corrigeren). In het huidige onderwijsbestel is er echter een derde speler bijgekomen: de onderwijsmanager (zie Arnold & van den Brink, 2012). Deze speler heeft, als gevolg van prestatieafspraken met de overheid (zie Ministerie van Onderwijs, Cultuur en Wetenschap, 2011), als opdracht studenten zo efficiënt mogelijk door een onderwijssysteem te laten lopen. Elke student die een horde niet weet te nemen vormt een kostenpost op de balans. Het aantal studenten dat zakt moet dus geminimaliseerd worden en dus is het aannemelijk dat een fout-negatief zwaarder weegt dan een fout-positief.

Om kort te gaan, de vraag of compensatoir dan wel conjunctief toetsen beter is, en dus of een fout-negatief erger is dan een fout-positief, hangt sterk af van het perspectief van waaruit er naar de vraag wordt gekeken. Zowel voor de onderwijsmanager als studenten[3] is efficiency belangrijk, terwijl voor de docent het voldoen aan leerdoelen en academische integriteit belangrijk zijn. Waar de docent vroeger te maken had met slechts één andere speler, heeft hij tegenwoordig in het spel twee tegenspelers, waardoor het vasthouden aan eisen zeer waarschijnlijk lastiger is geworden (zie Tweede Kamer der Staten-Generaal, 2011).

5 Is een beslissing over een gemiddeld cijfer hetzelfde als de gemiddelde beslissing over cijfers?

Is het zinvol om beslissingen te nemen op basis van gemiddelden? In zijn beroemde boek 'The flaw of averages: why we underestimate risk in the face of uncertainty' laat Savage (2012) zien dat managers er vaak ten onrechte vanuit gaan dat een beslissing op basis van een gemiddelde van een reeks relevante getallen dezelfde is als de gemiddelde beslissing over de afzonderlijke getallen. Bij het gebruik van compensatoir toetsen (waarbij beslissingen op basis van een gemiddeld cijfer worden genomen) vindt mogelijk een vergelijkbare fout plaats. Figuur 2 illustreert het principe. Het



Figuur 2. Een voorbeeld van cijfers van dezelfde student, geslaagd onder twee beslisregels. In het bovenste geval hoeft hij geen herkansingen te doen omdat steeds minimaal een 4 is behaald en zijn gemiddelde cijfer een 6 is. In het onderste plaatje moet hij voor elk vak met een cijfer lager dan een 5.5 een herkansing doen tot minimaal een voldoende is behaald.

figuur laat de cijfers van eenzelfde student zien onder zowel een compensatoir (boven) en conjunctief (onder) toetsregime; in het geval van het conjunctieve regime moet hij vakken met onvoldoende cijfers herkansen totdat er een voldoende is behaald. In het geval van het compensatoire regime hoeft hij niet te herkansen omdat hij steeds hoger dan een 4 haalt en gemiddeld een 6. De beslissing op basis van dit gemiddelde zou 'slagen' zijn. Echter wanneer we kijken naar de beslissingen over de afzonderlijke cijfers, zien we dat hij in slechts zes van de vijftien vakken een voldoende haalt en de gemiddelde beslissing dus 40% geslaagd is. Ook hier is de beslissing op basis van het gemiddelde niet gelijk aan de gemiddelde beslissing. Als een onderwijsinstelling hecht aan het bewaken van een minimaal niveau voor elke getoetste vaardigheid, dan hebben beslissingen op basis van een gemiddeld cijfer mogelijk ongunstige gevolgen. Als de vakken in latere jaren bovendien voorkennis vereisen die door de compensatoire beslisregel niet is opgedaan, bestaat het risico dat deficiënties in eerdere jaren persisteren en zelfs tot probleemsituaties leiden waar de student zelfs niet meer in staat is de minimaal vereiste 4 te behalen.

Een meer formeel punt komt uit de besliskunde. Bij compensatoir toetsen wordt een beslissing genomen op basis van de cijfers voor twee of meerdere vakken. Dit type beslissing wordt een multiattributieve beslissing genoemd. Cruciaal bij zulke beslissingen

is hoe het nut van elk van de afzonderlijke attributen wordt gecombineerd tot één waardering. Omdat bij compensatoir toetsen niet gekeken wordt naar de individuele cijfers maar naar het gemiddelde cijfer, worden de cijfers geacht additief utiliteitsonafhankelijk te zijn (Keeney & Raiffa, 1976, p. 231). Dit houdt in dat bij de waardering voor het eindresultaat het cijfer op de ene toets niet afhangt van het cijfer op de andere toets. Bijvoorbeeld, de bijdrage van een 8 voor Vak 1 aan de waardering voor het diploma zou identiek moeten zijn als er voor Vak 2 een 4 of een 6 op staat. Uit besliskundig onderzoek is gebleken dat de aanname van additieve onafhankelijkheid in veel beslissingssituaties, denk aan het voorbeeld van de mobiele telefoons, een ongewenst resultaat oplevert, en het is ook zeer de vraag of deze aanname geldt voor onderwijsinstellingen die met hun diploma's de garantie willen afgeven dat de afgestudeerde de discipline in al haar onderdelen beheerst. Het is aannemelijk dat de instelling een diploma met een 8 voor Vak 1 alleen positief evalueert als er voor Vak 2 minimaal een 6 op het diploma staat; in het geval van een 4 op Vak 2 is er sprake van een hiaat in kennis of vaardigheid en heeft de 8 op Vak 1 geen enkele waarde.

6 Gedraagt een student zich anders bij een compensatoire dan bij een conjunctieve regel?

In het voorgaande is ingegaan op het feit dat studenten die onder het conjunctieve model een onvoldoende halen een tentamen zullen moeten overdoen. Onder het compensatoire model hoeft een student dat niet, zolang zijn onvoldoende niet te laag is en het gemiddelde cijfer voldoende hoog. Dit verschil in gedrag komt tot stand doordat verschillende beslisregels tot verschillende maatregelen leiden en is dus reactief. Er is echter nog niet gesproken over mogelijke verschillen in strategisch gedrag onder de twee toetsregimes. Van Naerssen (1970) had het al over een speltheoretische situatie en de onderwijseconomische literatuur laat ook zien dat studenten niet alleen reageren op behaalde- maar ook anticiperen op verwachte resultaten (e.g., Babcock, 2010; Bonesrønning, 2004; Idson & Clark, 1991), en hun studiegedrag dus aanpassen aan de beslisregel. Studenten hebben voor elk vak kosten per studiepunt (Wilbrink, 1979; Idson & Clark, 1991), en voor vakken die voor de betreffende student moeilijk zijn zullen die kosten hoger zijn dan voor makkelijker vakken. In het geval van een compensatoire regel lijkt het dan efficiënt om de meeste tijd te steken in een makkelijk vak dat kan worden gebruikt om een moeilijk vak te compenseren. Voorstanders van compensatoir toetsen (zie, bijvoorbeeld Cohen-Schotanus, 1995; Arnold, 2011) hebben onderzocht of het in de praktijk zo was dat studenten met name moeilijke vakken compenseerden met makkelijke vakken en concludeerden dat dat niet het geval was. Het probleem bij dit onderzoek is dat de analyses te eenvoudig waren om zulke conclusies te kunnen trekken. Vakken werden vergeleken op de frequentie van compensatie en studenten die wel en niet compenseerden werden onderling vergeleken op het gemiddelde cijfer op overige vakken. Zulke aggregatie kan allerlei onderliggende effecten maskeren. In de eerste plaats vertoont mogelijk slechts een klein deel van de studenten opportunistisch gedrag waardoor dat niet duidelijk in de gemiddelden terugkomt. In de tweede plaats kunnen deelgroepen studenten verschillen in voor welke vakken hoge dan wel lage cijfers worden gehaald en heffen deze verschillen elkaar mogelijk op als er naar de gehele groep wordt gekeken. Om strategisch gedrag te onderzoeken had er in het bijzonder gekeken moeten worden naar cijfers op specifieke vakken bij specifieke studenten door, bijvoorbeeld, een clusteranalyse op de patronen van cijfers uit te voeren.

Studenten verschillen namelijk in de kosten per studiepunt, afhankelijk van hun vaardigheid (Grant & Green, 2013). Daarnaast is het zeer aannemelijk dat studenten erin verschillen in welke vakken ze goed dan wel slecht zijn, en zullen ze verschillen in welk vak de laagste kosten per studiepunten heeft en welk vak wordt gebruikt om een vak met hogere kosten te compenseren. Als gevolg van studiedruk en het strategische gedrag dat wordt gestimuleerd onder de compensatoire

regel ligt het voor de hand dat studenten tijdens hun studie de nadruk leggen op de vakken waar ze relatief goed in zijn en minder in de duurdere, moeilijkere vakken. Aangezien studenten verschillen in hun vaardigheden bestaat het risico dat afgestudeerden onder een compensatoir systeem niet alleen meer hiaten vertonen in hun kennis en vaardigheden, maar ook nog eens verschillen in waar de hiaten zitten. Het gevolg is dat de betekenis van het diploma voor de maatschappij heterogener is dan onder een conjunctief systeem (Academische Raad, KU Leuven, 2009). De student heeft gemiddeld voldaan, maar er zijn mogelijk hiaten, en het is onduidelijk waar die hiaten precies zitten.

7 Besluit

In de onderhavige bijdrage is gekeken naar de argumenten voor en implicaties van compensatoir toetsen. Een belangrijk argument om over te stappen op compensatoir toetsen is de onbetrouwbaarheid van studietoetsen. Bij dit argument verwisselt men feitelijk de aard en de precisie van beslisregels. Overstappen op een andere beslisregel leidt namelijk tot kwalitatief andere beslissingen en stelt dientengevolge andere eisen aan de student. Net als conjunctief toetsen heeft compensatoir toetsen last van onbetrouwbaarheid; de regimes verschillen echter in hoe vaak de twee typen beslisfouten (fout-positieven en fout-negatieven) voorkomen. Bij compensatoir toetsen zijn er in vergelijking met conjunctief toetsen minder fout-negatieven, maar ook meer fout-positieven. Ook wordt gesteld dat beslissingen over gemiddelde cijfers riskant kunnen zijn en dat compensatoir toetsen vanuit besliskundig oogpunt een bedenkelijke aanname doet.

Zowel conjunctief als compensatoir toetsen heeft last van meetfouten, maar de voorkeur voor compensatoir toetsen is gekoppeld aan een preferentie van fout-positieven boven fout-negatieven. Het is de taak van docenten om kwaliteit te garanderen door fout-positieven zo veel mogelijk te voorkómen en zij zullen daarom neigen vast te houden aan het oude toetssysteem. Voor veel studenten en onderwijsmanagers, daarentegen, betekent een fout-negatieve toetsuitslag een afname in efficiency en lijkt het compensatoire toetssysteem een oplossing te bieden. Echter, het compensatoir systeem is kwetsbaar voor strategisch gedrag: studenten zouden er onder dit regime voor kunnen kiezen om zich te richten op de vakken waar ze (al) goed in zijn en niet op de vakken waarbij het wegwerken van hiaten moeite kost[4]. In de mate waarin studenten zich onder het compensatoire systeem strategisch gedragen zal de waarde van het diploma achteruit gaan.

Deze discussiebijdrage liet ook zien dat er meer en beter wetenschappelijk onderzoek naar conjunctief en compensatoir toetsen nodig is. Indien men valide uitspraken wil doen over hoe vaak ongewenst gedrag voorkomt in een compensatoir toetssysteem is het nodig dat onderzoekers betere onderzoeksmethoden gebruiken dan tot nog toe is gedaan. Beter dan vakken en studenten te vergelijken op gemiddelden is het om te kijken naar patronen van cijfers van specifieke studenten op specifieke vakken, bijvoorbeeld door clusteranalyse, of wellicht beter, een latente-trek latente-klassen analyse (zie bijvoorbeeld Maij-de Meij, Kelderman & van der Flier, 2008) te gebruiken.

Ook zijn er meer simulatiestudies nodig die beide beslisregels vergelijken. Het is zinvol om zulke studies op twee punten af te laten wijken van de hier besproken studies. Ten eerste dienen er criteria te worden gebruikt die onder de verschillende regels dezelfde betekenis hebben. Bijvoorbeeld, de verdeling van (ware) scores van geslaagden onder beide regimes, of één van de twee systemen als het 'juiste' behandelen en dan bestuderen hoe vaak het andere ermee overeenkomt. Ook zou het interessant zijn om te onderzoeken bij welk gemiddelde (een 7?) het compensatoire systeem even streng is als het conjunctieve. Ten tweede kan het simulatieonderzoek sterk verbeterd worden door niet aan te nemen dat studenten zich onder alle toetssystemen hetzelfde gedragen, maar juist per systeem verschillen in hun gedrag en dit expliciet te modelleren. Het werk van Wilbrink (1995) lijkt daarvoor een goed beginpunt.

Noten

1 Een andere reden om voor studietoetsen niet te streven naar hoge betrouwbaarheden in de klassieke zin is dat dit kan leiden tot een slechte dekking van het beoogde kennisdomein (zie bijvoorbeeld, Schuwirth & van der Vleuten, 2006).

2 Een paar jaar voordat Lord zijn artikel schreef, werd de besliskunde door Cronbach en Gleser (1957) in de psychometrie geïntroduceerd. Omdat Lord deze niet heeft geïntegreerd in zijn artikel is het theoretisch gezien wat achterhaald.

3 Uiteraard geldt dat niet voor alle studenten; er zijn ook studenten die willen dat de kwaliteit van hun diploma gewaarborgd is (zie, Bakker, 2012; Peters & Verhoeks, 2012).

4 De Quality Assurance Netherlands Universities (QANU) heeft soortgelijke zorgen geuit naar aanleiding van de visitatie van de opleiding Psychologie aan de EUR (2012): 'De commissie heeft met het opleidingsmanagement gesproken over het risico dat studenten in dit toetssysteem onvoldoendes laten staan voor struikelvakken zoals statistiek, waardoor zij op dat vlak onvoldoende vaardigheden opdoen om de eindtermen te realiseren'.

Literatuur

Academische Raad Katholieke Universiteit Leuven. (2009). *Advies van de owr d.d. 3 juli 2008 m.b.t. de invoering van een nultolerantiesysteem* (Rapport nr. AR 389 - CvB 787 - Doc.nr. I.5.2/I.6.1). Leuven: Dienst Onderwijsbeleid Katholieke Universiteit Leuven. Vertrouwelijk document.

Adriaens, H. (2010). *Het ontstaan en de implementatie van het leerkrediet in het Vlaamse hoger onderwijs*. Masterthese, Universiteit Antwerpen, Antwerpen.

Arnold, I. J. M. (2011). Compensatorische toetsing en kwaliteit. *Tijdschrift voor Hoger Onderwijs*, *29*(1), 31–40.

Arnold, I. J. M. & van den Brink, W. A. (2012, 1 februari). Onrust over diploma halen met onvoldoendes onterecht. *De Volkskrant*.

Babcock, P. (2010). Real costs of nominal grade inflation? New evidence from student course evaluations. *Economic Inquiry*, *48*(4), 983–996.

Bakker, M. (2012, 30 januari). Vijven, en toch een UvA-diploma. *De Volkskrant*.

Bonesrønning, H. (2004). Do the teachers' grading practices affect student achievement? *Education Economics*, *12*(2), 151–167.

Bregman, R. (2013, 12 juli). Plofstudenten. *De Volkskrant*.

Chester, M. D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice*, *22*(2), 32–41.

Cohen-Schotanus, J. (1995). De praktijk van de compensatie. *Onderzoek van Onderwijs*, *24*(4), 60-62.

Crocker, L. M. & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston.

Cronbach, L. J. & Gleser, G. C. (1957). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

De Gruijter, D. N. M. (2008). *Toetsing en toetsanalyse* (Rapport). Leiden: Universiteit Leiden. Verkregen 16 oktober 2014, op http://media.leidenuniv.nl/legacy/toetsing-en-toetsanalyse.pdf.

Douglas, K. M. & Mislevy, R. J. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics*, *35*(3), 280–306.

Grant, D. & Green, W. B. (2013). Grades as incentives. *Empirical Economics*, 1–30.

Idson, T. D. & Clark, J. R. (1991). Student time allocation and scholastic ability. *Journal of Applied Business Research*, *7*(3), 83–91.

Instituut voor Psychologie, FSW. (2013). *Bacheloropleiding Psychologie.* [Studiegids 2013/2014]. Rotterdam: Erasmus Universiteit Rotterdam.

Keeney, R. L. & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. New York: Wiley.

Lord, F. M. (1962). Cutting scores and errors of measurement. *Psychometrika*, *27*, 19–30.

Maij-de Meij, A. M., Kelderman, H. & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, *32*(8), 611–631.

Mellenbergh, G. J. (1993). Beslissen en beoordelen. In P. Koele & J. van der Pligt (red.), (hfdst. Beslissen met Tests en Studietoetsen). Amsterdam: Boom.

Ministerie van Onderwijs, Cultuur en Wetenschap. (2008). *Van Bijsterveldt verscherpt de exameneisen* [Nieuwsbericht 23 oktober 2008]. Verkregen 24 oktober 2014, op http://www.rijksoverheid.nl/nieuws/2008/10/23/ van-bijsterveldt-verscherpt-de-exameneisen.html.

Ministerie van Onderwijs, Cultuur en Wetenschap. (2011). *Universiteiten sluiten hoofdlijnenakkoord met staatssecretaris Zijlstra* [Nieuwsbericht 9 december 2011]. Verkregen 24 oktober 2014, op http://www.rijksoverheid.nl/ nieuws/2011/12/09/universiteiten-sluiten-hoofdlijnenakkoord- met-staatssecretaris-zijlstra.html.

Peters, E. & Verhoeks, J. (2012, 7 februari). Met compensatie onvoldoendes begint verschraling hoger onderwijs. *De Volkskrant*.

Quality Assurance Netherlands Universities. (2012). *Rapport over de bacheloropleiding psychologie en de masteropleiding psychologie van de Erasmus Universiteit Rotterdam* (Projectnummer: Q313). Utrecht.

Rekveld, I. J. & Starren, J. (1994). Een examenregeling zonder compensatie in het Nederlandse hoger onderwijs? Een vergelijking tussen compensatie en conjunctie. *Tijdschrift voor Hoger Onderwijs*, *12*(4), 210–219.

Savage, S. L. (2012). *The flaw of averages: why we underestimate risk in the face of uncertainty*. Hoboken NJ: Wiley.

Schmidt, H. G., Rotgans, J. I. & Yew, E. H. J. (2011). The process of problem-based learning: What works and why. *Medical Education*, *45*(8), 792–806.

Schuwirth, L. W. T. & van der Vleuten, C. P. M. (2006). A plea for new psychometric models in educational assessment. *Medical Education*, *40*(4), 296–300.

Task Force Studiesucces. (2009). *Studiesucces: Rapport van de task force studiesucces* (Interne publicatie). Leiden: Task force Studiesucces, Universiteit Leiden.

Tweede Kamer der Staten-Generaal. (2011). *Vragen van het lid Jasper van Dijk (SP) aan de staatssecretaris van Onderwijs, Cultuur en Wetenschap over studenten van InHolland die via een truc een diploma kregen (ingezonden 13 juli 2010).* (Vragen gesteld door de leden der Kamer, met de daarop door de regering gegeven antwoorden. Vergaderjaar 2009-2010. Aanhangsel van de Handelingen nr. 3231).

Van Berkel, H. van, Bax, A. & Joosten-ten Brinke, D. (2014). *Toetsen in het hoger onderwijs*. Houten: Bohn Stafleu Van Loghum.

Van den Brink, W. P. (1982). *Binomiale modellen in de testleer* [PhD thesis]. Amsterdam: Dissertatiereeks, Faculteit der Psychologie, Universiteit van Amsterdam.

Van den Brink, W. P. & Mellenbergh, G. J. (1998). *Testleer en testconstructie*. Amsterdam: Boom.

Van Lankveld, T. & Draaijer, S. (2010). *Compensatorisch toetsen* (Interne publicatie). Amsterdam: Onderwijscentrum Vrije Universiteit.

Van Naerssen, R. F. (1970). *Over optimaal studeren en tentamens combineren* [Oratie Universiteit van Amsterdam]. Amsterdam: Swets & Zeitlinger.

Van Rijn, P. W., Béguin, A. A. & Verstralen, H. H. F. M. (2009). Zakken of slagen? de nauwkeurigheid van examenuitslagen in het voortgezet onderwijs. *Pedagogische Studiën*, *86*(3), 185–195.

Van Rijn, P. W., Béguin, A. A. & Verstralen, H. H. F. M. (2012). Educational measurement issues and implications of high stakes decision making in final examinations in secondary education in the Netherlands. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 117–136.

Vlaamse Onderwijsraad, Raad Hoger Onderwijs. (2007). *De meerwaarde en gevolgen van de flexibilisering van het hoger onderwijs: Een verkenning*. Antwerpen: Garant.

Werkgroep Studiesucces. (2009). *Studiesucces aan de Universiteit van Amsterdam* (Interne publicatie). Amsterdam: Universitaire Commissie Onderwijs, Universiteit van Amsterdam.

Wilbrink, B. (1979). Examenproblematiek. In K. D. Thio & P. Weeda (red.), (hfdst. Universitaire examenregeling: conjunctief of compensatorisch). Den Haag: SVO.

Wilbrink, B. (1995). Studiestrategieën die voor studenten en docenten optimaal zijn: het sturen van investeringen in de studie. In H. P. M. Creemers (red.), *Onderwijsonderzoek in Nederland en Vlaanderen 1995: proceedings van de Onderwijs Research Dagen 1995 te Groningen* (pp. 218–220). Groningen: Gronings Instituut voor Onderzoek van Onderwijs, Opvoeding en Ontwikkeling, Rijksuniversiteit Groningen.

Abstract

A comparison of compensatory and conjunctive testing in higher education

This paper evaluates compensatory testing and the arguments for using it. Although the main stated reason for considering compensatory testing is the low reliability of test scores, it suffers from unreliability as well; the difference with the conjunctive approach is the type of classification error which prevails. Therefore it is advised to separate the nature and precision of decision rules. Compensatory testing makes decisions based on averages, which may be risky, and it makes a questionable assumption from a decision-theoretic point of view. Another problem that may occur is that false negatives may be prevented at the expense of false positives, by which learning objectives may not be met.

Auteurs

Niels Smits was ten tijde van het onderzoek werkzaam bij de afdeling Methoden, Faculteit der Psychologie en Pedagogiek van de Vrije Universiteit Amsterdam en is nu als universitair docent werkzaam bij Programmagroep Methoden en Technieken, Afdeling Pedagogiek, Onderwijskunde en Lerarenopleiding, Faculteit der Gedrags- en Maatschappijwetenschappen, Universiteit van Amsterdam. **Jan Hoeksma** is als universitair hoofddocent werkzaam bij de afdeling Methoden, Faculteit der Psychologie en Pedagogiek van de Vrije Universiteit Amsterdam. **Henk Kelderman** is werkzaam als emeritus hoogleraar psychometrie bij de Faculteit der Psychologie en Pedagogiek van de Vrije Universiteit Amsterdam en de afdeling Methodologie & Statistiek, Instituut Psychologie, Faculteit der Sociale Wetenschappen, Universiteit Leiden.