Geslaagd beoordelen

Invloeden op de classificatie accuraatheid van examensets

MARIANNE HUBREGTSE

Een uitgave van
KCH Examens

Geslaagd beoordelen

In het mbo worden deelnemers met grote regelmaat beoordeeld. Dit kan tot doel hebben om de tussentijdse resultaten in beeld te brengen. Vastgesteld wordt of de resultaten van de deelnemers al op het niveau van de exameneisen zijn. Op welke gebieden moet er nog wat extra's geleerd worden? Door regelmatig feedback te krijgen op zijn kennis, vaardigheden en houding 'groeit' de deelnemer naar zijn examen toe. Dit formatieve beoordelingstraject beïnvloedt het opleidingstraject.

Bij examineren gaat het om een eindoordeel. De examinatoren maken hierbij gebruik van de exameninstrumenten. Na het afnemen van het examen blijkt of de deelnemer een bekwaam beginnend beroepsbeoefenaar is zoals dat is vastgelegd in het kwalificatiedossier. Het beoordelen van de deelnemer is een complex proces met ingrijpende gevolgen. De deelnemer wordt al dan niet bekwaam bevonden en ontvangt al dan niet zijn diploma.

Helaas is het complexe proces van beoordelen nooit 100% objectief. Zowel tijdens het formatieve als tijdens het summatieve traject zullen er meet- en beoordelingsfouten worden gemaakt. Maar hoe kunnen beoordelaars ervoor zorgen dat zij een zo juist mogelijk oordeel geven over de bekwaamheid van een deelnemer?

Deze publicatie is het resultaat van een deel van het promotieonderzoek van Marianne Hubregtse. Het onderzoek is gericht op de objectiviteit van de examinering. In haar verslag beschrijft ze de effecten van het werken met meerdere examens om te komen tot een diplomabeslissing. Het is een onderzoek dat bijdraagt aan een geslaagde beoordeling.

Kenniscentrum Handel maakt zich sterk voor goede beoordelingsinstrumenten en voor de scholing van gekwalificeerde beoordelaars. Het aantrekken van een promovenda die haar promotieonderzoek richt op de objectiviteit van de examinering draagt voor ons bij aan de continue kwaliteitsslag die gemaakt wordt in het examenservicesysteem (ESS). KCH Examens verwerkt de resultaten van het onderzoek in haar examenproducten.

Voor degenen die een verdieping zoeken naar aanleiding van deze publicatie is een uitgebreidere versie van dit onderzoek beschikbaar via www.kchexamens.nl. Op deze site vind u een link naar het onderzoek.

Kenniscentrum Handel
Marijke Backx
manager KCH Examens

Inhoud

Meten van examenkwaliteit

De kwaliteit van examens is belangrijk. Daar is iedereen het over eens. Dit geldt zeker voor summatieve examens, waar vaak veel van afhangt. Voor de kwaliteit van theorie-examens wordt gekeken naar de validiteit en betrouwbaarheid. De kwaliteit van praktijkexamens wordt meestal bekeken in termen van authenticiteit (bijvoorbeeld Gulikers, 2004) en de (inhouds)validiteit (bijvoorbeeld Linn, Baker & Dunbar, 1991). In tegenstelling tot theorie-examens, zijn praktijkexamens meestal geen gestandaardiseerde toetsen. Het is daarom niet mogelijk om de traditionele opvatting van betrouwbaarheid toe te passen op de praktijkexamens (Clauser, 2000; Dochy, 2009). Bovendien wordt een belangrijke summatieve beslissing nooit genomen op basis van één examen. Het is dus niet mogelijk om te kijken naar traditionele maten voor kwaliteit. Maar waarmee kunnen we dan wél aan de slag?

EXAMENSET

Diplomabeslissingen worden genomen aan de hand van een methodemix van examens. Voor bijvoorbeeld Verkoopspecialist Detailhandel bestaat de methodemix uit vijf theorie-examens, vier praktijkexamens en een werkstuk. Op basis van deze tien examens krijgt een student wel of geen diploma. Het is handig om met een enkel woord over al die examens te praten. Daarvoor gebruiken we het woord examenset. De term examenset omvat elke willekeurige samenstelling van minimaal twee examens. De examens in een examenset kunnen theorie-examens, praktijkexamens of een mix daarvan zijn.

CLASSIFICATIE ACCURAATHEID

Classificatie accuraatheid zegt niets anders dan: delen we de kandidaten in de juiste categorieën in? Twee belangrijke categorieën zijn bijvoorbeeld: verdient een diploma en verdient geen diploma. Een meer formele definitie van classificatie accuraatheid is: 'de mate van overlap tussen een beslissing gebaseerd op de geobserveerde score en de beslissing die zou zijn genomen op basis van een score zonder enige meetfout' (Hambleton & Novick, 1973).

Er bestaat alleen geen meting, en dus geen examen, zonder meetfout.

Er bestaat alleen geen meting, en dus geen examen, zonder meetfout. Daardoor zullen misclassificaties altijd voorkomen: kandidaten worden in een onjuiste categorie ingedeeld. Er zullen dus altijd valse negatieven zijn: studenten die zakken voor een examen, terwijl ze voldoende competentie bezitten. Maar ook zijn er altijd valse positieven: studenten die een examen halen, terwijl ze eigenlijk niet voldoende competentie bezitten. Dit noemen we misclassificaties. Deze misclassificaties zijn overal te vinden waar de scores van een examen afwijken van de scores van een examen zonder meetfout. Figuur 1 verduidelijkt dit. De twee vakjes 'goede diplomabeslissing' tellen samen op tot een percentage van alle ingedeelde studenten. Dit percentage is de classificatie accuraatheid.

Figuur 1 – Misclassificaties

	Student krijgt diploma	Student krijgt geen diploma
Student is competent	juiste diplomabeslissing	misclassificatie: vals negatief
Student is incompetent	misclassificatie: vals positief	juiste diplomabeslissing

KNIKKERS SORTEREN

In essentie is de beslissing die we in het mbo willen nemen dichotoom: een student verdient een diploma of een student verdient geen diploma. In statistische termen klinkt dat als: 'een student valt boven of onder een afgesproken grens'. Examens zijn daarom classificatie instrumenten. Men gebruikt examens om te bepalen of een student boven of onder de grens geplaatst moet worden.

Je kunt dit vergelijken met een sorteermachine voor knikkers. Knikkers staan hier voor studenten en jij bent de assessor, die cijfers uitdeelt. Stel dat je een grote bak met knikkers hebt. De knikkers hebben veel verschillende maten en je wilt alle knikkers met een doorsnede van halve centimeter of kleiner rood verven. Alle knikkers die groter zijn, wil je blauw verven. Je moet de knikkers dan dus sorteren. Je laat de knikkers dan een 'examen' afleggen: je meet ze of je schat in of ze groter of kleiner dan een halve centimeter zijn (dit is vergelijkbaar met de assessor die studenten beoordeeld.) Alle kleine knikkers horen na je examen in het rode bakje te liggen en alle grote knikkers in het blauwe bakje. Omdat je examen een zekere meetfout bevat, ligt een aantal knikkers in het verkeerde bakje. Als het belangrijk is dat er weinig knikkers in de verkeerde bakjes liggen, kun je de meting een aantal keer uitvoeren. Je bekijkt dan voor elke knikker aan het eind hoe vaak er 'rood' en hoe vaak er 'blauw' uit het examen kwam. Je kiest dan voor de bak (kleur) die het vaakst werd aangeraden.

Omdat kennis, vaardigheden en competenties een stuk moeilijker te meten zijn dan de doorsnede van een knikker, ligt het voor de hand dat daar meetfouten worden gemaakt. Daarom wordt een diplomabeslissing gebaseerd op een examenset en niet op een enkel examen. Het classificatieprobleem wordt hierdoor anders. De classificatie die van belang is, ligt dan niet meer bij een enkel examen, maar bij de examenset. Het is daarom interessant om de classificatie accuraatheid van de examenset te bekijken. Dit gebeurt eigenlijk nooit, omdat we zijn gefocust op de losse examens. Het belang van de examenset als geheel wordt niet altijd gezien.

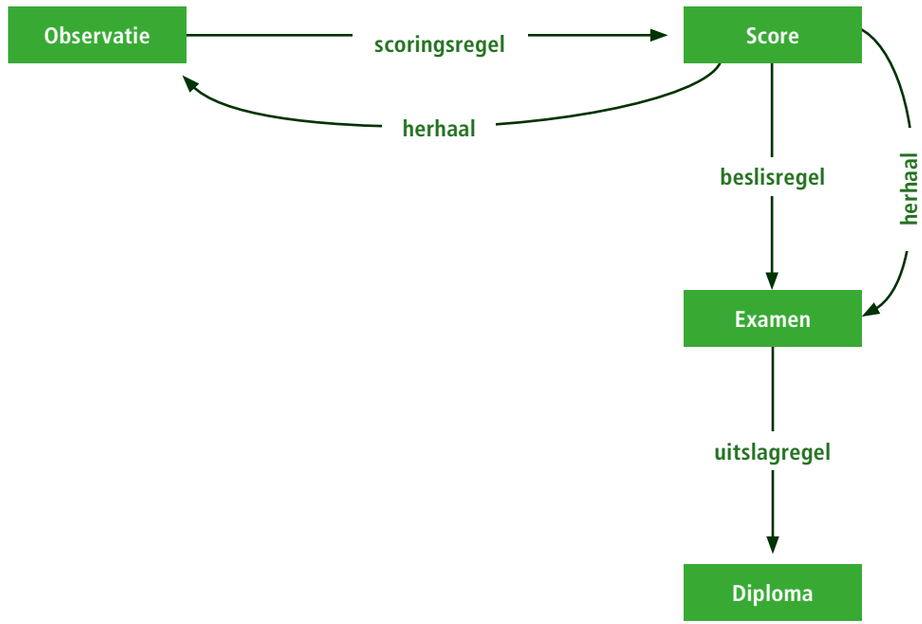
DE UITSLAGREGEL

In figuur 2 is te zien met welke drie regels een examenmaker te maken heeft. Een examenset leidt in de figuur tot een diploma. De figuur begint met een aantal observaties. Observaties kunnen zowel beantwoorde vragen als praktijkobservaties zijn. Op basis van losse observaties nemen we nooit direct een diplomabeslissing. Je kunt het vergelijken met een portfolio waaraan nog geen punten zijn toegekend. Er zijn misschien los beantwoorde vragen en videofragmenten van praktijktoetsen.

Een aantal observaties vatten we daarom samen in een score. Een **scoringsregel** beschrijft hoe je van een specifieke observatie tot een score komt. Voor praktijktoetsen zijn dit criteria, die de beoordelaar kan vergelijken met de observatie (voor theorietoetsen is de scoringsregel vastgelegd in het scoringsvoorschrift). Deze stap wordt zo vaak herhaald als er observaties in een examenset zitten. Vervolgens heb je een ruwe score. Vanuit een score wordt bepaald of een student een examen heeft gehaald of niet. De **beslisregel** specificeert bij welke score je welk cijfer of oordeel toekent, en dus bij welke score de student een voldoende krijgt. Deze stap wordt herhaald voor elk examen in een examenset.

Er is nog een laatste regel in het schema. Dit is een regel die altijd vastligt, maar waarmee men veel minder vaak te maken heeft. De **uitslagregel** bepaalt of iemand op basis van de resultaten van alle losse examens in de examenset een diploma krijgt of niet. De regel beschrijft dus hoe je van alle examens naar een diplomabeslissing gaat. Deze laatste regel wordt soms 'vergeten' en alleen impliciet gebruikt.

Figuur 2 – Schema van hoe observaties tot een diplomabeslissing leiden



Onderzoek

Om dit onderzoek te kunnen doen, heeft een aantal onderwijsinstellingen[1] dataverzameling mogelijk gemaakt. Er zijn data verzameld van de praktijktoetsen van de opleidingen Verkoopspecialist, Eerste Verkoper en Manager Handel. Deze opleidingen hebben voor een deel overlap in de praktijkbeoordelingen. Alleen de overlappende delen zijn gebruikt. De examenset bestaat uit vier praktijkexamens voor de vier kerntaken van de Verkoopspecialist.

De examenset bestaat uit vier praktijkexamens voor de vier kerntaken van de Verkoopspecialist.

DATAVERZAMELING

De examenset waarvan data zijn verzameld, bestaat uit vier praktijkexamens.

- Kerntaak 1 – Verzorgt de ontvangst en verwerking van goederen
- Kerntaak 2 – Verkoopt, adviseert en verleent service
- Kerntaak 3 – Handelt verkooptransacties af en/of leidt deze
- Kerntaak 4 – Optimaliseert verkoop en assortiment

Elk praktijkexamen bestaat weer uit een set observaties van beoordelingscriteria. Deze criteria worden gebruikt als de items van het rekenmodel waarop de analyse is gebaseerd[2]. Figuur 3 laat een voorbeeld van een deel van een van de praktijkexamens zien.

Figuur 3 – Voorbeeld van observatiecriteria van het praktijkexamen(vormen samen één item)s

	Omcirkel per competentie uw oordeel:	
Competentie: Kwaliteit leveren	onvoldoende	0
	voldoende	1
De deelnemer ...	goed	2
• controleert zorgvuldig en systematisch de voorraad.	zeer goed	3
• controleert zorgvuldig en systematisch de bijbehorende gegevens.		
• signaleert afwijkingen op tijd.		
• plaatst de bestelling zorgvuldig.		
• zorgt ervoor dat er geen verschil is tussen de ingeschatte artikelen die nodig zijn en de daadwerkelijk bestelde artikelen.		

1 Grote dank gaat uit naar het ROC Nijmegen, het ROC van Amsterdam, ROC Friese Poort, Albeda College en ROC Landstede. Zonder data is er tenslotte geen onderzoek.

2 Voor de uitgebreide uitleg van dataverzameling en analyse verwijzen we de lezer naar het volledige onderzoeksverslag op http://www.kch.nl/

Cesuren

Met cesuur wordt de beslisregel uit figuur 2 bedoeld. Het belangrijkste aspect daarbij is de grens tussen onvoldoende en voldoende voor losse examens. De beslissing is in dit geval of een student aan de onvoldoende of voldoende kant van de grens valt. Dit lijkt op de beslissing voor een examenset, maar dan voor slechts één examen. De cesuur voor de examens is ook van invloed op de classificatie accuraatheid van de examenset.

NORM-GEREFEREERDE EN CRITERIUM-GEREFEREERDE CESUREN

De cesuren zijn een van tevoren vastgestelde grens tussen voldoende en onvoldoende voor een bepaald examen. Er bestaan twee soorten cesuren: een norm-gerefereerde cesuur en een criterium-gerefereerde cesuur. Een norm-gerefereerde cesuur bepaalt hoeveel procent van de studenten moet slagen. Een criterium-gerefereerde cesuur bepaalt hoeveel procent van de vragen de studenten goed moeten hebben om te slagen.

In het mbo wordt normaal gesproken gewerkt met criterium-gerefereerde cesuren. We willen graag dat studenten een bepaald deel van de stof beheersen. Over de stof maken we met toetsmatrijzen een verdeling. Een examen gaat dan voor een vast percentage over bepaalde onderdelen van de stof en het aantal vragen ligt meestal ook vast. De cesuur vertelt ons dan hoeveel vragen van het theorie-examen voldoende gemaakt moeten worden. Voor praktijkexamens vertelt de cesuur ons hoeveel criteria voldoende beoordeeld moeten zijn om het examen af te leggen met een voldoende. De huidige cesuur ligt bij 60% van de leerstof (er slaagt ongeveer 80% van de studenten bij deze cesuur).

REKENTOETSEN

Als een cesuur erg extreem is, zijn er geen misclassificaties (Lee, 2008). Met erg extreem wordt bedoeld dat de cesuur buiten het bereik van de vaardigheid van de doelpopulatie ligt. Bijvoorbeeld; stel je voor dat je de rekenvaardigheid van vierdejaars mbo-ers wil meten. Je gebruikt daar een rekentoets voor die eigenlijk voor kleuters was gemaakt. De mbo-ers zijn de doelpopulatie. Als je de cesuur voor de kleuters aanhoudt, dan verwacht je dat alle mbo-ers een voldoende krijgen. Er zouden geen mbo-ers moeten zijn die nog op kleuterniveau rekenen. In dat geval zijn er dus geen misclassificaties.

Ook andersom zijn die er niet. Stel dat je wilt weten hoe goed kleuters zijn in het samenstellen van een bedrijfsplan. Je gebruikt hiervoor een praktijkexamen voor vestigingsmanager groothandel. Dat is natuurlijk een te moeilijk examen voor de kleuters. Je verwacht dus dat ze allemaal een onvoldoende halen. De vaardigheid ligt ver onder de cesuur. Ook hier is er 100% classificatie accuraatheid. Alle kleuters zijn ingedeeld aan de juiste kant van de cesuur.

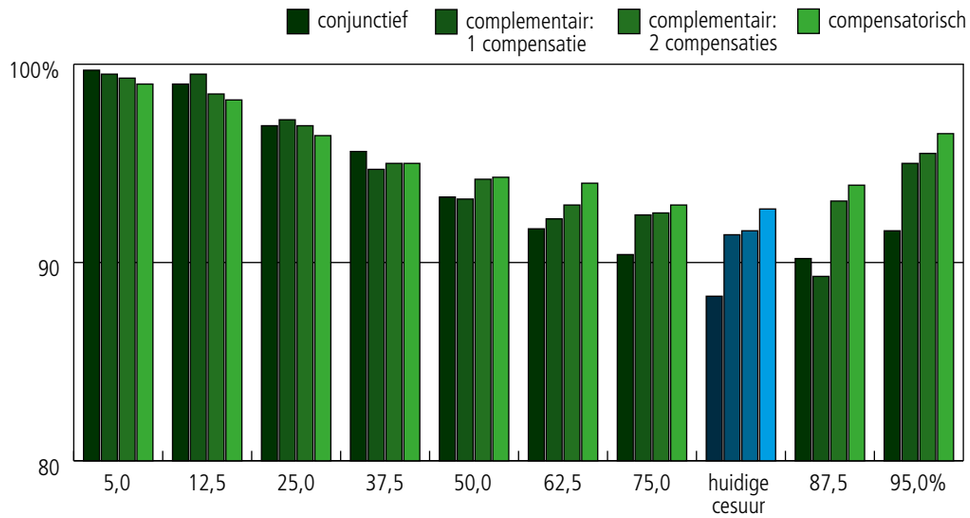
VAARDIGHEID VAN DE STUDENT

Studenten die een vaardigheid hebben die ver van de cesuur af ligt, zijn gemakkelijker aan de goede kant van de cesuur te plaatsen dan studenten die een vaardigheid hebben die dicht bij de cesuur ligt (Martineau, 2007). Van sommige studenten kun je goed zien dat ze een bepaalde competentie nog niet beheersen. Duidelijk is dat zij nog onvoldoende competent zijn. Als je zo'n student examineert, is dat gemakkelijk en er zal niet snel een foute beslissing worden genomen. Net als bij de kleuters is er geen twijfel mogelijk dat ze onvoldoende

competentie bezitten. Ook als een student juist uitblinkt in een bepaalde competentie, is het gemakkelijk om een praktijkexamen in te vullen en is het onwaarschijnlijk dat de student een onvoldoende krijgt. Dat lijkt op de mbo-ers die de rekentoets voor kleuters doen.

Als een student een vaardigheid heeft die dicht bij de cesuur ligt, wordt het moeilijker om hem te beoordelen. Een beoordelaar moet voor een juiste beoordeling dan ook vertrouwen op de omschrijvingen van de criteria in het praktijkexamen. Het is moeilijk om van tevoren in te schatten wat het oordeel zal zijn. De student hangt op net een onvoldoende of net een voldoende. Bij zo'n student is het gemakkelijker voor te stellen dat er een verkeerde beslissing genomen wordt. Een enkel criterium dat anders wordt beoordeeld, kan al leiden tot een andere beslissing.

Figuur 4 – Resultaten gerangschikt naar opgelegde cesuur



Let op dat de grafiek begint bij 80%, en niet bij 0%. De blauwe balkjes geven de classificatie accuraatheid weer voor de cesuur die KCH Examens op dit moment hanteert. De cesuur is in absolute termen 60% van de leerstof. In de doelpopulatie slaagt bijna 80% bij deze cesuur voor een diploma. De termen conjunctief, complementair en compensatorisch worden in het volgende hoofdstuk uitgelegd.

EFFECT VAN CESUUR OP CLASSIFICATIE ACCURAATHEID

Figuur 4 laat de classificatie accuraatheid zien van de hele examenset. De cesuren voor de losse examens zijn steeds aan elkaar gelijk. Het is te zien dat we de hoogste classificatie accuraatheid vinden bij de extremen, waar of iedereen, of juist niemand zijn diploma haalt. We hadden dit ook al voorzien. Tenslotte maak je geen classificatie fouten met een examenset die ver boven, of juist ver onder, de vaardigheid van de doelpopulatie ligt. De voorbeelden zijn een kleutertoets aan mbo-ers en een mbo-toets aan kleuters. We verwachtten al dat bij de hele lage cesuren (waar maar 5% een diploma krijgt) en hele hoge cesuren (waar 95% een diploma krijgt) de classificatie accuraatheid hoog is.

We vinden de hoogste classificatie accuraatheid bij de extremen, waar of iedereen, of juist niemand zijn diploma haalt.

De cesuur die KCH Examens op dit moment hanteert is in figuur 4 blauw gekleurd. Bij deze cesuur (60% van de leerstof) haalt ongeveer 80% van de leerlingen zijn diploma. Het valt direct op dat rond die cesuur ook de laagste classificatie accuraatheid wordt gezien. We verwachtten dit, omdat studenten die een vaardigheid dicht bij de cesuur hebben moeilijker op de goede plek te plaatsen zijn. Dat de cesuur en de dip samenvallen kan twee achterliggende oorzaken hebben. Doordat het goed bekend is wat er van de studenten wordt verwacht, kan het zo zijn dat studenten naar de cesuur toe leren. Dit kan zowel vanuit de studenten gebeuren, maar ook onderwijsinstellingen en leerbedrijven kunnen studenten stimuleren om dit niveau te halen. Het kan ook zo zijn dat de examensets goed passen bij wat de studenten van nature aan vaardigheden en kennis opdoen tijdens hun studie.

Zolang de cesuur van de examensets het gewenste niveau weerspiegelt, hoeft dit geen probleem te zijn. Sterker nog, aangezien er zeker in het mbo sprake is van opbrengstgericht werken (in het Engels ook wel 'teaching to the test') genoemd, is dit een gewenst effect. Het brengt echter de vraag met zich mee hoeveel misclassificaties we dan maken en of dat aanvaardbaar is of niet. Mocht het aantal misclassificaties ontoelaatbaar hoog zijn, dan zou naar maatregelen gekeken kunnen worden. Dit kan bijvoorbeeld inhouden dat er wordt ingezet op een hoger niveau van de studenten. Om een meer accurate meting te krijgen, zouden onderwijsinstellingen dus kunnen proberen om studenten veel hoger op te leiden dan het doelniveau.

Een zeer extreme cesuur, waarbij bijvoorbeeld iedereen slaagt, is meestal niet wenselijk in het onderwijs, omdat dit een erg hoog niveau van de studenten vereist, of een onredelijk lage cesuur. Het eerste lijkt op het geven van mbo examens (en diploma's) aan hbo-ers, het tweede is in de trant van kleutertoetsen als mbo afsluiting. Beide gevallen doen geen recht aan de studenten. Een andere oplossing zou kunnen zijn om niemand meer een diploma te geven. Dat zou analoog zijn met het geven van mbo examens aan kleuters. Ook dit levert niet het gewenste resultaat. We willen immers adequaat opgeleide beginnende beroepsbeoefenaren.

Uitslagregels

Omdat we de cesuur maar tot op beperkte hoogte kunnen veranderen, is er in het onderzoek ook gekeken naar uitslagregels voor de examenset. Er is daar al eerder naar gekeken. Dat betrof de verschillende mogelijkheden voor compensatie binnen de centrale examens van het middelbaar onderwijs (Van Rijn, Béguin & Verstralen, 2009; Verstralen, 2009a). Uit deze studies blijkt dat classificatie accuraatheid wordt beïnvloed door de gebruikte uitslagregel. Hoe meer mogelijkheid tot compenseren, hoe hoger de classificatie accuraatheid. Het is de verwachting dat dit altijd geldt. Als het zo is, dan zouden we het effect moeten terugzien, ongeacht welke cesuur wordt gehanteerd. Ergens is het ook logisch dat het zo werkt. Wat er met compensatie namelijk gebeurt, is dat je de examens min of meer samenneemt en gaat zien als één groot examen. De betrouwbaarheid van de examenset gaat dan omhoog, omdat de toetslengte van de examenset omhoog gaat (Gatti & Buckendahl, 2006). Een betrouwbaarder examen meet preciezer, waardoor je minder misclassificaties maakt en dus een hogere classificatie accuraatheid bereikt.

Een langere examenset meet preciezer, waardoor je minder misclassificaties maakt en dus een hogere classificatie accuraatheid bereikt.

TYPEN UITSLAGREGELS

De uitslagregel specificeert hoe de resultaten op de verschillende examens gecombineerd moeten worden tot één beslissing voor de hele examenset. Deze regels schrijven voor hoeveel compensatie er mogelijk is. Dat wil zeggen: of en hoe een student een onvoldoende voor een examen kan ophalen met een ander examen. Er zijn drie 'smaken' van uitslagregels: conjunctieve, complementaire en compensatorische uitslagregels (zie ook Van Rijn et al., 2009). **Conjunctieve regels:** je mag nooit compenseren. Studenten moeten voor elk examen in de examenset een voldoende halen. **Complementaire regels:** er is enige mate van compensatie mogelijk, maar er is precies omschreven hoeveel compensatie nog toelaatbaar is en bij hoeveel onvoldoendes de student geen diploma meer krijgt. **Compensatorische regels:** studenten moeten voor alle examens in een examenset gemiddeld een voldoende halen.

GEBRUIK UITSLAGREGELS

Conjunctieve en complementaire regels worden vaak gebruikt als er een minimumniveau van de studenten wordt verwacht. Compensatorische regels worden gebruikt op het moment dat men kan stellen dat het redelijk is om een bepaalde competentie met een andere te compenseren. Er zijn veel verschillende compensatorische regels denkbaar, afhankelijk van het aantal examens in de examenset en de bereidheid om compensatie toe te laten (Hambleton, Jaeger, Plake & Mills, 2000).

In het onderzoek werden alle drie de soorten regels gebruikt, oplopend van weinig naar veel compensatiemogelijkheid. Op deze manier kunnen we zien hoe compenseren tussen examens binnen een examenset, de classificatie accuraatheid beïnvloedt. In totaal werden vier verschillende uitslagregels met elkaar vergeleken: één conjunctieve regel, twee complementaire regels en één compensatorische regel. Geen compensatie betekent dat voor elk examen een voldoende moet worden gehaald, maar er hoeven geen perfecte scores te worden behaald. Bij volledige compensatie hoeft een student alleen gemiddeld voor alle examens een voldoende te halen. Daartussenin mochten studenten ofwel één, ofwel twee punten compenseren.

Hieronder staan de vier omschrijvingen van de uitslagregels verder uitgelegd. Een student doet een examen voor elk van de vier kerntaken[3]. Voor elk examen kan hij minimaal 0 en maximaal 5 punten halen. De cesuur voor alle examens is op 3 punten gelegd. Heeft een student 3 punten, dan heeft hij een voldoende voor het examen. Heeft een student 2 punten, dan heeft hij een onvoldoende. Per uitslagregel wordt er uitgelegd of er iets gecompenseerd mag worden en hoeveel er gecompenseerd mag worden. De voorbeelden zijn denkbeeldige studenten die op basis van hun vier cijfers (voor de vier examens) wel of geen diploma zouden krijgen.

Conjunctieve uitslagregel – geen compensatie

Voorbeeld wel een diploma

Voorbeeld geen diploma

De student moet voor elk examen minimaal 3 punten halen, en dus minimaal 12 punten voor de examenset. Als hij voor één examen of meerdere examens een onvoldoende haalt, krijgt hij geen diploma (zelfs niet bij meer dan 12 punten totaal).

3 zie onder kopje 'Onderzoek'

Complementaire regel 1 – 1 compensatie mogelijk

Voorbeeld wel een diploma

Voorbeeld geen diploma

De student moet in totaal minimaal 12 punten halen. Er mag één 2 gecompenseerd worden. Daar staat minimaal één 4 of 5 tegenover.

Complementaire regel 2 – 2 compensaties mogelijk

Voorbeeld wel een diploma

Voorbeeld geen diploma

De student moet in totaal minimaal 12 punten halen. Er mag één 1 of twee 2-en worden gecompenseerd. Daar staat dan minimaal één 5 of twee 4-en tegenover.

Compensatorische uitslagregel – gemiddeld voldoende

Voorbeeld wel een diploma

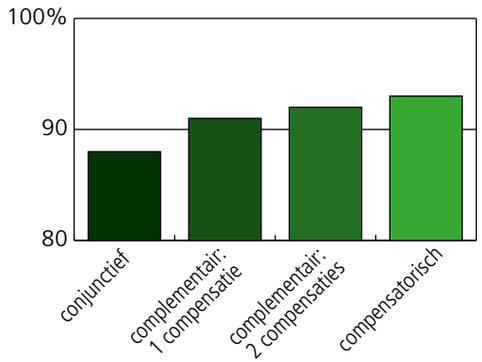
Voorbeeld geen diploma

De student moet in totaal gemiddeld 12 punten halen. Ook een 0 mag worden gecompenseerd.

EFFECT VAN UITSLAGREGEL OP CLASSIFICATIE ACCURAATHEID

Om te kijken naar de invloed van de uitslagregel, is het interessant om in te zoomen op de huidige cesuur. Dan kunnen we zien wat de kleinste maatregel, namelijk aanpassen van het compensatiebeleid, doet met de classificatie accuraatheid. Figuur 5 laat dit zien. Het ligt voor de hand om naar de huidige cesuur te kijken, omdat er opbrengstgericht wordt gewerkt.

Figuur 5 – Resultaten voor uitslagregels bij de huidige cesuur



De condities zijn van links naar rechts gerangschikt van geen compensatie naar volledige compensatie. De balkjes geven de classificatie accuraatheid weer voor de cesuur die KCH Examens op dit moment hanteert. De cesuur is in absolute termen 60% van de leerstof. In de doelpopulatie slaagt bijna 80% bij deze cesuur voor een diploma.

Als we kijken naar figuur 5, kunnen we goed zien dat de classificatie accuraatheid voor alle condities behoorlijk hoog is. In het meest ongunstige geval (geen compensatie tussen examens) is de classificatie accuraatheid 88%. Dat betekent dat zo'n 12% van de studenten onterecht slaagt of onterecht zakt. Dit kan nog worden verminderd met 5%. In de meest compensatorische conditie, stijgt de classificatie accuraatheid naar 93%. Hier krijgt dus nog maar 7% van de studenten onterecht een diploma of onterecht geen diploma. Ook interessant is dat de grootste winst wordt geboekt van 'geen enkele compensatie' naar 'enige compensatie' (verschil van 91% - 88%= 3%). Dit is geen verrassend resultaat omdat compensatie in essentie de examens verlengt en dus betrouwbaarder meten mogelijk maakt. Door de hogere betrouwbaarheid is de examenset beter in staat studenten in het goede 'bakje' te leggen.

Dit is geen verrassend resultaat omdat compensatie in essentie de examens verlengt en dus betrouwbaarder meten mogelijk maakt.

Dit zou een reden kunnen zijn om meer compensatie tussen examens toe te staan. Het is hierbij de vraag wat er voorrang moet krijgen: een minimum niveau van de studenten per examen, of meer studenten die juist worden geclassificeerd. Dat laatste houdt in dat er zowel meer studenten zullen zijn die terecht een diploma krijgen, als dat er minder studenten zijn die onterecht een diploma krijgen. Zeker gezien het aantal studenten, kan het de moeite waard zijn om, in ieder geval deels, compensatie toe te staan binnen één examenset. Echter, examenmakers moeten goed afstemmen met de onderwijsinstellingen en de beroepspraktijk of het wenselijk is dat bepaalde onderdelen onvoldoende kunnen worden gemaakt. Inhoudelijke argumenten (en dus inhoudsvaliditeit) wegen meestal zwaarder dan classificatie accuraatheid.

Invloeden op classificatie accuraatheid

De resultaten laten zien dat classificatie accuraatheid van examensets wordt beïnvloed door de cesuren van de losse examens en de gehanteerde uitslagregel. Deze drie invloeden werken hetzelfde voor losse examens als voor de examenset. Dat is goed nieuws, want dat betekent dat we een groot aantal aanbevelingen direct kunnen overnemen. Uiteraard blijft de kwaliteit van de examenset altijd nauw verbonden met de kwaliteit en de hoeveelheid van examens in de examenset.

EXTREME CESUREN

Het blijkt verder dat extreme cesuren (waarbij bijna geen enkele student een diploma krijgt, of juist bijna iedereen) de classificatie accuraatheid sterk verhogen. In de praktijk is dit niet realistisch. Hooguit als er een kleine groep studenten geselecteerd wordt, bijvoorbeeld de top 10 studenten, is het interessant om met zo'n cesuur te werken. In het onderwijs is een vaststaande criteriumgerichte cesuur de meest voorkomende. Het zou erg veel werk, en erg inefficiënt zijn om examensets te maken die geheel buiten de bedoelde vaardigheid nog goed kunnen meten.

LATEN COMPENSEREN TUSSEN EXAMENS...

Verder laat dit onderzoek zien dat compenseren tussen de verschillende examens in de examenset de classificatie accuraatheid verhoogt. Compenseren, zowel binnen één examen, als binnen de hele examenset, is een factor waar examenmakers of onderwijsinstellingen altijd invloed op hebben. Er zijn verschillende redenen waarom je wel of geen compensatie zou willen toelaten. Een reden om compensatie toe te staan is het verhogen van de classificatie accuraatheid. Daarnaast lijkt het redelijk om de meetfout in de relatief korte examens die we in het mbo tegenkomen te verminderen door compensatie toe te staan. Bovendien kan het oneerlijk lijken om een zeer goede student zijn of haar diploma niet te laten halen, omdat er één examen niet goed ging. Een examenset meet over het algemeen accurater dan losse examens. Focussen op de examenset betekent dat de vraag hoe vaardig studenten exact zijn ondergeschikt raakt aan de vraag of studenten een diploma verdienen of niet.

...OF NIET LATEN COMPENSEREN

Toch zijn er ook redenen om conjunctieve examensets te gebruiken. In sommige beroepen is het zeer belangrijk om alle onderdelen van de examens voldoende te maken. In de zorg en bij de politie is dit bijvoorbeeld sterk het geval. We willen immers geen verpleger aan ons bed die niet weet hoe een infuusnaald moet worden ingebracht. Verder is het heel gemakkelijk om uitslagen te berekenen van een conjunctieve examenset. (En tijd is geld.) Niet onbelangrijk is dat een conjunctieve beslisregel eenvoudig is uit te leggen. Beoordelaars maken minder optelfouten. Studenten weten meteen wat er wordt verwacht.

Een examenset meet over het algemeen accurater dan losse examens.

MEER LEZEN

Het volledige onderzoeksverslag is te lezen op www.kchexamens.nl. De engelstalige wetenschappelijke publicatie is terug te vinden als hoofdstuk 10 in Psychometrics in Practice at RCEC, op: http://www.rcec.nl/publicaties/overige%20publicaties/boekPS.pdf.

Literatuur

Clauser, B. (2000). Recurrent Issues and Recent Advances in Scoring Performance Assessments. **Applied Psychological Measurement,** 24(4), 310-324.

Dochy, F. (2009). The Edumetric Quality of New Modes of Assessment: Some Issues and Prospects. In G. Joughin (Ed.), **Assessment, Learning and Judgement in Higher Education** (pp. 85-114). Springer Science.

Gatti, G. G., & Buckendahl, C. W. (2006). On Correctly Classifying Examinees. In **Annual Meeting of the American Educational Research Association** (San Francisco, CA). Retrieved April 26, 2011 from http://www.unl.edu/buros/biaco/pdf/pres06gatti01.pdf.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A Five-Dimensional Framework for Authentic Assessment. **Educational Technology Research and Development, 52**(3), 67-85.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting Performance Standards on Complex Educational Assessments. **Applied Psychological Measurement, 24**(4), 355-366.

Hambleton, R., & Novick, M. (1973). Toward an Integration of Theory and Method for Criterion-Referenced Tests. **Journal of Educational Measurement, 10**(3), 159-170.

Holden, J. E., & Kelley, K. (2008). **Effects of Misclassified Data on Two Methods of Classification Analysis: A Monte Carlo Simulation Study.** Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Lee, W. C. (2008). **Classification Consistency and Accuracy for Complex Assessments Using Item Response Theory.** Iowa City: Center for Advanced Studies in Measurement and Assessment.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, Performance-Based Assessment: Expectations and Validation Criteria. **Educational Researcher, 20**(8), 15-21.

Martineau, J. A. (2007). An Expansion and Practical Evaluation of Expected Classification Accuracy. **Applied Psychological Measurement, 31**(3), 181-194.

Van Rijn, P., Béguin, A., & Verstralen, H. (2009). Zakken of Slagen? De Nauwkeurigheid van Examenuitslagen in het Voortgezet Onderwijs. **Pedagogische Studiën, 86**, 185-195.

Verstralen, H. (2009a). **Quality of Certification Decisions.** Arnhem: Cito.

VOORBEHOUD

De resultaten van dit onderzoek zijn niet alleszeggend. Er is onderzoek gedaan bij een subpopulatie van het handelsonderwijs (n=188). Bij een strengere cesuur, zouden minder studenten slagen en de gemeten classificatie accuraatheid dalen.

Geslaagd beoordelen

Invloeden op de classificatie accuraatheid van examensets

EVEN VOORSTELLEN

Drs. Marianne Hubregtse (29) werkt als promovenda bij Kenniscentrum Handel. Na een bachelor opleiding aan University College Utrecht heeft zij de Research Master Methodologie en Statistiek aan de Universteit Utrecht afgerond. Haar promotieonderzoek richt zich op de kwaliteit van de beoordeling van praktijkgericht examineren in het mbo. Dit onderzoek is deel van het promotieonderzoek.

11 / 2012 / 500