Zakken of slagen? De nauwkeurigheid van examenuitslagen in het voortgezet onderwijs

P. van Rijn, A. Béguin en H. Verstralen

Samenvatting

Een essentieel aspect van iedere toets is de meetnauwkeurigheid die wordt uitgedrukt in de betrouwbaarheid. Bij een combinatie van meerdere, uiteenlopende toetsen, zoals in het Nederlandse examensysteem in het voortgezet onderwijs, is de meetnauwkeurigheid van de beoordeling echter lastiger te bepalen. In dit artikel wordt een methode uitgewerkt waarbij meetnauwkeurigheid voor deze situatie wordt gekwantificeerd in termen van het percentage onjuist geclassificeerde kandidaten. De methode is gebaseerd op klassieke testtheorie en gebruikt modelmatige simulatie. Toegepast op examengegevens worden verschillende uitslagregels op de examens met elkaar vergeleken op basis van het percentage kandidaten dat zakt voor het examen en het percentage misclassificaties. Hierbij is uitgegaan van statistieken van examengegevens uit 2004, 2005 en 2006 en aangenomen dat leerlingen zich niet anders gaan gedragen. Uit de analyses blijkt dat het percentage gezakte kandidaten aanzienlijk toeneemt bij bijna alle alternatieve uitslagregels. De verschillen in percentage misclassificaties zijn ook substantieel maar subtieler, waarbij het blijkt dat de compensatorische eigenschappen van een uitslagregel gunstig zijn voor de meetnauwkeurigheid.

1 Inleiding

Meetnauwkeurigheid is een essentieel onderdeel van de standaarden voor kwalitatief goede toetsing die zowel nationaal (Evers, Van Vliet-Mulder, Resing, Starren, Van Alphen de Veer, & Van Boxtel, 2002) als internationaal (AERA, APA & NCME, 1999) zijn opgesteld. Vaak wordt in de verantwoording van een toets een schatting van de betrouwbaarheid zoals Cronbach's α (Cronbach, 1951) gerapporteerd. Een maat voor de betrouwbaarheid is echter minder relevant wanneer de beslissing over een persoon niet genomen wordt op basis van de resultaten van één toets, maar op basis van de combinatie van een aantal toetsen. De uiteindelijke nauwkeurigheid zal dan afhangen van de nauwkeurigheid van elk van die toetsen, de samenhang tussen de toetsresultaten en de manier waarop die resultaten worden samengenomen in de beslissing over de persoon. Iets dergelijks gebeurt bij de eindexamens in het voortgezet onderwijs in Nederland. Het eindexamen wordt afgelegd in een verzameling vakken, en bestaat voor een deel uit een schoolexamen en een centraal examen. De nauwkeurigheid, in de vorm van een betrouwbaarheidscoëfficiënt, is bekend voor de afzonderlijke examenvakken op het centraal examen. De betrouwbaarheid van elk vak wordt jaarlijks berekend op basis van een steekproef van examengegevens, en gerapporteerd in de examenverslagen van Cito (Alberts, 2008). Voor het schoolexamen zijn geen gegevens voorhanden om de betrouwbaarheid te berekenen, maar het is realistisch om te veronderstellen dat de betrouwbaarheid van het schoolexamen van dezelfde orde van grootte is als de betrouwbaarheid van het centraal examen (Verstralen & Van Rijn, 2008). In dit artikel wordt het begrip nauwkeurigheid verder uitgewerkt en vertaald naar de situatie van een examen dat uit meerdere vakken bestaat en uit een schoolexamen en centraal examen. Via een modelmatige aanpak kunnen we vervolgens vaststellen in hoeverre leerlingen op juiste wijze worden geclassificeerd als gezakt of als geslaagd op basis van hun examenresultaten en een bepaalde uitslagregel. Met andere woorden, hoe nauwkeurig is een uitspraak over zakken of slagen op basis van een verzameling examenresultaten? Tevens zal ingegaan worden op de vraag welk effect de toegepaste uitslagregel heeft op de nauwkeurigheid van de beslissing.

2 Combineren van examenresultaten

2.1 Uitslagregels

Het combineren van de resultaten op meerdere examens om tot een examenuitslag te komen kan op veel verschillende manieren gebeuren, waarbij een typering zoals in Tabel 1 kan worden aangebracht (Chester, 2003; Douglas, 2007). Wanneer voor alle examens een voldoende resultaat dient te worden behaald, dan wordt de uitslagregel conjunctief genoemd. Als niet op alle examens een voldoende resultaat behaald moet worden, dan wordt de uitslagregel bestempeld als complementair. Wanneer mindere resultaten kunnen worden verdisconteerd met betere resultaten, dan heet de uitslagregel compensatorisch.

Tabel 1

Typeringen van uitslagregels

Typering	Omschrijving
Conjunctief	Voldoende resultaat op alle examens
Complementair	Voldoende resultaat op een aantal examens
Compensatorisch	Voldoende resultaat op totaal

De uitslagregel die de afgelopen jaren is toegepast in het Nederlandse examensysteem voor het havo en vwo is als volgt: Om te slagen voor het examen mag een leerling maximaal één eindcijfer vier en één eindcijfer vijf halen, met maximaal één onvoldoende op de vakken van het gekozen profiel Natuur en Techniek (N&T), Natuur en Gezondheid (N&G), Economie en Maatschappij (E&M) of Cultuur en Maatschappij (C&M). Met ingang van het schooljaar 2008-2009 voor havo en het schooljaar 2009-2010 voor vwo gaat de uitslagregel veranderen. De eis van maximaal één onvoldoende op de profielvakken komt te vervallen. Echter, wanneer een leerling als eindcijfer één vier, twee vijven, of één vier en één vijf heeft, dan dient het gemiddelde eindcijfer ten minste 6,0 te zijn. Om te slagen voor het vmbo-examen mag een leerling maximaal één eindcijfer vier of twee eindcijfers vijf halen. In het geval van één vier of twee vijven dient tevens gecompenseerd te worden met minimaal één eindcijfer zeven op een ander vak. De oude uitslagregel voor havo en vwo is zowel compensatorisch als complementair. Compensatorisch, omdat resultaten op het schoolexamen en centraal examen per vak worden gemiddeld. Complementair, omdat een gering aantal onvoldoendes op de eindlijst is toegestaan. In de oude uitslagregel bij havo en vwo is er wel een specifiek element aan het complementaire aspect: onvoldoende resultaten op de profielvakken zijn beperkt. In de nieuwe uitslagregel bij havo en vwo, en bij de uitslagregel in het vmbo is er nog een compensatorisch aspect tussen vakken, omdat onvoldoendes voor bepaalde vakken op de eindlijst dienen te worden gecompenseerd met ruime voldoendes op andere vakken.

2.2 Alternatieve uitslagregels

De genoemde uitslagregels kunnen worden vergeleken met alternatieve regels om de uitslag te bepalen. Dit is relevant en actueel, omdat het ministerie van Onderwijs, Cultuur en Wetenschappen (OCW) recent vanuit verschillende hoeken is geadviseerd over het veranderen van het examensysteem ten behoeve van kwaliteitsverbetering dan wel niveauverhoging. Zo pleiten de door de voormalige minister van OCW Van der Hoeven ingestelde Profielcommissies Natuur & Techniek/Natuur & Gezondheid en Economie & Maatschappij/Cultuur & Maatschappij (2007) voor het ontkoppelen van het schoolexamen en centraal examen om recht te doen aan de functie en inhoud van beide examens. Waar het centraal examen van belang is voor het maatschappelijk vertrouwen in diploma's, toetst het schoolexamen minder doorstroomrelevante zaken (Profielcommissies Natuur en Techniek / Natuur en Gezondheid & Economie en Maatschappij / Cultuur en Maatschappij, 2007, pp. 62-63). De Lange en Dronkers (2007) bevinden in hun onderzoek dat cijfers op het schoolexamen in toenemende mate hoger zijn dan op het centraal examen en concluderen dat de waarde van het diploma achteruit gaat. De gevonden verschillen tussen schoolexamen en centraal examen worden gebruikt als argument voor

ontkoppeling van beide examens door onder andere de Commissie Parlementair Onderzoek Onderwijsvernieuwingen (2008). De Onderwijsraad (2007) pleit op haar beurt voor verplichte voldoendes op het havo- en vwo-examen voor de kernvakken Nederlands, Engels en wiskunde om de basisbagage te waarborgen. Ook de Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008) adviseert tot het onderzoeken van een invoering van een centrale danwel decentrale toetsing van doorstroomrelevante aspecten van taal en rekenen.

3 Methode

3.1 Klassieke testtheorie en meetnauwkeurigheid

De schoolexamens en centrale examens waar wij over spreken kunnen beschouwd worden als een verzameling toetsen. Om de statistische eigenschappen van deze verzameling toetsen te bestuderen maken we hier gebruik van de zogeheten klassieke testtheorie (zie bijv. Lord & Novick, 1968; Mellenbergh & Van den Brink, 1998). Met behulp van klassieke testtheorie, kunnen we het Nederlandse examensysteem modelmatig bestuderen.

In het beschrijven van klassieke testtheorie, gebruiken we de meer algemene term toets in plaats van examen. Klassieke testtheorie veronderstelt dat er bij een toetsafname een meetfout kan optreden en dat deze meetfout beschouwd kan worden als een aselecte trekking uit een kansverdeling. Hierdoor valt af te leiden dat voor een aselect getrokken persoon, de score op een toets (X) is opgebouwd uit een ware score (T, *true score*) en een meetfout (E, *measurement error*),

$$X = T + E.$$

De verwachte waarde van de geobserveerde examenscores is gelijk gesteld aan die van de ware scores:

$$E(X) = E(T) \text{ of } \mu_X = \mu_T.$$

Door aan te nemen dat de meetfout niet samenhangt met de ware score, kan de variantie van de toetsscore X worden geschreven als:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2.$$

Het belangrijkste concept uit de klassieke testtheorie is dat van betrouwbaarheid. De betrouwbaarheid van een toets is gedefinieerd als de gekwadrateerde correlatie tussen de geobserveerde score en de ware score:

$$\rho_{XT}^2 = \frac{\sigma_{XT}^2}{\sigma_X^2 \sigma_T^2} = \frac{\sigma_T^2}{\sigma_X^2},$$

waar σ_{XT}^2 de gekwadrateerde covariantie is tussen de geobserveerde score X en de ware score T. De betrouwbaarheid wordt doorgaans gezien als de (populatie-afhankelijke) meetnauwkeurigheid van een toets.

In het algemeen geldt voor de Nederlandse centrale examens dat de toetsscore een behaald puntenaantal behelst en dat deze toetsscore is begrensd. De behaalde toetsscore wordt vervolgens omgezet in een cijfer middels een lineaire transformatie. In het geval van schoolexamens kunnen we niet eenvoudigweg spreken van een toetsscore, omdat het schoolexamen vaak uit meerdere toetsen bestaat en ook herkansingsmogelijkheden bevat. Wij beperken ons echter tot de examen*cijfers* als we het hebben over toetsscores, omdat dat de enige beschikbare gegevens zijn. Dit is dus geen exacte modellering van schoolexamens en centrale examens, maar een noodzakelijke benadering.

Er geldt natuurlijk dat de cijfers op verschillende examens met elkaar samenhangen. Een positieve correlatie zorgt er voor dat een cijfer dat boven het gemiddelde ligt op een examen Frans ertoe leidt dat de verwachte score op het examen Engels ook boven het gemiddelde zal liggen. Er wordt nu verondersteld dat de samenhang tussen de cijfers op verschillende examens verloopt via een samenhang in ware scores en niet via een samenhang in meetfout. Verder geldt dat de zojuist genoemde maat van betrouwbaarheid gebruikt kan worden om te schatten welk deel van de score gebaseerd is op de ware score en welk deel gebaseerd is op meetfout.

Wanneer we de variantie-covariantiematrix van een verzameling van *n* examens $X' = (X_1, X_2, ..., X_n)$ noteren als Σ_X, kunnen we deze schrijven als de som van de variantie-covariantie matrices van de ware scores

en de meetfouten:

$$\Sigma_X = \Sigma_T + \Sigma_E$$

We kunnen nu de variantie-covariantie matrices van de verzameling ware scores T en meetfouten E als volgt schrijven:

$$\Sigma_T = R_X^2 \Sigma_X,$$
$$\Sigma_E = (1\text{-}R_X^2) \Sigma_X,$$

waar R_X^2 een diagonaalmatrix is met op de diagonaal de betrouwbaarheden van de betreffende examens en I een $n \times n$-identiteitsmatrix is. De matrix van de meetfouten is dus diagonaal, dat wil zeggen de meetfouten voor de verschillende examens hangen niet samen met elkaar.

3.2 Opzet simulaties

Om de examens en verschillende uitslagregels modelmatig te bestuderen is gebruik gemaakt van een gegevensbestand met cijfers voor schoolexamen en centraal examen voor de jaren 2004, 2005 en 2006. De gemiddelden, varianties en covarianties van alle examenvakken uit dit gegevensbestand zijn berekend om het model van de klassieke testtheorie toe te kunnen passen en zodoende de meetnauwkeurigheid van de verschillende uitslagregels te kunnen bepalen. De betrouwbaarheden van de verschillende centrale examens zijn afkomstig uit de examenverslagen 2004, 2005 en 2006 van Cito en zijn gemiddeld over de drie jaar. Van de schoolexamens is de betrouwbaarheid in principe onbekend en waarschijnlijk verschillend per school. Wel is het mogelijk om een schatting te maken van deze betrouwbaarheden via een aanname over de correlatie tussen de ware scores van het schoolexamen en het centraal examen. Door deze aanname over de correlatie kan de correctie voor attenuatie worden gebruikt om de onbekende betrouwbaarheid van het schoolexamen te schatten. De correctie voor attenuatie geeft de gekwadrateerde correlatie tussen de ware scores van twee toetsen (T_1 en T_2) met behulp van de gekwadrateerde correlatie tussen de geobserveerde scores (X_1 en X_2) en de betrouwbaarheden van beide toetsen, dus:

$$\rho^2_{T_1T_2} = \frac{\rho^2_{X_1X_2}}{\rho^2_{X_1}\rho^2_{X_2}}.$$

Op basis van een kleinschalige studie lijkt het een aannemelijke keuze om de betrouwbaarheden van de schoolexamens gelijk te nemen aan de betrouwbaarheid van de centrale examens in dezelfde vakken (Verstralen & Van Rijn, 2008). Voor de vakken waar helemaal geen centraal examen voor is, is de betrouwbaarheid genomen van het vak wat het er het meest op lijkt (bijv. de betrouwbaarheid voor het vak natuurkunde wordt gebruikt voor het vak algemene natuurwetenschappen).

Door aan te nemen dat de examengegevens normaal verdeeld zijn, kunnen we ware scores, meetfouten en geobserveerde scores verkrijgen door middel van simulatie. Dit hebben we uitgevoerd voor het meest voorkomende vakkenpakket in elk van de vier profielen in het havo en vwo en elk van de vier sectoren in de gemengde en theoretische leerwegen van het vmbo.

3.3 Evaluatiecriteria

Verschillende uitslagregels worden vergeleken op basis van twee indicatoren. De eerste indicator is het percentage gezakte leerlingen in de examengegevens 2004-2006. De tweede indicator is het percentage misclassificaties. Dit percentage kan voor een verzameling examens als volgt worden berekend. In het geval van één toets en een gehanteerde cesuur (de grens tussen onvoldoende en voldoende) kan een beslissingstabel worden opgesteld zoals weergegeven in Tabel 2. Wanneer zowel de ware als de geobserveerde score tot dezelfde uitslag leiden, onvoldoende dan wel voldoende, is er sprake van een juiste classificatie. Wanneer de ware en geobserveerde score leiden tot verschillende uitslagen, is er sprake van een misclassificatie.

De beslissingstabel kan worden geïllustreerd met behulp van Figuur 1. Hierin is een verdeling van examencijfers afgebeeld met als gemiddelde examencijfer 6,44, standaardafwijking 1,20 en betrouwbaarheid 0,79. De verticale lijn geeft de minimale voldoende van 5,5 weer, de cesuur. De stippellijnen geven de conditionele verdelingen

Tabel 2

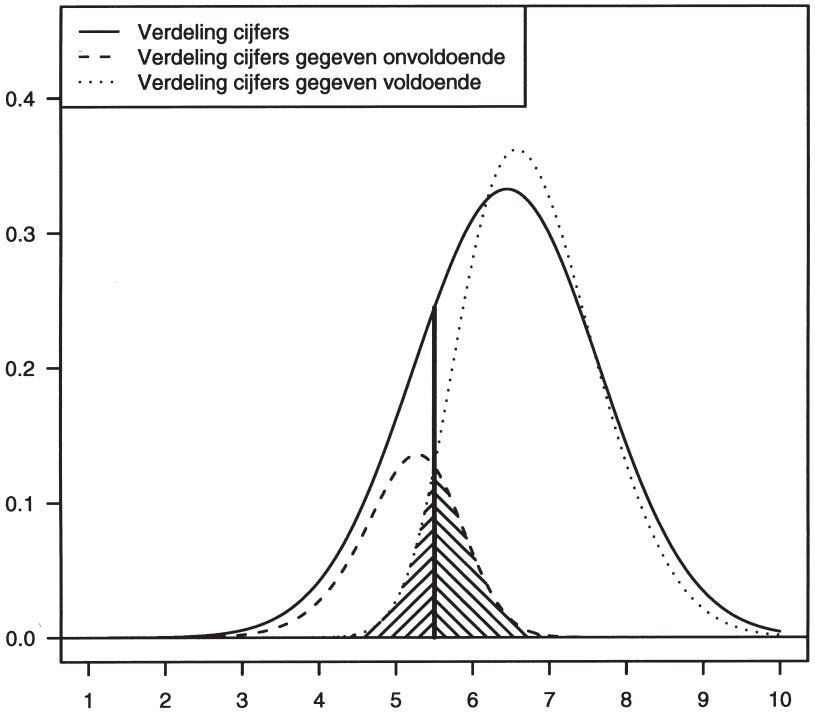
Beslissingstabel voor bepalen van nauwkeurigheid van één toets

		Ware score T	
		Onvoldoende	Voldoende
Geobserveerde score X	Onvoldoende	Juiste classificatie	Misclassificatie
	Voldoende	Misclassificatie	Juiste classificatie

weer voor de situaties dat de ware score onder dan wel boven de cesuur ligt. Het gearceerde oppervlak geeft het percentage misclassificaties weer voor dit examen. Het gearceerde oppervlak onder de cesuur kan geïnterpreteerd worden als het percentage leerlingen dat onjuist geclassificeerd wordt als onvoldoende en het gearceerde oppervlak boven de cesuur geeft het percentage leerlingen aan dat onjuist geclassificeerd wordt als voldoende.

De beslissingstabel voor het gehele examen (Tabel 3) is een variant op de tabel voor een enkel vak. Nu is de geobserveerde uitkomst niet een score maar een uitslag op het gehele examen, namelijk gezakt of geslaagd. Deze uitkomst kan worden bepaald door de uitslagregel toe te passen op de examenresultaten van zowel schoolexamen als centraal examen van alle vakken die de kandidaat heeft gedaan. Op eenzelfde manier wordt ook de ware score vervangen door, in dit geval, de ware uitkomst gegeven het vaardigheidsniveau van de kandidaat.

De in dit onderzoek bestudeerde uitslagregels zijn weergegeven in Tabel 4 en worden vergeleken op basis van het percentage gezakte leerlingen en het percentage misclassificaties. Het uitgangspunt is de uitslagregel die van toepassing was in de periode waarover examengegevens beschikbaar zijn. Op twee na zijn de onderzochte uitslagregels verscherpingen van de toen geldige uitslagregel. Bij deze uitslagregels worden immers extra eisen toegevoegd aan de bestaande uitslagregel. Alleen de vierde en de zevende uitslag-



Figuur 1. Verdeling van examencijfers en misclassificaties.

Tabel 3

Beslissingstabel voor bepalen van nauwkeurigheid van heel examen

		Vaardigheidsniveau	
		Onvoldoende	Voldoende
Examenuitslag	Gezakt	Juiste classificatie	Misclassificatie
	Geslaagd	Misclassificatie	Juiste classificatie

regel zijn niet noodzakelijk een verscherping. Leerlingen kunnen zakken op basis van de oude uitslagregel, maar slagen wanneer uitslagregel vier of zeven wordt toegepast, en andersom. Het dient opgemerkt te worden dat de benaming oude en nieuwe uitslagregel niet van toepassing is voor het vmbo, omdat hier vooralsnog niets gewijzigd is.

4 Resultaten

4.1 Beschrijvende statistieken

Om inzicht te krijgen in de recente resultaten op het schoolexamen en centraal examen zijn de examengegevens van de jaren 2004, 2005 en 2006 geanalyseerd. In Tabel 5 staat een overzicht van beschrijvende statistieken van deze examengegevens voor de schooltypen vwo, havo en vmbo-gemengde leerweg (gl) en vmbo-theoretische leerweg (tl). Per schooltype en per jaar worden de gemiddelde score en de standaarddeviatie van de scores op het schoolexamen, van de scores op het schoolexamen waar ook een centraal examen voor is en van de scores op het centraal examen gegeven samen met het gemiddelde verschil tussen schoolexamen en centraal examen. Dit gemiddelde verschil is berekend door steeds per leerling het verschil tussen cijfers te nemen voor vakken die worden afgesloten met zowel een schoolexamen als een centraal examen en vervolgens te middelen. Uit Tabel 5 blijkt dat het percentage leerlingen dat zakt voor het examen het grootst is in het havo en het kleinst in vmbo-gl/tl. Het verschil tussen schoolexamen en centraal examen in het vwo loopt van 0,29 in 2004 tot 0,40 in 2006, terwijl dit verschil voor havo en vmbo-gl/tl wat kleiner is en fluctueert tussen de 0,17 en 0,28.

Resultaten voor havo en vmbo-gl/tl worden verder niet gepresenteerd, omdat de resultaten met betrekking tot het vergelijken van de uitslagregels in grote lijn vergelijkbaar zijn met de resultaten voor vwo. We richten ons in het vervolg van deze sectie dus op het vwo. Tabel 6 geeft een nader gespecificeerd overzicht van de beschrijvende statistieken uit Tabel 5, maar dan voor de verschillende profielen in het vwo. Van de overige 31 kandidaten met een andere profielcombinatie zijn geen statistieken in de tabel opgenomen.

4.2 Percentages gezakte en geslaagde leerlingen bij vwo

In Tabel 7 worden de percentages gezakte en geslaagde leerlingen gegeven voor de vier profielen aangevuld met het gecombineerde profiel Natuur & Techniek en Natuur & Gezondheid voor de verschillende uitslagregels

Tabel 4

Verschillende uitslagregels

Uitslagregel	
Oude uitslagregel	0
Oude uitslagregel + gemiddelde CE-cijfer voldoende	1
Oude uitslagregel + CE-cijfer NL, EN en WI voldoende	2
Oude uitslagregel + eindcijfer NL, EN en WI voldoende	3
Oude uitslagregel op SE-cijfers én op CE-cijfers	4
Oude uitslagregel + Gem. CE voldoende met max. één 5 als CE-cijfer voor NL, EN, WI	5
Oude uitslagregel + Gem. CE voldoende met max. één 5 als eindcijfer voor NL, EN, WI	6
Nieuwe uitslagregel[1]	7

[1] Oude en nieuwe uitslagregel zijn alleen in het havo en vwo ongelijk

Tabel 5

Beschrijvende statistieken van gebruikte examengegevens

Schooltype	Jaar	Aantal Leer-lingen[1]	Gezakt[2] (%)	Gem. SE	Sd. SE	Gem. SE met CE	Sd. SE met CE	Gem. CE	Sd. CE	Gem. verschil
Vwo	2004	29.393	6,0	6,89	0,86	6,80	0,85	6,51	1,20	0,29
	2005	30.745	6,1	6,90	0,87	6,81	0,86	6,42	1,19	0,39
	2006	31.576	7,4	6,90	0,87	6,80	0,86	6,40	1,20	0,40
Havo	2004	40.394	8,7	6,53	0,78	6,46	0,79	6,29	1,09	0,17
	2005	41.623	10,0	6,52	0,79	6,45	0,79	6,22	1,11	0,23
	2006	42.882	11,3	6,52	0,79	6,44	0,79	6,16	1,10	0,28
Vmbo-gl/tl	2004	47.881	5,6	6,60	0,83	6,53	0,81	6,36	1,07	0,17
	2005	47.947	5,8	6,61	0,82	6,54	0,81	6,35	1,12	0,20
	2006	49.271	5,1	6,63	0,82	6,56	0,81	6,37	1,13	0,19

[1] Aantal leerlingen dat gezakt óf geslaagd is, geen particuliere scholen en vavo´s.
[2] Herkansingen waren al verwerkt in de data.

Tabel 6

Beschrijvende statistieken van gebruikte vwo examengegevens

Profiel	Aantal Leer-lingen	Gezakt (%)	Gem. SE	Sd. SE	Gem, SE met CE	Sd. SE met CE	Gem CE	Sd. CE	Gem. Ver-schil
Natuur en Techniek	11.208	6,9	6,94	0,92	6,87	0,93	6,62	1,29	0,25
Natuur en Gezondheid	27.370	7,6	6,94	0,89	6,78	0,86	6,47	1,20	0,31
Economie en Maatschappij	30.382	6,4	6,80	0,83	6,73	0,81	6,33	1,14	0,41
Cultuur en Maatschappij	18.337	5,7	6,91	0,84	6,83	0,85	6,39	1,19	0,43
Natuur en Techniek & Natuur en Gezondheid	3.841	3,1	7,20	0,94	7,14	0,93	6,95	1,26	0,19

Tabel 7

Percentage leerlingen dat zakt bij elk van de verschillende uitslagregels per profiel in het VWO

	Uitslagregel	N&T	N&G	E&M	C&M	N&T N&G	Totaal
0	Oude uitslagregel[1]	6,8	7,4	6,3	5,5	2,9	6,4
1	Oude uitslagregel + gem. CE-cijfer voldoende	12,1	12,9	13,5	12,7	5,2	12,6
2	Oude uitslagregel + CE-cijfer NL, EN en WI voldoende	45,0	46,1	49,5	44,3	32,8	46,2
3	Oude uitslagregel + eindcijfer NL, EN en WI voldoende	25,4	26,6	27,6	23,2	15,2	25,6
4	Oude uitslagregel op SE-cijfers én op CE-cijfers	23,8	28,9	32,1	31,7	15,5	29,3
5	Oude uitslagregel + Gem. CE voldoende met max. één 5 als CE-cijfer voor NL, EN, WI	23,0	24,1	26,0	23,3	14,2	24,0
6	Oude uitslagregel + Gem. CE voldoende met max. één 5 als eindcijfer voor NL, EN, WI	13,9	14,6	15,5	14,2	6,4	14,4
7	Nieuwe uitslagregel	6,6	6,9	6,4	5,6	2,7	6,2

[1] Het percentage gezakte leerlingen wijkt licht af van dat in Tabel 6, omdat alleen leerlingen in de analyses zijn opgenomen bij wie het toepassen van de uitslagregel hetzelfde resultaat opleverde als de geregistreerde uitslag.

op basis van de examengegevens 2004-2006. De percentages gezakte kandidaten stijgen aanzienlijk wanneer de alternatieve uitslagregels worden toegepast, van 6,4% bij de oude uitslagregel tot 46,2% als ook elk van de vakken Nederlands, Engels en wiskunde voldoende moet zijn op het centraal examen (regel 2). Regel 1 heeft het kleinste effect, maar leidt toch al bijna tot een verdubbeling van het percentage kandidaten dat zakt. Hoewel regel 4 strikt genomen niet automatisch een verzwaring van de exameneisen is, blijkt op basis van onze analyse deze regel wel te leiden tot een hoger percentage gezakte kandidaten. Het toepassen van de nieuwe uitslagregel, regel 7, op de examengegevens leidt tot een lichte daling van het percentage gezakte kandidaten.

4.3 Percentage misclassificaties bij vwo

In Tabel 8 worden de percentages misclassificaties gegeven voor de verschillende profielen in het vwo. Alleen regel 1 en 6 leiden niet tot een substantiële toename van het percentage misclassificaties. De aanvullende eis dat het gemiddelde cijfer op het centraal examen voldoende moet zijn (regel 1) leidt in totaal tot een zeer kleine stijging van het percentage misclassificaties (6,4%). Wanneer deze eis verder aangescherpt wordt door maar voor één van de kernvakken; Nederlands, Engels en wiskunde, een onvoldoende eindcijfer toe te staan dat niet lager mag zijn dan het cijfer 5 (regel 6), stijgt het percentage misclassificaties naar 6,7%. Als in plaats daarvan gekeken zou worden naar de cijfers op het centraal examen in deze kernvakken (regel 5) wordt het percentage misclassificaties 11,3%. De aanvullende eis dat geen van de kernvakken onvoldoende mag zijn leidt tot 9.8% misclassificaties als gekeken wordt naar het eindcijfer (regel 3) en tot 17,7% misclassificaties als naar het centraal examen (regel 2) wordt gekeken. Loskoppelen van het schoolexamen en centraal examen door op beide examens de oude uitslagregel toe te passen leidt tot 11,7% misclassificaties. De nieuwe uitslagregel, regel 7, leidt als enige tot een vermindering van het percentage misclassificaties (5,2%) ten opzichte van de oude uitslagregel (6,0%).

5 Discussie

5.1 Het effect van compensatie

De resultaten van de modelmatige vergelijking van de verschillende uitslagregels kunnen voor een belangrijk deel worden verklaard vanuit de mate van compensatie die tussen resultaten kan plaatsvinden. Wanneer hogere eisen worden gesteld aan de prestatie op individuele vakken zal het percentage onvoldoendes en het percentage misclassificaties hoger zijn dan als hogere eisen worden gesteld aan een combinatie van vakken. Wanneer de eisen worden gesteld aan het centraal examen in plaats van het eindcijfer wordt dit effect versterkt doordat er geen compensatie tussen het schoolexamen en het centraal examen plaats kan vinden. Dit effect voorspelt een stijgende reeks in percentage onvoldoende en percentage misclassificaties bij de

Tabel 8

Percentage misclassificaties voor de verschillende uitslagregels per profiel in het vwo

	Uitslagregel	N&T	N&G	E&M	C&M	N&T N&G	Totaal
0	Oude uitslagregel	6,0	6,5	6,5	4,5	6,2	6,0
1	Oude uitslagregel + gem. CE-cijfer voldoende	6,3	6,5	6,9	5,2	6,3	6,4
2	Oude uitslagregel + CE-cijfer NL, EN en WI voldoende	17,,0	16,7	18,6	18,3	17,0	17,7
3	Oude uitslagregel + eindcijfer NL, EN en WI voldoende	9,8	10,3	11,3	6,6	10,1	9,8
4	Oude uitslagregel op SE-cijfers én op CE-cijfers	9,7	11,3	13,2	11,3	9,9	11,7
5	Oude uitslagregel + Gem. CE voldoende met max. één 5 als CE-cijfer voor NL, EN, WI	10,9	11,0	12,0	11,0	10,8	11,3
6	Oude uitslagregel + Gem. CE voldoende met max. één 5 als eindcijfer voor NL, EN, WI	6,6	6,9	7,4	5,3	6,7	6,7
7	Nieuwe uitslagregel	**5,6**	**5,9**	**5,4**	**3,5**	**5,7**	**5,2**

uitslagregels 0, 1, 6 en 5. Bij deze regels worden respectievelijk de eisen:

- het gemiddelde CE moet voldoende zijn;
- Eis 1 en er mag maximaal één onvoldoende niet lager dan vijf worden gehaald op het eindcijfer van Nederlands, Engels en wiskunde, en
- Eis 1 en er mag maximaal één onvoldoende niet lager dan vijf worden gehaald op het centraal examen van Nederlands, Engels en wiskunde toegevoegd aan de oude uitslagregel.

Deze eisen worden steeds iets strenger en geven minder mogelijkheid tot compenseren. Op dezelfde manier kan de ordening van de uitslagregels 0, 3 en 2 worden voorspeld. Doordat in de nieuwe uitslagregel meer compensatie mogelijk is dan in de oude uitslagregel (de eis op de profielvakken komt te vervallen) én het percentage gezakte leerlingen niet toeneemt, is de gevonden vermindering van het percentage misclassificaties ook hier te verklaren.

5.2 Interpretatie van het percentage gezakte kandidaten

Uit de uitgevoerde analyses blijkt dat het percentage gezakte kandidaten aanzienlijk toeneemt bij elk van de alternatieve uitslagregels. Dit wil echter niet automatisch zeggen dat dit percentage kandidaten echt zou zakken als deze uitslagregel in de praktijk zou worden ingevoerd. In dat geval kan immers verwacht worden dat kandidaten rekening houden met de specifieke uitslagregel en hun inspanning anders verdelen over vakken en examens. Ook kunnen kandidaten hun totale inspanning voor het examen of voor hun opleiding verhogen als er strengere eisen worden gesteld. Een andere manier waarop het aantal gezakte kandidaten zou kunnen dalen is dat de standaard op het examen word verlaagd door de hiervoor bevoegde instantie. Dit zou betekenen dat met eenzelfde prestatie op het examen bij de nieuwe uitslagregel een hoger cijfer wordt gegeven dan werd gedaan bij de oude uitslagregel. Ook op deze manier kan in de praktijk het percentage gezakte kandidaten lager uitvallen dan op basis van onze analyses wordt voorspeld.

Voordat eventueel over gegaan kan worden tot het invoeren van een andere uitslagregel is het zinvol om te onderzoeken wat het effect is van een wijziging van de uitslagregel op het percentage kandidaten dat zakt binnen specifieke groepen kandidaten. Uit diverse onderzoeken blijkt dat de relatieve prestaties op het schoolexamen en centraal examen van jongens en meisjes verschillen. Ook blijkt dat de prestaties van allochtone leerlingen vaak grotere verschillen te zien geven tussen het schoolexamen en het centraal examen (Centraal Bureau voor de Statistiek, 2002; Rekers-Mombarg & Harms, 2008). Een wijziging van uitslagregel kan dus meer of juist minder invloed hebben op deze verschillende groepen leerlingen. Dit aspect is niet meegenomen in de analyses in het huidige onderzoek, maar kan wel relevante informatie zijn, die nodig is om een afgewogen keuze te maken voor een wijziging van de uitslagregel.

5.3 Interpretatie van het percentage misclassificaties

Een absolute interpretatie van het percentage misclassificaties als maat voor nauwkeurigheid en kwaliteit van de beslissing is niet in alle gevallen gerechtvaardigd. De hoogte van dit percentage hangt namelijk af van de populatie waarbinnen dit percentage is bepaald. Ook spelen de vorm en de plaats van de cesuur binnen de verdeling van kandidaten een rol. Een kandidaat van wie de ware score dicht bij de cesuur ligt zal een hogere kans hebben op misclassificatie dan een kandidaat waarvan de ware score ver van de cesuur ligt. In het uiterste geval waarbij de ware score van een kandidaat op de cesuur ligt zal de kandidaat een kans van 50% hebben om te slagen of te zakken. Logischerwijs is de kans op misclassificatie dan ook 50%. In het huidige onderzoek werd het percentage misclassificaties binnen dezelfde populatie gebruikt als vergelijkingsbasis voor verschillen in nauwkeurigheid bij beslissingen op basis van verschillende uitslagregels. De relatieve interpretatie is gerechtvaardigd doordat steeds binnen dezelfde populatie kandidaten wordt vergeleken. Wel kan een verzwaring van de eisen het effect hebben dat de relatieve positie van de cesuur in de populatie verschuift. Dit heeft dan gevolgen voor het percentage misclassificaties. Wanneer de cesuur zo verschuift dat er meer leerlingen een ware

score in de buurt van de cesuur hebben zal dit leiden tot een hoger percentage misclassificatie. Omgekeerd zal het percentage misclassificaties dalen als er minder leerlingen een ware score hebben in de buurt van de cesuur.

Een methode om het percentage misclassificaties te verkleinen is het vergroten van de betrouwbaarheid van de examens waarop de uitslagregel wordt toegepast. Een standaardmethode voor het vergroten van de betrouwbaarheid is het langer maken van de toets. Met deze methode is het mogelijk om het percentage misclassificaties te verkleinen voor de uitslagregels waarin verplichte voldoendes voor bepaalde vakken worden vereist. De randvoorwaarden waarbinnen de huidige examens worden gemaakt met betrekking tot de lengte, de duur en de vorm kunnen echter problematisch zijn voor de methode van toetsverlenging.

Tot slot, in de huidige analyses is geen rekening gehouden met het effect van een eventuele herkansing. Logischerwijs zal de mogelijkheid van herkansing leiden tot een lager percentage kandidaten dat zakt. Het effect op het percentage misclassificaties is onduidelijk. Hier speelt een rol dat zowel kandidaten een herkansing doen die ten onrechte zijn gezakt (fout negatief) als kandidaten die terecht waren gezakt. Kandidaten uit de eerste categorie die alsnog slagen leiden tot een daling van het aantal misclassificaties. Het omgekeerde geldt voor de kandidaten die terecht gezakt waren. Als zij na herkansing alsnog slagen dan stijgt daarmee het aantal fout positieve classificaties en daarmee het aantal misclassificaties. Onbekend is in welke proportie terecht en onterecht gezakte kandidaten deelnemen aan een examen en bij gevolg weten we niet wat het effect van een herkansing is op het percentage misclassificaties. Wel moet opgemerkt worden dat een herkansing leidt tot een daling van het percentage onterecht gezakte kandidaten maar tot een stijging van het percentage onterecht geslaagde kandidaten.

De beoogde effecten van voorgestelde uitslagregels zoals niveauverhoging of verhoging van het maatschappelijk vertrouwen gaan dus niet altijd hand in hand met een verbetering van de meetnauwkeurigheid van het eindexamen. De resultaten van de huidige studie zijn natuurlijk in beperkte mate generaliseerbaar naar toekomstige eindexamens vanwege de aannames over bijvoorbeeld de betrouwbaarheid en onveranderd leerlinggedrag. Aangezien het eindexamen echter een bepalend moment is voor ieder individu in het voortgezet onderwijs, hoort een weloverwogen afweging over de meetnauwkeurigheid thuis in de discussie over eventuele aanpassingen in de uitslagregel van het eindexamen.

Literatuur

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Alberts, R. V. J. (2008). *Verslag van de examencampagne 2008: voortgezet onderwijs*. Arnhem, Nederland: Cito.

Central Bureau voor de Statistiek. (2002). *Jaarboek onderwijs in cijfers, 2002. Feiten en cijfers over het onderwijs in Nederland*. Den Haag, Nederland: Centraal Bureau voor de Statistiek.

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

Commissie Parlementair Onderzoek Onderwijsvernieuwingen. (2008). *Tijd voor Onderwijs. Eindrapport.* Den Haag, Nederland: Sdu Uitgeverij.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Douglas, K. M. (2007). *General method for estimating the classification reliability of complex decisions based on configural combinations of multiple assessment scores*. Dissertatie. University of Maryland, College Park, MD, Verenigde Staten.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen. (2008). *Over de drempels met taal en rekenen.* Enschede, Nederland: Expertgroep Doorlopende Leerlijnen Taal en Rekenen.

Evers, A., Vliet-Mulder, J. C. van, Resing, W. C. M., Starren, J. C. M. G., Alphen de Veer, R. J. van, & Boxtel, H. van. (2002). *COTAN Test-*

boek voor het onderwijs. Amsterdam: NDC-Boom.

Lange, M. de, & Dronkers, J. (2007).*Hoe gelijkwaardig blijft het eindexamen tussen scholen in Nederland? Discrepanties tussen de cijfers voor het schoolonderzoek en het centraal examen in het voortgezet onderwijs tussen 1998 en 2005* (EUI working papers SPS No. 2007/03).

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

Onderwijsraad. (2007) *Versteviging van kennis in het onderwijs II*. Den Haag, Nederland: Onderwijsraad.

Profielcommissies Natuur en Techniek / Natuur en Gezondheid & Economie en Maatschappij / Cultuur en Maatschappij. (2007). *Eindadvies: Kennis, kwaliteit en keuze in de tweede fase.* Harderwijk, Nederland: Flevodruk.

Rekers-Mombarg, L. T. M., & Harms, G. J. (2008) *Meten met twee maten? De discrepantie tussen de cijfers op het schoolexamen en het centraal examen VO van allochtone leerlingen.* Groningen, Nederland: Gion.

Verstralen, H. H. F. M., & Rijn, P. W. van. (2008, juni). *De betrouwbaarheid van het schoolexamen.* Paper gepresenteerd op de Onderwijsresearchdagen, Eindhoven, Nederland

Manuscript aanvaard: 3 februari 2009

Auteurs

Peter van Rijn, **Anton Béguin** en **Huub Verstralen** zijn als onderzoekers werkzaam bij het psychometrisch onderzoekcentrum van Cito.

Correspondentieadres: Peter van Rijn, Postbus 1034, 6801 MG Arnhem. Email: peter.vanrijn@cito.nl.

Abstract

Failing or passing? Measurement precision of examinations in secondary education

Measurement precision is an essential aspect of an examination or any other test and is commonly quantified by an estimate of the reliability of the test or the standard error of measurement. While measurement precision is easy to determine for single tests, it is much more difficult to determine for multiple tests on different subjects as in the Dutch examination system for secondary education. In this paper a method for quantifying measurement precision for multiple tests is presented. Measurement precision is quantified in terms of misclassifications, that is, the number or percentage of candidates that is either correctly or incorrectly classified by the tests being administered. This method is used to assess the effect of different pass/fail decision rules on the number of misclassifications. Applying the method to examinations from 2004, 2005, and 2006, where is assumed that students do not change their behaviour, it is shown that compensatory decision rules are to be preferred over conjunctive decision rules.