

Is het ene examen het andere? Over de stabiliteit van schoolonderzoek en centraal examen

H. van den Bergh, E. Rohde en M. Zwarts

Samenvatting

Examens zijn zowel voor de betrokken leerlingen als voor de samenleving van groot belang. Om over een vaste maatstaf te beschikken, zijn met de Mammoetwet de centrale examens (CE) ingevoerd, die in combinatie met de schoolonderzoeken (SO) (nu school-examens) de uiteindelijke cijfers op het examen bepalen. Over de waarde van het centraal examen om alle leerlingen van een bepaalde jaargang op een vergelijkbare wijze te beoordelen, is weinig discussie. Over de mogelijkheden om ook over de jaren heen de kwaliteit te bewaken is meer controverser. Dat geldt ook voor de waarde van het schoolonderzoek bij het beoordelen van de prestaties van leerlingen.

In dit artikel beantwoorden we drie vragen:

- 1 In hoeverre geven verschillen in de SO-cijfers tussen scholen en jaren verschillen in de vaardigheid van de leerlingen op die scholen weer?
- 2 In hoeverre geven verschillen in de CE-cijfers in opeenvolgende jaren verschillen in de vaardigheid van de leerlingen weer?
- 3 Heeft het gebruik van normhandhavingsonderzoek dat sinds 1996 plaatsvindt, geleid tot een grotere gelijkwaardigheid van het CE-cijfer over jaren?

De opzet van het schoolonderzoek als een schoolgebonden waardering voor de prestaties van leerlingen laat een vergelijking tussen scholen en jaren niet toe. Voor het bepalen van de cijfers gebruiken scholen vermoedelijk de schoolpopulatie en de prestaties in voorgaande jaren als referentiekader. Dit leidt tot over de jaren heen stabiele resultaten die eventuele verschillen in vaardigheid van de leerlingen tussen jaren maskeren.

De constructie, de correctie, de normering en de afname van de centrale examens vormen een garantie dat de examenresultaten gebruikt kunnen worden om de prestaties van leerlingen binnen een jaar op een gelijkwaardige wijze te beoordelen. Die garantie geldt

niet voor de examens van verschillende jaren. De resultaten tussen de verschillende jaren wijken sterker van elkaar af dan redelijkerwijs toegeschreven kan worden aan veranderingen in het vaardigheidsniveau van leerlingen. Het normhandhavingsonderzoek dat vanaf 1994 door het Cito wordt uitgevoerd en is bedoeld om tot een meer gelijkwaardige cijfergeving te komen, heeft niet geleid tot afname van de verschillen tussen de gemiddelde CE-cijfers in opeenvolgende jaren. De centrale examens zijn dus geen betrouwbaar middel om ontwikkelingen van de vaardigheden van leerlingen over de jaren te volgen. Dit betekent ook dat het voor de slagingskans van leerlingen uitmaakt in welk jaar ze examen doen.

1 Inleiding

Examens hebben voor de verschillende belanghebbenden een aantal functies (Inspectie van het Onderwijs, 2001). Voor de leerling is de toegang tot het vervolgonderwijs (de paspoortfunctie) het meest belangrijk. Voor de school bieden de examens de mogelijkheid om de realisering van hun eigen doelen en programma te meten (interne kwaliteitsmeting en -borging). Voor de centrale overheid is de sturingsfunctie van belang: de mogelijkheid om het niveau van het onderwijs te volgen en die informatie te gebruiken voor haar beleid. Die functies kunnen alleen vervuld worden als de examens aan bepaalde eisen voldoen. Scholen kunnen hun eigen kwaliteit slechts meten als de resultaten met andere jaren en met andere scholen vergeleken kunnen worden. Ook de overheid kan het niveau van het onderwijs slechts volgen als aan die eisen van vergelijkbaarheid voldaan wordt. Vergelijkbaarheid houdt hier in dat eventuele verschillen tussen scholen of jaren toegeschreven kunnen worden aan verschillen in de kennis en vaardigheden van leerlingen, en niet veroorzaakt worden door andere factoren, zoals de eigenschappen van de ge-

bruikte toetsen. Bij de mogelijkheden om zinvolle vergelijkingen te maken, zijn al sinds de invoering van het systeem vraagtekens gezet.

In 1992 (Inspectie van het Onderwijs, 1992) schreef de inspectie over de ontwikkeling van de normen bij de centrale examens. De directe aanleiding was toen de zorg van de Tweede Kamer voor een daling van het niveau van de prestaties die niet in cijfers tot uiting zou komen. Samen met het Cito heeft de inspectie dit onderzocht op basis van de examens uit 1981 tot en met 1991. De conclusie was dat er geen sprake was van daling van de prestaties, maar wel dat er zich grote verschillen voordeden in de moeilijkheid van de examens. Deze verschillen werden niet voldoende door de normering (cijfergeving) gecompenseerd. Daardoor zijn de examens over de jaren heen niet voldoende gelijkwaardig, in de zin dat gelijke prestaties met een gelijk cijfer worden gewaardeerd. Mede naar aanleiding van dit onderzoek is het Cito gestart met normhandvingsonderzoek (Alberts, 1997, 2001) dat vanaf 1994 moest leiden tot meer gelijkwaardige examens.

De examens bestaan uit twee onderdelen: het centraal examen (CE) dat voor alle leerlingen gelijk is, en een schoolonderzoek (SO) dat door de school wordt geconstrueerd en dus van school tot school kan verschillen. Pijl (1991) onderzocht de discrepantie tussen SO en CE. Hij voerde zijn onderzoek uit naar aanleiding van het examenverslag dat de inspectie in 1988 uitbracht. Daarin sprak de inspectie haar zorg uit over de grote verschillen tussen SO en CE, die bovendien niet voor alle scholen even groot waren. Pijl constateert op basis van de examengegevens uit 1990:

- dat er betrekkelijk kleine gemiddelde verschillen zijn tussen SO en CE;
- dat er wel grote verschillen zijn tussen scholen en
- dat deze verschillen niet verklaard kunnen worden door verschillen in de vaardigheden van de leerlingen die deze scholen bevolken.

Sinds 1995 beschikt de inspectie over de examenresultaten van individuele leerlingen. Vermeulen (2000) heeft deze gegevens geanalyseerd, waarbij hij vooral keek naar het

verloop over de jaren. Hij concludeert dat op het SO systematisch (iets) hogere cijfers worden gegeven dan op het CE en dat over de jaren heen het SO stabiel is. Verder blijkt de samenhang tussen SO en CE afhankelijk van het vak. Zijn analyse is op een 'round table' besproken met belanghebbenden (Inspectie van het Onderwijs, 2001). De voornaamste conclusies van deze conferentie zijn dat de gemiddelde verschillen tussen SO en CE geen reden zijn voor zorg, maar dat nader onderzoek nodig is naar de scholen die sterk afwijken. Dezelfde conclusies trok Pijl ook in 1991.

Recent heeft Dronkers (2000) aandacht besteed aan het schoolonderzoek. Hij vraagt zich af of de eindexamencijfers waaraan het schoolonderzoek een bijdrage levert, gelijkwaardig zijn tussen scholen. Zijn voornaamste conclusie is dat de becijfering van het schoolonderzoek van school tot school verschilt, ook als gecorrigeerd wordt voor een aantal schoolkenmerken. Hij beantwoordt de gestelde vraag dan ook negatief: de becijfering van het schoolonderzoek ondermijnt (...) de gelijkwaardigheid van het eindexamen.

Van Schooten en De Gloppe (2002) hebben voor de moderne vreemde talen onderzocht in hoeverre de gegevens van de centrale examens bruikbaar zijn voor de kwaliteitsbewaking op systeemniveau. Zij komen tot de conclusie dat voor de moderne vreemde talen aan die bruikbaarheid getwijfeld kan worden. Daarbij stellen ze ook het normhandvingsonderzoek, dat verricht wordt om tot over de jaren gelijkwaardige cijfers te komen, ter discussie.

In dit artikel beantwoorden we drie vragen die betrekking hebben op de vergelijkbaarheid van het schoolonderzoek en het centraal examen:

- 1 In hoeverre geven verschillen in de SO-cijfers tussen scholen en tussen jaren verschillen in de vaardigheid van de leerlingen op die scholen weer?
- 2 In hoeverre geven verschillen in de CE-cijfers in opeenvolgende jaren verschillen in de vaardigheid van de leerlingen weer?
- 3 Heeft het gebruik van het normhandvingsonderzoek dat sinds 1994 plaatsvindt, geleid tot een grotere gelijkwaardigheid van het CE-cijfer over jaren?

De vraag naar de vergelijkbaarheid van schoolonderzoek en centraal examen komt in dit artikel maar zijdelings aan de orde.

2 Schoolonderzoek en centraal examen

In deze paragraaf kijken we eerst naar de verschillen in de inhoud, de constructie en de beoordeling van het schoolonderzoek en het centraal examen. Daarna gaan we dieper in op de wijze waarop het SO-cijfer tot stand komt en de betekenis die dit cijfer daarmee heeft.

2.1 Inhoud van schoolonderzoek en centraal examen

De inhoud van het centraal examen en het schoolonderzoek ligt vast in de examenprogramma's die de overheid vaststelt. Voor sommige vakken is het programma voor het schoolonderzoek en het centraal examen hetzelfde, bijvoorbeeld wiskunde. Voor andere vakken heeft het schoolonderzoek betrekking op andere onderdelen dan het centraal examen. Zo zijn bij Nederlands spreken en luisteren onderdeel van het schoolonderzoek en is tekstbegrip onderdeel van het centraal examen.

Behalve dat de inhoud van schoolonderzoek en centraal examen kan verschillen, kan de inhoud ook van jaar tot jaar anders zijn; dat geldt met name voor aardrijkskunde en geschiedenis.

De verschillen in inhoud tussen schoolonderzoek en centraal examen zijn ten dele het gevolg van het feit dat sommige onderdelen moeilijk in een centraal examen getoetst kunnen worden. Zo zijn de praktische vaardigheden in de kunstvakken, de practica in de exacte vakken en de mondelinge vaardigheden in de talen onderdeel van het schoolonderzoek. Werkstukken zijn ook altijd onderdeel van het schoolonderzoek.

Het feit dat de inhoud van het schoolonderzoek en het centraal examen voor een groot aantal vakken niet dezelfde is, betekent dat in het algemeen voor die vakken de betekenis van het cijfer van SO en CE niet dezelfde is.

2.2 Constructie van schoolonderzoek en centraal examen

Tussen 1995 en 1999 heeft de inspectie (Inspectie van het Onderwijs, 1995, 1996, 1997, 1999) de kwaliteit van het schoolonderzoek voor een aantal vakken in verschillende opleidingen beoordeeld. Aandachtspunten daarbij waren de organisatie, de inhoud, de constructie en de beoordeling van de schoolonderzoeken. De belangrijkste conclusies zijn:

- Het aantal schoolonderzoeken loopt sterk uiteen. Afhankelijk van school en vak ligt dit aantal meestal tussen de drie en de tien.
- Inhoudelijk dekken de schoolonderzoeken vaak niet het gehele programma. Ook daarbij verschillen de scholen sterk van elkaar.
- De wijze waarop de schoolonderzoeken zijn geconstrueerd, laat vaak te wensen over.
- Het nakijken van de schoolonderzoeken gebeurt over het algemeen correct. Cijfers worden echter meestal pas achteraf vastgesteld.

Het schoolonderzoek als geheel is dus samengesteld uit een per school wisselend aantal unieke onderdelen die sterk verschillen in inhoud en kwaliteit.

Het centraal examen is één schriftelijke toets, die alle leerlingen van een opleidingstype op hetzelfde tijdstip en onder vergelijkbare omstandigheden maken. Het Cito construeert het centraal examen onder verantwoordelijkheid van de Commissie Vaststelling Examenopgaven (CEVO). De constructie is omgeven door een groot aantal kwaliteitsgaranties. De kwaliteit wordt steeds gecontroleerd en voldoet aan de gebruikelijke psychometrische eisen. Het Cito rapporteert daar jaarlijks over in haar examenverslag.

2.3 Beoordeling van schoolonderzoek en centraal examen

Het resultaat op een schoolonderzoek of een centraal examen is een waardering in de vorm van een cijfer. In principe komt dat cijfer in twee stappen tot stand. In eerste instantie vindt correctie van de opgaven plaats. Daarbij worden aan de antwoorden punten

toegekend die aangeven in welke mate de opgave goed is gemaakt. De som van deze punten vormt de score. De regels die gebruikt worden voor de correctie vormen het correctievoorschrift. In tweede instantie wordt de score omgezet in een cijfer van 1 tot en met 10, meestal afgerond op één decimaal. Het normeringsvoorschrift is de regel die aangeeft welk cijfer bij een bepaalde score hoort.

Voor het nakijken van het schoolonderzoek is meestal een correctievoorschrift beschikbaar. Voor het geven van het cijfer wordt in de regel geen expliciet normeringsvoorschrift gebruikt. Voor zover beschikbaar wordt dit meestal na het nakijken vastgesteld.

Het schoolonderzoek levert meerdere cijfers op. Het uiteindelijke cijfer op het schoolonderzoek is over het algemeen een gewogen gemiddelde over de verschillende onderdelen. Zo telt het laatste onderdeel soms zwaarder dan de andere onderdelen. Vaak hebben leerlingen ook de mogelijkheid om één of meer schoolonderzoeken te herkansen. Bij het centraal examen vindt de correctie plaats op basis van een vastgesteld correctievoorschrift. Bij examens met open vragen kijken docenten het examen na. Voor vrijwel alle examens met open vragen vindt een tweede correctie plaats door een leraar van een andere school. Examens met meerkeuzevragen werden tot 2000 op het Cito nagekeken.

De CEVO stelt de norm vast die gebruikt wordt om de cijfers voor de centrale examens te berekenen. In de hier onderzochte periode werden twee procedures gehanteerd (Alberts, 1999). De eerste procedure geldt voor examens met uitsluitend gesloten opgaven, zoals de moderne vreemde talen. Voor die examens stelt de CEVO de norm vast na afname van het examen. Daarbij wordt, indien beschikbaar, gebruikgemaakt van een normhandhavingsonderzoek, de reacties van docenten en de gegevens over de resultaten van het betrokken jaar in vergelijking met de resultaten in eerdere jaren. Het normhandhavingsonderzoek is bedoeld om tot gelijkwaardige cijfers in verschillende jaren te komen. Verderop in het artikel gaan we hierop in. De tweede procedure wordt gehanteerd bij de examens die geheel of gedeeltelijk uit open vragen bestaan. Bij de constructie van deze examens wordt gestreefd naar een cesuur bij 55 van de

100 maximaal toe te kennen punten. Bij deze constructie worden, indien beschikbaar, ook de resultaten van het normhandhavingsonderzoek gebruikt. De definitieve norm wordt vastgesteld na analyse van de resultaten van een steekproef van leerlingen en de reactie van docenten. De cesuur wordt aangepast als er fouten zijn in het examen of het correctievoorschrift of als de resultaten aanzienlijk slechter zijn dan in de voorafgaande jaren.

2.4 Betekenis van cijfers op het schoolonderzoek

Dronkers (2000) noemt een aantal redenen waarom de cijfers voor het schoolonderzoek en het centraal examen van elkaar zouden afwijken. Twee ervan hebben te maken met het verschil in inhoud en het aantal metingen. Daarop zijn we hierboven al ingegaan. Als andere punten noemt Dronkers de relatieve normering binnen scholen, de cijfercultuur binnen een school en binnen vakken, en de concurrentie tussen scholen. Op deze punten gaan we hier verder in.

Posthumus heeft al in 1940 betoogd dat de prestaties van leerlingen vaak niet op zich worden beoordeeld, maar in vergelijking met prestaties van de andere leerlingen, in ons geval van leerlingen die hetzelfde schoolonderzoek hebben gemaakt. De hoogste en laagste scores bepalen in hoge mate wat als goed en slecht wordt beschouwd. Verder ligt het voor de hand dat, onder normale omstandigheden, een gemiddelde leerling een voldoende haalt. In deze situatie ontstaat de veel voorkomende (scheve) verdeling van vrijwel alle cijfers tussen de 3 en de 8 met een gemiddelde tussen de 6 en de 7. Het is overigens niet zo dat bewust naar zo'n verdeling wordt gestreefd, maar resultaten die daarvan afwijken worden gesignaleerd en zo nodig gecorrigeerd. Bij een schoolonderzoek dat desastreus is gemaakt, kan bijvoorbeeld een punt extra worden gegeven, of het kan gevolgd worden door een aanzienlijk gemakkelijker schoolonderzoek.

Scholen vergelijken niet alleen leerlingen, maar ook jaargangen. Een heel andere uitkomst van een schoolonderzoek in vergelijking met voorafgaande jaren zal opvallen en, als er geen duidelijk aanwijsbare oorzaak is,

leiden tot bijstellingen. Dit leidt over de jaren heen tot stabiele resultaten.

Een derde vergelijking die de cijfergeving mede bepaalt, is de vergelijking met andere scholen. Voor zover scholen zich vergelijkbaar achten met andere scholen, zullen afwijkende resultaten opvallen en eventueel tot actie leiden. Dronkers noemt hierbij de concurrentieverhoudingen tussen scholen.

Naast relatieve normering en vergelijking met andere jaren en andere scholen, speelt bij het geven van cijfers ook de traditie een belangrijke rol. Vakken verschillen aanzienlijk in het gemiddelde cijfer en daarmee samenhangend in het percentage voldoende. Dit geldt zowel voor het SO als het CE. Zo zijn de prestaties op Nederlands en Engels voor alle opleidingen aanzienlijk beter dan voor wiskunde of economie. Dat ligt niet aan de aard van de vakken of aan de begaafdheid van de leerlingen die deze vakken kiezen. Zo worden in het vwo wiskunde B en economie II gekozen door de meest begaafde leerlingen. Toch zijn dit de vakken waarbij zowel voor het SO als het CE de meeste onvoldoendes vallen. De Groot heeft in zijn onvolprezen "Vijven en Zessen" (De Groot, 1966) al uitgelegd dat dit vooral te maken heeft met de status van een vak en met de invloed die een vak heeft op selectiebeslissingen zoals overgang en diplomering. Wat de verklaring ook moge zijn, de traditie van cijfergeving binnen een vak bepaalt in belangrijke mate de normen die in het schoolonderzoek (en het centraal examen) worden gehanteerd.

Binnen algemene kaders is de betekenis van het schoolonderzoekcijfer dus schoolgebonden. Als kaders functioneren de schoolpopulatie, de prestaties op andere scholen, de resultaten op de centrale examens en - niet in de laatste plaats - de traditie binnen de vakken. Dit alles leidt tot voor de school hanteerbare en over de jaren stabiele uitkomsten, die echter niet vergelijkbaar zijn met andere scholen, en het verschil tussen opeenvolgende generaties leerlingen kunnen maskeren.

2.5 Conclusies

Voor het SO-cijfer hebben we geconstateerd dat het gebaseerd is op een - per school unieke - set onderdelen. Het aantal daarvan wisselt per school en er zijn grote kwaliteitsverschil-

len. Verder is de cijfergeving ondoorzichtig. Er zijn redenen om aan te nemen dat de gehanteerde normen sterk beïnvloed worden door de schoolsamenstelling en dat ook prestaties in eerdere jaren en van andere scholen als vergelijkingsmaatstaf een rol spelen. De conclusie is dan ook dat het SO-cijfer een schoolgebonden waardering is voor de prestaties van leerlingen die niet gebruikt kan worden voor de vergelijking van scholen of van prestaties in verschillende jaren.

Het centraal examen is een goed geconstrueerd instrument met bekende eigenschappen. De scoring van de opgaven en de toegepaste normering zijn voor alle kandidaten gelijk. Het CE-cijfer is daarmee bruikbaar om scholen in een bepaald jaar met elkaar te vergelijken. De wijze waarop de normen worden vastgesteld, bieden echter geen garantie voor de gelijkwaardigheid van de CE-cijfers van verschillende jaren. Weliswaar beschikt de CEVO sinds 1994 over de resultaten van het normhandavingsonderzoek door het Cito, maar de wijze waarop dit wordt gehanteerd en het gebruik van andere informatie doen afbreuk aan de transparantie van de procedure. Met name het feit dat in de onderzochte periode alleen gecorrigeerd werd in het voordeel van de leerlingen, vormt een bedreiging van de vergelijkbaarheid.

3 Analyse

3.1 Gegevens

Voor de analyse is gebruikgemaakt van twee datasets. De eerste dataset bestaat uit de schoolgemiddelden van alle scholen van 1990 tot en met 2000. De tweede dataset bestaat uit de SO- en CE-cijfers van individuele leerlingen. De CE-cijfers betreffen de resultaten op het eerste tijdvak. In 1995 heeft de inspectie een zelfwegende steekproef van havo- en vwo-leerlingen getrokken. In 1996 betrof dat een zelfwegende steekproef van mavo-leerlingen en bijna de gehele populatie van vbo-leerlingen. Vanaf 1998 beschikt de inspectie over de integrale gegevens die elektronisch door de Informatie Beheer Groep zijn verzameld en verwerkt.

Om de bestanden met de individuele examengegevens vanaf 1995 geschikt te maken

Tabel 1

Aantallen leerlingen in de populatie, in de steekproef en in feite betrokken bij de analyse

Examenjaar	Integrale leerlingentelling	Steekproef		Analyse	
		N	%	N	%
1995	41571	20397	49	19021	46
1998	42930	40575	95	39794	93
1999	44537	40529	90	40247	90
2000	41898	40131	96	39369	94

voor de analyses, moesten een aantal vakken met bijbehorende cijfergegevens en een aantal leerlingen met bijbehorende vakken en cijfergegevens worden verwijderd. Dit gold in het bijzonder voor het vbo, omdat bij deze opleiding vakken mogen worden afgesloten op A- en B-niveau, waarvoor geen centraal examen hoeft te worden afgelegd.

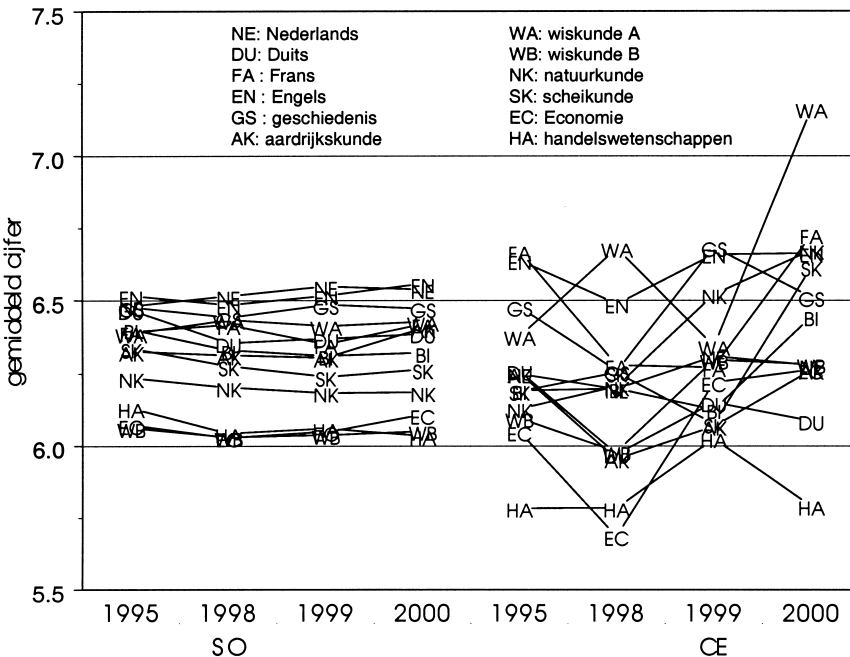
In dit artikel presenteren we alleen de resultaten voor het havo. De resultaten voor de andere opleidingen staan in Van den Bergh, Rohde en Swarts (2003). Deze leiden tot dezelfde conclusies als in dit artikel worden getrokken. In Tabel 1 staat per jaar aangegeven welke deel van de havo-populatie in de steekproef is betrokken en welk deel is geanalyseerd.

3.2 Gelijkwaardigheid over jaren

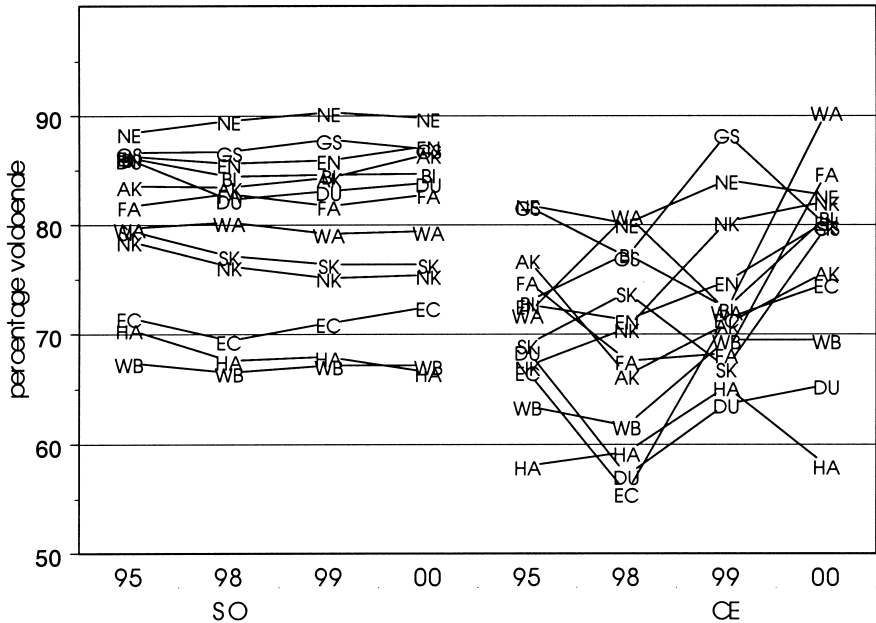
We hebben ons de vraag gesteld of de verschillen in CE-cijfers van opeenvolgende jaren, verschillen in vaardigheid van de leerlingen weergeven. We beantwoorden die vraag aan de hand van de tweede dataset, omdat deze individuele leerlingengegevens bevatten.

Figuur 1 geeft een indruk van het verloop van de SO- en CE-cijfers over de betrokken jaren. De figuur bevat de gemiddelden van de meest gekozen vakken in het havo. De overeenkomstige gemiddelden zijn door een lijn verbonden.

Uit de figuur blijkt duidelijk dat de gemiddelde SO-cijfers minder variëren dan de CE-cijfers. Alle SO-gemiddelden liggen tus-



Figuur 1. Gemiddeld SO- en CE- cijfer voor havo.



Figuur 2. Percentage voldoende op SO en CE voor havo.

sen 6.0 en 6.6. Het gemiddelde voor het CE varieert van 5.6 tot 7.2. Verder is de stabiliteit over de jaren heen voor het SO groter dan voor het CE. Voor het SO lopen de lijnen die de vakken verbinden vrijwel parallel en blijft de volgorde van de vakken over de jaren redelijk in tact. De CE-gemiddelden laten dalende en stijgende, elkaar kruisende lijnen zien.

Zoals Figuur 2 laat zien, veranderen deze conclusies niet als in plaats van de gemiddelden het percentage voldoende als maatstaf wordt gebruikt. De percentages voor het SO liggen dicht bij elkaar en lopen vrijwel parallel. Voor het CE zijn er grote verschillen tussen de vakken in het percentage voldoende, tot bijna 20% toe.

Voor de analyse van deze gegevens hebben we meerniveau-analyse gebruikt (Goldstein, 1995). Bij deze statistische techniek wordt rekening gehouden met de aard en de structuur van de gegevens. In ons geval vormen jaren en scholen een 'random' kruisclassificatie en zijn leerlingen genest binnen scholen en jaren; een leerling doet maar op één school in enig jaar examen. Het model is meer formeel weergegeven in Appendix I.

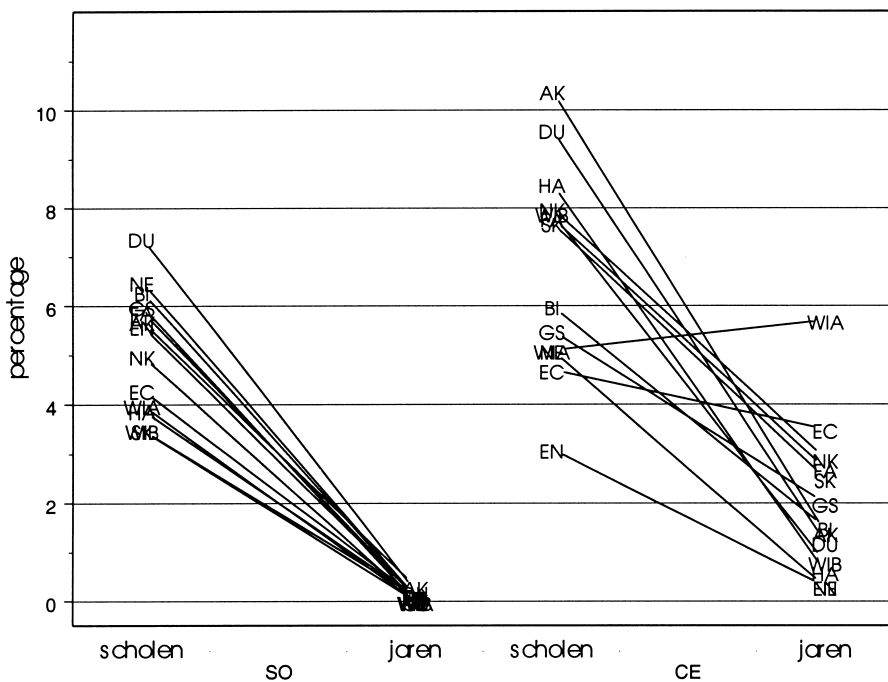
In een statistisch model worden parameters geschat die een compacte beschrijving

geven van de data en is het mogelijk om hypothesen over deze parameters te toetsen. De parameters in ons model zijn de gemiddelden over jaren en de varianties tussen leerlingen, scholen en jaren. De schattingen van deze parameters staan in Appendix II.

Voor de beantwoording van onze vraag naar de vergelijkbaarheid van de examens zijn vooral de varianties van belang, met name van de percentages die toegeschreven kunnen worden aan de verschillen tussen scholen en aan de verschillen tussen jaren. In Figuur 3 zijn deze gegevens weergegeven voor de SO- en de CE-cijfers.

In Figuur 3 wordt kwantitatief weergegeven wat al bleek uit Figuur 1: voor het schoolonderzoek zijn de verschillen tussen jaren verwaarloosbaar, voor het centraal examen is dat duidelijk niet het geval. Weliswaar zijn ook daar vakken waarbij de jaren minder dan 1% van de variantie voor hun rekening nemen, maar voor veel vakken ligt het percentage in de orde van 3 à 4% met uitschieters tot bijna 6% bij wiskunde A.

De vraag is nu of het aannemelijk is dat de verschillen tussen jaren toegeschreven kunnen worden aan verschillen in vaardigheden van leerlingen in opeenvolgende jaren. Zijn de eindexamenkandidaten op het havo in



Figuur 3. Percentage variantie tussen scholen en jaren voor het CE- en SO-cijfer voor havo.

2000 veel beter in wiskunde A, of is het examen in dit jaar wellicht makkelijker geweest?

We mogen veronderstellen dat de gemiddelde vaardigheid van leerlingen van jaar tot jaar maar weinig zal verschillen als de inhoud en de vorm van het onderwijs niet grondig veranderen. In de onderzochte periode zijn de examenprogramma's voor vrijwel alle vakken niet of nauwelijks veranderd. Ook didactisch hebben zich geen opvallende ontwikkelingen voorgedaan. De grote veranderingen die plaatsvonden met de invoering van de tweede fase in havo en vwo, en van de leerwegen in het vmbo hebben pas na 2000 hun beslag gekregen. In 2000 zijn alleen voor de voorhoedescholen havo nieuwe examens afgenomen. De leerlingen die daarbij zijn betrokken, zijn niet in de analyse meegenomen. We mogen dus verwachten dat de verschillen tussen jaren klein zullen zijn.

Door het onderzoek naar de effectiviteit van scholen (Blok & Eiting, 1988) weten we dat betrekkelijk kleine percentages van de totale variantie (i.c. de intraklassecorrelatie) al op substantiële effecten wijzen. Voor 6 van de 13 vakken die in de analyse zijn betrokken, is de variantie tussen jaren meer dan 2%

van de totale variantie; bij één vak is dat meer dan 5%. We constateren dan ook dat de verschillen tussen de jaren aanzienlijk groter zijn dan verwacht mag worden in een onveranderde omgeving; de verschillen in prestaties in verschillende jaren lijken groter dan verklaard zou kunnen worden door verschillen in vaardigheid van leerlingen.

We kunnen de verschillen tussen jaren ook vergelijken met de verschillen tussen scholen. Omdat de centrale examens voor alle scholen in alle jaren dezelfde zijn, kunnen deze verschillen geïnterpreteerd worden als verschillen in kwaliteit tussen de scholen. Figuur 3 laat zien dat die verschillen voor bijna alle vakken meer dan 4% van de totale variantie verklaren. Verschillen tussen jaren van dezelfde grootte zou betekenen dat het verschil tussen het jaar met de hoogste en laagste score even groot is als het verschil tussen de school met - gemiddeld over jaren - de beste en de slechtste prestaties. Voor zes vakken is de variantie tussen jaren meer dan eenderde van de variantie tussen scholen, voor twee ervan is de variantie meer dan tweederde van de variantie tussen scholen. Het is uiterst onwaarschijnlijk dat zich zulke

vaardigheidsverschillen voordoen, terwijl er maar weinig in vorm en inhoud van het onderwijs verandert.

Er zijn meer redenen om te twijfelen aan de veronderstelling dat de verschillen tussen jaren, verschillen in vaardigheidsniveau zijn. Onder stabiele omstandigheden zullen veranderingen in vaardigheden zich geleidelijk voltrekken. Figuur 1 en Figuur 4 laten zien dat er voor vrijwel alle vakken geen lijn zit in de jaargemiddelden. Ook lijkt het redelijk, aan te nemen dat bij vakken met een soortgelijke inhoud en onderwijsvorm, zoals de moderne vreemde talen, veranderingen zich op een gelijke wijze voltrekken. Ook deze veronderstelling wordt door de feiten gelogenstraft.

Al met al is het niet erg waarschijnlijk dat de verschillen tussen jaren uitsluitend of zelfs voor een groot deel toegeschreven kunnen worden aan verschuivingen in de vaardigheden van leerlingen. Ook andere factoren moeten een rol spelen. De factor die het meest voor de hand ligt, is het verschil in moeilijkheid van de examens en de wijze waarop de cijfers worden toegekend.

3.3 Effect van moeilijkheid op kans van slagen

Het cijfer van het centraal examen op één vak is maar één van de minimaal 12 cijfers (zes SO en zes CE) die een rol spelen bij de beslissing of een leerling wel of niet slaagt en zijn diploma krijgt. Het is dan ook de vraag of verschuivingen in de moeilijkheid van het examen voor een enkel vak wel invloed hebben op het al of niet slagen voor een examen. Om hier zicht op te krijgen, is voor de leerlingen die in 2000 examens deden de uitslag

Tabel 2
Percentage leerlingen dat meer of minder slaagt bij verschuiving van het cijfer op het CE met een half punt, en aantal leerlingen dat examen doet in het vak

Vak	Verschuiving		Aantal leerlingen
	+ .5	- .5	
Nederlands	+ .9	- 3.0	39357
Engels	+ 1.0	- 2.5	38753
Wiskunde A	+ .9	- 2.6	27362
Wiskunde B	+ 1.0	- 3.7	9986
Geschiedenis	+ .8	- 2.7	12092

herberekend, nadat voor één vak het cijfer op het CE met een half punt is verhoogd of verlaagd.

In Tabel 2 staan de verschuivingen in de slagingspercentages van de leerlingen die in het betrokken vak examen hebben gedaan. Een plusteken voor het percentage betekent een stijging van het percentage geslaagden, een minteken betekent een daling.

Uit de tabel blijkt dat een verschuiving van een half punt duidelijk effect heeft op het percentage leerlingen dat slaagt voor het havo. Als het cijfer Nederlands een half punt opgevoerd wordt, dan slaagt ('ceteris paribus') 0.9%, oftewel ruim 350 leerlingen extra. Een negatieve verschuiving van een half punt bij Nederlands heeft tot gevolg dat (ceteris paribus) 3.0% minder slaagt. Dat wil zeggen: bijna duizend leerlingen zouden dan hun havo-diploma niet behaald hebben.

Er kan dus geconcludeerd worden dat kleine verschillen in normstelling tussen examens - in moeilijkheid of in spreiding - grote gevolgen kunnen hebben in termen van aantallen leerlingen. Natuurlijk betreft het hier een enigszins theoretische vergelijking, want bij een echt examen gaat het niet om verschuivingen voor één vak en ook niet altijd om een verschil van een half punt. Voor elk vak zal er een verschil in moeilijkheid zijn met het voorafgaande jaar dat niet aan verschil in vaardigheid toegeschreven kan worden. Sommige van die afwijkingen zullen positief zijn en andere negatief. Over het geheel genomen zullen deze verschuivingen elkaar uitmiddelen. Er zal er dan ook maar weinig verandering optreden in het percentage leerlingen dat slaagt. Het zal dan echter voor een niet onaanzienlijk deel om andere leerlingen gaan: voor de slagingskans van een individuele leerling maakt het wel degelijk uit in welk jaar hij of zij examen doet. Wel wordt hiermee aangegeven dat kleine wijzigingen in de moeilijkheid grote effecten kunnen hebben.

3.4 Effecten van het normhandavingsonderzoek

De derde vraag die we stellen, is of de invoering van normhandavingsonderzoek geleid heeft tot grotere gelijkwaardigheid van de CE-cijfers in verschillende jaren. Het normhandavingsonderzoek is opgezet als steun

bij het bepalen van een grens tussen voldoende en onvoldoende die gelijkwaardig is met de grens op voorafgaande examens, zodat gelijke prestaties gelijk gewaardeerd worden. De procedure die het Cito daarvoor hanteert, staat in Appendix III. Daarin staat ook voor welke vakken en vanaf welk moment de resultaten van normhandavingsonderzoek zijn gebruikt bij het vaststellen van de normen.

Verondersteld mag worden dat de procedure voor de vakken die het betreft, geleid heeft tot afname van de verschillen tussen jaren in gemiddeld cijfer. We hebben deze veronderstelling onderzocht op basis van de schoolgemiddelden voor de periode 1990 tot en met 2000.

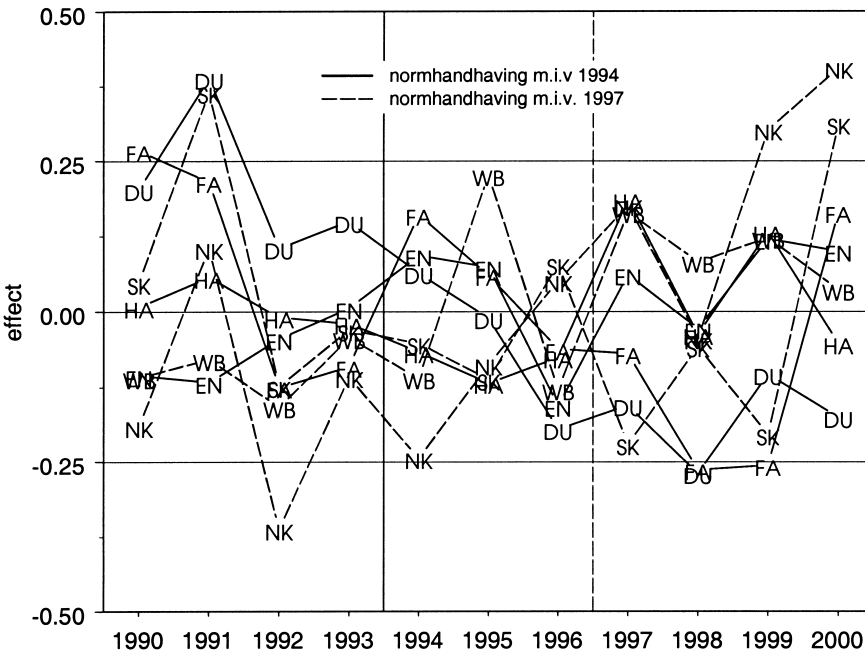
Voor de analyse van het effect hebben we dataset 1 gebruikt en hetzelfde model als voor de voorafgaande analyse, zij het dat er nu geen sprake is van variantie tussen leerlingen. Het model staat als Model II in Appendix I. Om na te gaan of de verschillen zijn afgenomen, kijken we naar het jaareffect als afwijking van het algemeen gemiddelde. In Figuur 4 staan deze effecten afgebeeld voor de vakken waarbij normhandavingsonderzoek is gebruikt. In de figuur geven de verticale strepen het moment aan waarbij de

procedure is ingevoerd. Voor de moderne vreemde talen is dat 1994, voor de andere vakken 1997.

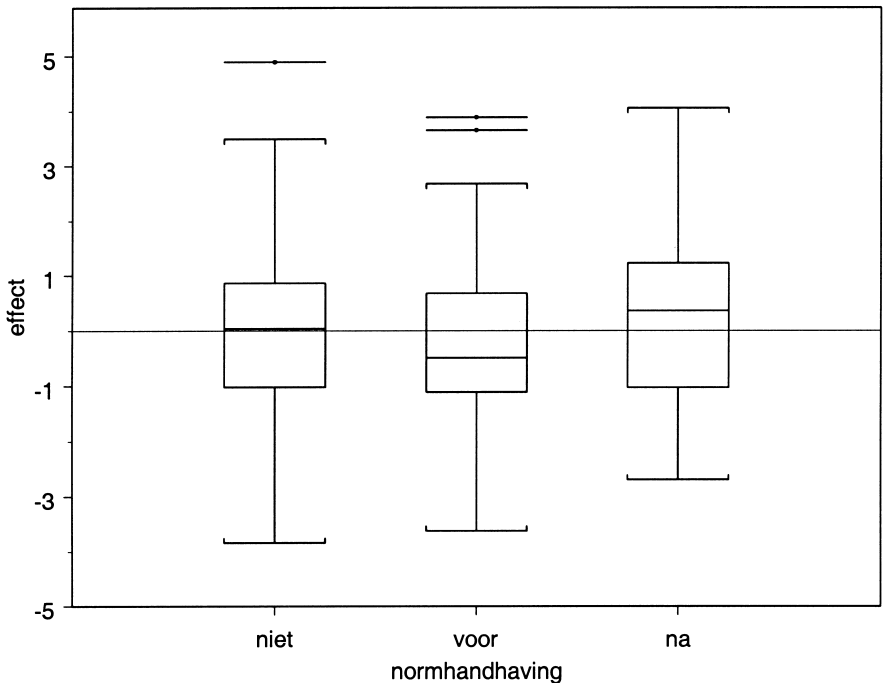
Uit Figuur 4 is niet zonder meer duidelijk dat vanaf 1994 of 1997 de jaareffecten kleiner zijn dan daarvoor. Figuur 5 geeft daarover meer duidelijkheid. In deze figuur zijn door middel van 'boxplots' de afwijkingen van de jaargemiddelden van het algemeen gemiddelde (jaareffecten) aangegeven voor de vakken waarbij wel of geen normhandavingsonderzoek heeft plaatsgevonden. Voor de vakken waarbij de procedure wel is toegepast, is onderscheid gemaakt tussen de jaren voor en na invoering.

In de figuur geeft de rechthoek (box) aan binnen welk bereik 50% van de effecten valt. De horizontale lijn die de rechthoek in tweeën deelt geeft de plaats van de mediaan aan. De haken geven aan binnen welk bereik het grootste deel van de effecten ligt. De lijnen daarboven en daaronder vormen uitschieters.

Uit de figuur blijkt dat er nauwelijks verschillen zijn in de verdelingen van de jaareffecten, noch tussen vakken met of zonder normhandavingsonderzoek, noch voor of na de invoering van deze procedure.



Figuur 4. Jaareffecten voor havo.



Figuur 5. Boxplots van de jaareffecten voor havo.

4 Conclusies en discussie

Bij het onderzoek zijn we uitgegaan van drie functies die het examen zou moeten vervullen:

- de paspoortfunctie voor de leerlingen;
- de kwaliteitsborging voor de scholen, en
- de sturingsfunctie voor de overheid.

Het realiseren van deze functies veronderstelt de gelijkwaardigheid van cijfers: eenzelfde prestatie moet met eenzelfde cijfer gewaardeerd worden.

Aan de hand van twee vragen hebben we onderzocht of de cijfers voor schoolonderzoek en centraal examen gelijkwaardig zijn. We zijn tot de conclusie gekomen dat voor het schoolonderzoek de cijfers van verschillende scholen en verschillende jaren niet gelijkwaardig zijn. Voor het centraal examen zijn de cijfers van leerlingen op verschillende scholen in eenzelfde examenjaar wel gelijkwaardig. Het is echter niet aannemelijk dat de cijfers in verschillende jaren gelijkwaardig zijn. We hebben ook geconstateerd dat het normhandhavingsonderzoek dat sinds 1994 voor een aantal vakken wordt uitgevoerd niet aantoonbaar heeft bijgedragen aan een verbetering van de gelijkwaardigheid.

In termen van de genoemde functies betekenen deze conclusies dat het voor de kans van slagen van leerlingen uitmaakt in welk jaar ze examens doen. Het paspoort is het ene jaar iets anders waard dan het andere.

Voor scholen is het schoolonderzoek niet geschikt voor kwaliteitscontrole: vergelijkingen met voorafgaande jaren of andere scholen is niet goed mogelijk. Het centraal examen staat vergelijking met andere scholen wel toe, maar de uitkomsten kunnen niet vergeleken worden met voorafgaande jaren. Omdat de jaren niet betrouwbaar met elkaar vergeleken kunnen worden, is het centraal examen niet bruikbaar om de ontwikkeling van het prestatieniveau te volgen en kan de overheid daarop dus ook niet sturen.

4.1 Discussie

Door de aard van de beschikbare gegevens hebben we niet bewezen dat de verschillen tussen jaren niet aan veranderingen in vaardigheid zijn toe te schrijven. We hebben wel aangetoond dat dit niet aannemelijk is. Voor een bewijs is het nodig dat eindexamenleerlingen examens van verschillende jaren maken, zodat exact te bepalen is wat aan verschil in

moelijkheid en wat aan verschil in vaardigheid is toe te schrijven. Een dergelijk onderzoek heeft de inspectie samen met het Cito in 1992 uitgevoerd. Om zekerheid te krijgen over verschillen in de examens zou zo'n onderzoek opnieuw uitgevoerd moeten worden.

We kunnen evenmin bewijzen dat het normhandavingsonderzoek niet leidt tot gelijkwaardigheid van de cijfers in verschillende jaren. Er is echter geen enkele indicatie dat dit wel het geval is.

In de onderzochte periode is bij het vaststellen van de norm alleen rekening gehouden met de prestaties van de leerlingen als dit niet zou leiden tot verhoging van het percentage onvoldoendes. Dit houdt in dat voor te gemakkelijk examens niet, of niet in voldoende mate is gecorrigeerd. Sinds 2001 heeft de CEVO ook de mogelijkheid te corrigeren voor, in haar ogen, te gemakkelijk examens. Het is mogelijk dat deze verruiming tot een grotere stabiliteit zal leiden.

4.2 Mogelijkheden voor verbetering

De aard van het schoolonderzoek als schoolgebonden waardering van de prestaties van leerlingen brengt beperkingen met zich mee die het vrijwel onmogelijk maken een vorm te vinden die het mogelijk maakt om ze te gebruiken voor de kwaliteitscontrole door scholen. Uiteraard zijn inhoudelijke en psychometrische verbeteringen mogelijk, maar dat zal alleen leiden tot een betere beoordeling van de leerlingen in een school, maar niet tot een grotere gelijkwaardigheid van de cijfers tussen scholen.

Voor het centraal examen is het niet gemakkelijk om aan te geven of er verbeteringen mogelijk zijn. Daarvoor spelen te veel factoren een rol en is een aantal aspecten te ondoorzichtig. Voor een goede vergelijkbaarheid zijn de volgende aspecten van belang:

- 1 De omschrijving van de exameneisen in het examenprogramma moet voldoende informatie bevatten om eenduidig toetsen op te baseren.
- 2 De vertaling van het examenprogramma in examenopgaven en de samenstelling van het examen moet, bijvoorbeeld in de vorm van een toetsmatrijs, duidelijk zijn vastgelegd. Dit moet leiden tot inhoudsvalide examens die nominaal parallel zijn.

- 3 De correctievoorschriften moeten een objectieve scoring mogelijk maken.
- 4 Door middel van onderzoek moet bepaald zijn wat de verschillen in moeilijkheid zijn.
- 5 Het normeringsvoorschrift moet transparant zijn en corrigeren voor de verschillen in inhoud en moeilijkheid van het examen.

Ons onderzoek laat uitspraken over de eerste drie aspecten niet toe. Het is echter wel zo dat bij vakken waarbij het examenprogramma van jaar tot wisselt, zoals aardrijkskunde en geschiedenis, het vrijwel onmogelijk is om tot gelijkwaardige cijfergeving te komen.

Ons onderzoek laat ook geen uitspraken toe over het normhandavingsonderzoek dat het Cito uitvoert. We hebben alleen geconstateerd dat de invoering ervan niet heeft geleid tot afname van de verschillen tussen jaren. Duidelijk is wel dat het normhandavingsonderzoek niet onder optimale omstandigheden plaatsvindt. Onduidelijk is of deze beperkingen zodanig afbreuk doen aan de uitkomsten dat een schatting van de verschillen in moeilijkheid niet goed mogelijk is, maar uitgesloten is dat evenmin. Bovendien kan de gehanteerde normhandavingsprocedure niet het gewenste eindresultaat - gelijke monniken, gelijke kappen - opleveren, omdat men zich slechts richt op de cesuur van 5.5. Verschillen in variantie in moeilijkheid van de examenopgaven leiden tot een verandering die niet in de normeringsprocedure meegenomen wordt, maar die wel degelijk leidt tot andere cijfers voor leerlingen met een gelijk vaardigheidsniveau.

De procedure voor het vaststellen van de normen is op zich helder omschreven (Melissen, 2002). Het bevat echter een aantal subjectieve elementen die de uitvoering minder doorzichtig maakt. Dat betreft op de eerste plaats het gebruik van reacties uit het veld. Het is waarschijnlijk dat bij deze reacties ook andere factoren dan inhoudelijke fouten en problemen mee zullen spelen. Op de tweede plaats betreft dit het bijstellen van de cesuur als het percentage onvoldoendes te ver afwijkt van de verwachting. Voor de vakken waarbij normhandavingsonderzoek is verricht, betekent dit in feite een doublure die de helderheid niet bevordert.

Literatuur

- Alberts, R. V. J. (1997). *Verslag normhandhaving examens vbo/avo/vwo 1997*. Arnhem: Cito.
- Alberts, R. V. J. (1999). *Examenverslag vbo/mavo/havo/vwo*. Arnhem: Cito.
- Alberts, R. V. J. (2001). Equating exams as a prerequisite for maintaining standards: experience with Dutch centralised secondary examinations. *Assessment in Education*, 8, 353-367.
- Bergh, H. van den, Rohde, E., & Zwarts, M. (2003). *Schoolonderzoek en centraal examen*. Utrecht: Inspectie van het Onderwijs.
- Blok, H., & Eiting, M. H. (1988). De grootte van schooleffecten: Hoe verschillend presteren leerlingen van verschillende scholen. *Tijdschrift voor Onderwijsresearch*, 13, 16-30.
- Dronkers, J. (2000). *Is het eindexamen wel gelijkwaardig tussen scholen? Discrepantie tussen de cijfers voor schoolonderzoek en centraal examen in het voortgezet onderwijs*. Paper. Amsterdam: Universiteit van Amsterdam.
- Glas, C. A. W., & Béguin, A. A. (1999). Normhandhaving bij examens in het voortgezet onderwijs: een toepassing van item response theorie. In J. M. Pieters, T. Plomp, & L. E. Oden-thal, *20 jaar Toegepaste Onderwijskunde*. Enschede: Twente University Press.
- Goldstein, H (1995). *Multilevel statistical models*. Londen: Edward Arnold.
- Groot, A. D. de (1996). *Vijven en zessen. Cijfers en beslissingen: het selectieproces in ons onderwijs*. Groningen: Wolters.
- Inspectie van het Onderwijs (1992). *Examens op punten getoetst*. Zoetermeer: Inspectie van het Onderwijs.
- Inspectie van het Onderwijs (1995). *Het schoolonderzoek onderzocht*. Utrecht: Inspectie van het Onderwijs.
- Inspectie van het Onderwijs (1996). *Het schoolonderzoek onderzocht II*. Utrecht: Inspectie van het Onderwijs.
- Inspectie van het Onderwijs (1997). *Het schoolonderzoek onderzocht III*. Utrecht: Inspectie van het Onderwijs.
- Inspectie van het Onderwijs (1999a). *Het schoolonderzoek onderzocht IV*. Utrecht: Inspectie van het Onderwijs.
- Inspectie van het Onderwijs (2001). *Round table exams. Verslag van een bijeenkomst over examens voortgezet onderwijs*. Utrecht: Inspectie van het Onderwijs.
- Melissen, M (2002). *De normering van de centrale examens. De besluitvormingsprocedure*. Inleiding op de expertmeeting 'Normhandhaving centrale examens voortgezet onderwijs'.
- Posthumus, K (1940). Middelbaar onderwijs en schifting. *De Gids*, 104, 24-42.
- Pijl, Y. J. (1991). *Discrepantie tussen het schoolonderzoek en het centraal schriftelijk examen in havo en vwo*. Groningen: Gion.
- Schooten, E. van, & Glopper, K. de (2002). Dalen de leerlingprestaties op de centraal schriftelijke examens Duits, Engels en Frans in mavo, havo en vwo? *Pedagogische Studiën*, 79, 5-17.
- Venables, W. N., & Ripley, B. D. (1997). *Modern applied statistics with S-PLUS*. New-York: Springer.
- Vermeulen, C. (2000). *Discrepantie tussen so-cijfer en CE-cijfer in het voortgezet onderwijs*. Utrecht: Inspectie van het Onderwijs.

Auteurs

Huib van den Bergh is werkzaam als universitair hoofddocent bij het Utrechts Instituut voor Linguïstiek van de Universiteit van Utrecht.

Elma Rohde is werkzaam als statistisch analist bij de afdeling Analyse & Macrorapportage van de Inspectie van het Onderwijs.

Michel Zwarts is werkzaam als inspecteur kennis bij de Inspectie van het Onderwijs

Correspondentieadres: Huib van den Bergh, Instituut voor Linguïstiek Universiteit Utrecht, Trans 10, 3512 JK Utrecht, e-mail: huib.vandenbergh@let.uu.nl

Abstract

On the equivalence of central and school exams in the Netherlands

In this article the question is posed whether it is sensible to compare grades for school exams and central exams over years. As the content for school exams differ from school to school, as well as the way they are rated, there is a severe limitation in the comparison of the grades of students from different schools. As school exams differ from year to year, a comparison over years does not seem to be warranted. Nevertheless, school exams do produce highly stable grades; the mean grade does hardly change between years. For the central exams, which are the same for students of each school type (track) a comparison of grades within years seems a worthwhile en-

terprise. For many subjects a standard setting procedure is used to make grades over years comparable. However, the differences between years still appear to be (relatively) large. The differences between exams are much larger than can be expected, due to either sample fluctuations, or to differences in the population of students in different years. Hence, the standard setting procedures do not seem to function optimally, and consequently in some years it is easier to pass the exams, whereas in other years it might be more difficult to pass. Hence, the same grade does indicate different abilities in different years. Hence, the grades of students in different years can hardly be compared sensibly.

Although the differences in grades between years seem to be small, they nevertheless indicate substantial effects, in terms of number of students failing or passing the exams.

Appendix I

Multiniveaumodellen

Model I

Het model betreft de score (SO of CE) op enig vak van een leerling i , op school j in examenjaar k . Elke leerling is dus genest binnen de combinatie van school en examenjaar. Stel, $y_{i(jk)}$ is de score van leerling i ($i = 1, 2, \dots, N_{jk}$) op school j ($j = 1, 2, \dots, J$) in jaar k ($k = 1, 2, \dots, K$). Het model kan beschreven worden als: $y_{i(jk)} = b_0 + [e_{i(jk)} + u_{00(jk)} + u_{10j0} + u_{200k}]$.

De 'fixed' parameter (b_0) is een schatting van het algemeen gemiddelde. In het random deel van het model (tussen vierkante haken) worden vier residuele termen onderscheiden: $e_{i(jk)}$, $u_{00(jk)}$, u_{10j0} en u_{200k} . Deze residuen kunnen als volgt geïnterpreteerd worden: de eerste term ($e_{i(jk)}$) geeft de afwijking van leerling i van het schoolgemiddelde in school j in jaar k ; $u_{00(jk)}$ is de afwijking van de jaargroep van het school- en jaargemiddelde; u_{10j0} representeert de afwijking van school j van het algemeen gemiddelde, en u_{200k} geeft de afwijking van jaar k van het algemeen gemiddelde.

We nemen aan dat de residuele scores normaal verdeeld zijn met een gemiddelde van 0 en een variantie van respectievelijk: $S^2_{e_{i(jk)}}$, $S^2_{u_{10jk}}$, $S^2_{u_{10j0}}$ en $S^2_{u_{200k}}$.

Model II

Het model betreft het SO- of CE-gemiddelde van een school j op enig vak in examenjaar k . Stel, y_{jk} is het gemiddelde van school j ($j = 1, 2, \dots, J$) in jaar k ($k = 1, 2, \dots, K$). Het model kan beschreven worden als: $y_{jk} = b_0 + [e_{jk} + u_{1j0} + u_{20k}]$.

De fixed parameter (b_0) is een schatting van het algemeen gemiddelde. In het random deel van het model (tussen vierkante haken) worden drie residuele termen onderscheiden: e_{jk} , u_{1j0} , u_{20k} . Deze residuen kunnen als volgt geïnterpreteerd worden: de eerste term (e_{jk}) geeft de afwijking van het schoolgemiddelde in jaar k van het jaar- en schoolgemiddelde; u_{1j0} representeert de afwijking van school j van het algemeen gemiddelde en u_{20k} geeft de afwijking van jaar k van het algemeen gemiddelde.

We nemen aan dat de residuele scores normaal verdeeld zijn met een gemiddelde van 0 en een variantie van respectievelijk: $S^2_{e_{jk}}$, $S^2_{u_{1j0}}$ en $S^2_{u_{20k}}$.

Appendix II

Parameterschattingen onder Model I

Toelichting:

Gem. gemiddelde

Var. totale variantie

S^2_{II} percentage variantie verklaard door verschillen tussen leerlingen

S^2_{Jg} percentage variantie verklaard door verschillen tussen jaargroepen

S^2_{sch} percentage variantie verklaard door verschillen tussen scholen

S^2_{jaar} percentage variantie verklaard door verschillen tussen jaren

Jaar/Sch. ratio van de variantie tussen jaren en variantie tussen scholen

Schattingen voor het so (havo)

Vak	Gem.	Var.	Percentage verklaarde variantie				Jaar/Sch.
			S^2_{II}	S^2_{Jg}	S^2_{sch}	S^2_{jaar}	
Nederlands	65	61	90.1	3.4	6.5	.1	.01
Frans	64	90	90.7	3.2	5.9	.1	.02
Duits	64	80	89.0	3.5	7.4	.2	.02
Engels	65	82	92.4	1.9	5.6	.1	.02
Geschiedenis	65	71	89.7	4.3	6.0	.0	.00
Aardrijkskunde	63	63	89.8	4.2	5.7	.3	.04
Wiskunde A	64	112	92.1	4.0	4.0	.0	.00
Wiskunde B	60	122	92.3	4.3	3.5	.0	.00
Natuurkunde	62	89	91.3	3.7	5.0	.0	.01
Scheikunde	63	99	93.3	3.2	3.5	.1	.02
Biologie	63	61	89.9	3.6	6.3	1	.02
Economie	60	91	91.8	3.8	4.3	.1	.02
Handel	60	122	93.3	2.7	3.9	.1	.01

Schattingen voor het CE (havo)

Vak	Gem.	Var.	Percentage verklaarde variantie				Jaar/Sch.
			S^2_{II}	S^2_{Jg}	S^2_{sch}	S^2_{jaar}	
Nederlands	62	60	91.9	2.7	5.1	.3	.06
Frans	65	170	86.9	2.6	7.8	2.7	.34
Duits	61	163	87.7	1.5	9.6	1.2	.13
Engels	66	190	95.9	.7	3.1	.3	.08
Geschiedenis	65	113	85.0	7.5	5.5	2.0	.37
Aardrijkskunde	61	101	82.6	5.6	10.4	1.4	.13
Wiskunde A	66	183	83.4	5.9	5.1	5.7	1.11
Wiskunde B	61	202	88.2	3.2	7.9	.8	.10
Natuurkunde	64	161	84.9	4.3	8.0	2.9	.36
Scheikunde	63	157	86.7	3.2	7.7	2.5	.33
Biologie	62	96	89.9	2.6	6.0	1.5	.24
Economie	60	144	87.8	4.0	4.7	3.5	.74
Handel	58	154	86.6	4.2	8.5	.6	.07

Appendix III

Normhandhavingsonderzoek

Twee examens zijn equivalent of gelijkwaardig als gelijke prestaties gelijk gewaardeerd worden. Dit betekent dat een leerling die twee verschillende examens maakt in principe hetzelfde cijfer krijgt. In feite zal dat niet altijd het geval zijn, omdat ook het toeval een rol speelt.

Om na te gaan of twee examens gelijkwaardig zijn, is normhandhavingsonderzoek nodig. In dit onderzoek wordt vastgesteld wat de prestaties zijn van een groep leerlingen op een referentie-examen met een vastgelegde norm. Daaruit kan een normering voor het te equivaleren examen worden afgeleid waardoor een leerling die op het referentie-examen voldoende haalt, dit ook haalt op het te equivaleren examen. Deze norm wordt als adviesnorm aan de CEVO beschikbaar gesteld bij het vaststellen van de norm. Voor een gedetailleerde bespreking zie Alberts (2001) en Glas en Beguin (1999). In 1994 is het voor het eerst gebruikt bij de moderne vreemde talen in vbo- en mavo-D, en havo. In 1997 is het ingevoerd voor de moderne vreemde talen havo en voor een aantal bèta- en gammavakken in havo en vwo. In de volgende tabel is aangegeven voor welke vakken normhandhavingsonderzoek wordt uitgevoerd.

Vakken betrokken bij het normhandhavingsonderzoek met jaar van ingang

Vakken betrokken bij het normhandhavingsonderzoek met jaar van ingang

Vak	Jaar						
	1994	1995	1996	1997	1998	1999	2000
Nederlands D	•	•	•	•	•	•	•
Frans	•	•	•	•	•	•	•
Duits	•	•	•	•	•	•	•
Engels	•	•	•	•	•	•	•
Wiskunde B				•	•	•	•
Scheikunde				•	•	•	•
Natuurkunde				•	•	•	•
Handel				•	•	•	•

Voor de moderne vreemde talen en Nederlands D wordt het normhandhavingsonderzoek uitgevoerd op basis van data die direct na afname van het examen bij een andere doelgroep dan de examenkandidaten wordt verzameld. Voorbeeld: het Engels havo-examen wordt gecombineerd met het referentie-examen Engels afgenomen bij leerlingen uit 5 vwo.

Voor Nederlands-havo, de exacte vakken en de mens- en maatschappijvakken is het Cito aangewezen op de feitelijke doelgroep. De benodigde gegevens worden in het kader van het laatste SO of als examen-voorbereidingstoets afgenomen bij leerlingen die datzelfde schooljaar nog voor hun CE opgaan.