

Kwaliteit van de Data bij Normhandhaving

A. Béguin

M. de Groot

Kwaliteit van de Data bij Normhandhaving

A. Béguin
M. de Groot

Citogroep
Arnhem, oktober 2002

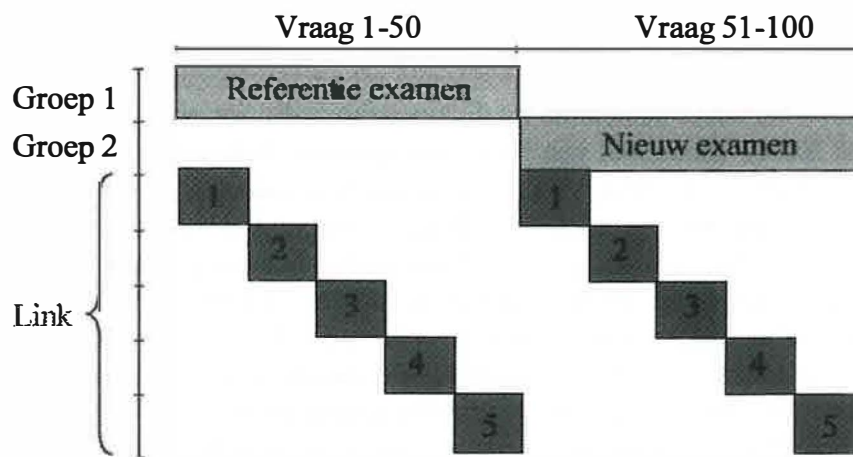
© Arnhem 2002

R&D Notities zijn voor intern gebruik bedoelde notities van medewerkers van de Citogroep. Aanhalingen uit deze notities of verwijzingen naar deze notities vereisen de toestemming van de auteur(s).

Recent is enige twijfel ontstaan over de kwaliteit van de (posttest)data die verzameld wordt bij de normhandhavingprocedure van moderne vreemde talen (MVT). Deze twijfel heeft zowel theoretische als empirische gronden. De theoretische twijfel aan de kwaliteit van de posttest is gebaseerd op de recente veranderingen in de examenpraktijk. Zoals daar zijn: het *gebruik van woordenboeken* bij het examen, de *introdunctie van open vragen*, *introdunctie van het studiehuis* en daarmee samenhangend het *onderscheid tussen deelvak en totaalvak* en de *introdunctie van eindterm 1*. Empirisch is het de afgelopen examencampagnes opgevallen dat de prestaties op de aanvullende data in het posttestdesign mogelijk aan verandering onderhevig zijn en dat de verzamelde data daardoor mogelijk niet meer voldoen om te worden gebruikt in de normhandhavingprocedure. In dit onderzoek zal geprobeerd worden te achterhalen of de kwaliteit van de data voldoende is om de normhandhaving correct uit te voeren. Alvorens de vraag naar de kwaliteit van de data te beantwoorden zal kort uiteengezet worden wat voor data gebruikt worden bij de normhandhaving MVT.

Het gebruikte design

Het design waarmee de data bij de normhandhavingprocedure MVT verzameld worden wordt ook wel een posttestdesign genoemd. Naast de antwoorden op de opgaven gegeven door de examenkandidaten van het referentie-examen en een nieuw examen bestaat het posttestdesign uit antwoorden op aanvullende toetsen die zijn afgenomen om na te kunnen gaan hoe het referentie-examen zich verhoudt tot het nieuwe examen. In de praktijk worden de aanvullende toetsen afgenomen bij groepen leerlingen die geacht mogen worden de opgaven op een redelijk niveau te kunnen maken, maar die niet deelnemen aan het nieuwe of de afgenomen oude examens. Bijvoorbeeld gegevens over de VMBO examens kunnen verzameld worden bij leerlingen in de 3^e klas van het VWO.



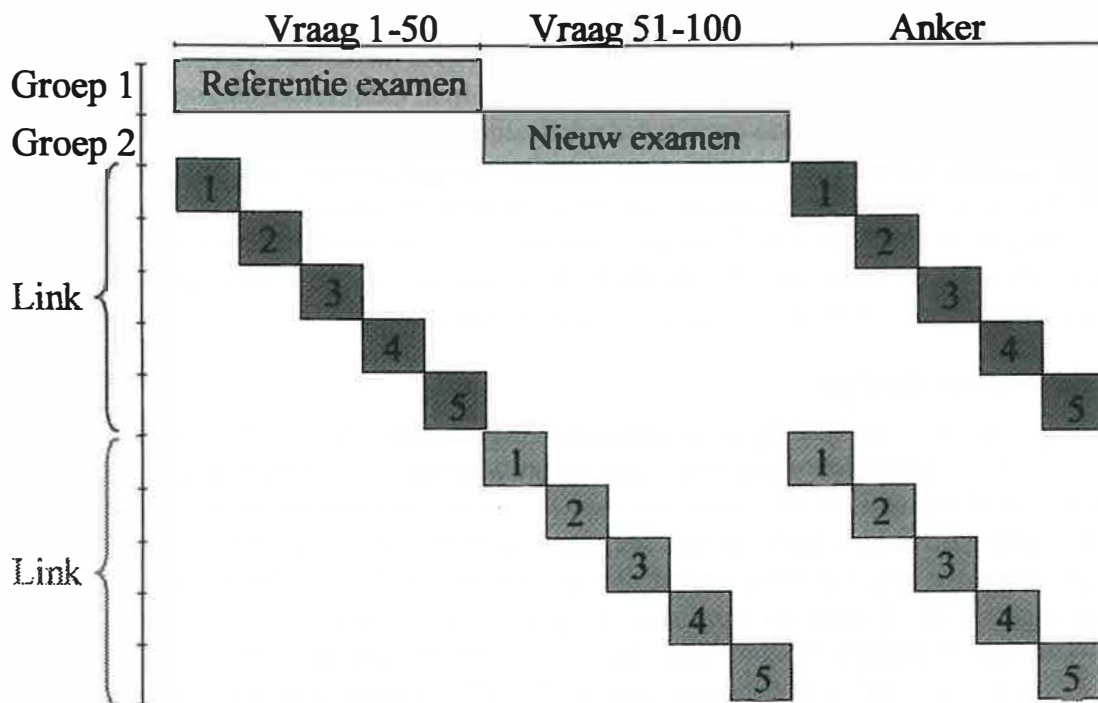
Figuur 1 Schematische weergave van het posttestdesign zoals dat tot en met 1998 werd gebruikt. Het vlak achter de aanduiding 'Groep 1' geeft een steekproef uit de antwoorden op het referentie examen weer. Achter 'Groep 2' staan de antwoorden op het nieuwe examen. De groepjes met de labels 1 tot 5 geven de aanvullende toetsen aan waarbij opgemerkt kan worden dat elk van deze toetsen bestaat uit vragen uit beide examens.

Tot en met 1998 bestonden de aanvullende toetsen uit zowel opgaven uit het nieuwe examens als opgaven uit het referentie-examen. (zie voor een uitgebreidere beschrijving: Béguin, Alberts & Kremers, 2002). Een schematische weergave van dit design wordt gegeven in Figuur 1.

Posttestdesign met een ankertoets

Na 1998 worden aanvullende data verzameld met toetsen die opgaven uit het nieuwe examen en opgaven uit een zogeheten ankertoets bevatten. De opgaven uit deze ankertoets komen uit derdetijdvak-examens HAVO en VWO. Voor deze verandering is destijds gekozen omdat

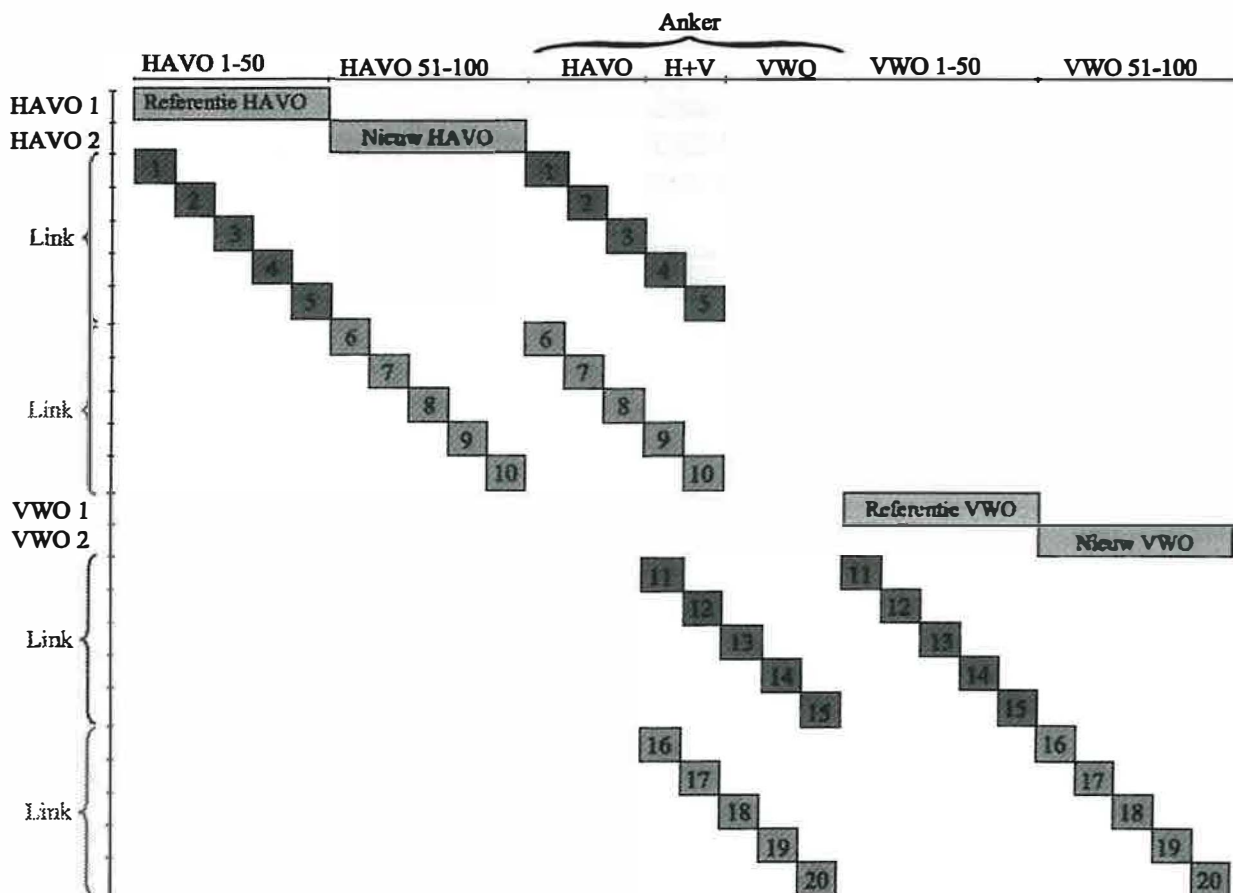
leerlingen bekend konden zijn met de opgaven van het referentie-examen. Het referentie-examen is namelijk openbaar beschikbaar en bijvoorbeeld opgenomen in examenbundels en op cd-rom's met oefenopgaven. De nieuwe procedure maakt gebruik van een geheime ankertoets en dit garandeert dat de opgaven niet in oefenmateriaal zijn opgenomen.



Figuur 2. Schematische weergave van het posttestdesign met een ankertoets.

In Figuur 2 zijn de scores op twee examens weergegeven. Behalve de scores van de twee oorspronkelijke groepen leerlingen, één bestaande uit de kandidaten die het referentie-examen gemaakt hebben en de ander bestaande uit de kandidaten die het nieuwe examen gemaakt hebben, zijn ook scores verzameld van 10 afnamegroepen. Iedere afnamegroep heeft een stukje van een examen en een stukje van de ankertoets gemaakt. Deze afnamegroepen dienen om de resultaten van de twee examens met elkaar in verband te brengen. De groepen genummerd 1 tot 5 hebben een stukje van het referentie-examen gemaakt in combinatie met een stuk van de ankertoets. De groepen genummerd 6 tot 10 hebben een stukje van het nieuwe examen gemaakt samen met een stukje van de ankertoets. Merk op dat de opgaven uit de ankertoets zowel in combinatie met opgaven uit het nieuwe examen als met opgaven uit het referentie-examen gemaakt worden. Bijvoorbeeld afnamegroep 1 en 6 hebben dezelfde opgaven uit de ankertoets gemaakt maar groep 1 deed dat in combinatie met een deel van het referentie-examen en groep 6 deed dat in combinatie met een deel van het nieuwe examen. Indirect via de ankertoets wordt op deze manier dus de link gelegd tussen beide examens. Net als bij het gewone posttestdesign geldt dat het vaardigheidsniveau van de afnamegroepen niet hetzelfde hoeft te zijn als die van de examengroepen. Hun scores kunnen in het model weergegeven door een aparte vaardigheidsverdeling. Wel moeten hun scores in het model passen. Dat is niet het geval als ze op hun toetsjes bijvoorbeeld puur zouden gokken. Het is dus belangrijk dat de leerlingen van de afnamegroepen op hun toetsjes serieus, en gebruikmakend van hun vaardigheid, antwoord geven.

Bij zowel HAVO en VWO examens MVT wordt gebruikt gemaakt van een posttestdesign met een ankertoets. De posttestdesigns van HAVO en VWO vertonen enige overlap doordat een aantal opgaven van de ankertoets van HAVO en VWO gelijk zijn. Het totale design voor de HAVO en VWO examens MVT wordt gegeven in Figuur 3.



Figuur 3 Schematische weergave van het posttestdesign met een ankertoets zoals dat gebruikt wordt bij HAVO en VWO examens. Een gedeelte van de opgaven uit de ankertoets zijn opgaven specifiek op VWO niveau (de gedeeltes van de ankertoets die voorkomen in de toetsen met nummers 13, 14, 15, 18, 19 en 20), een ander gedeelte zijn opgaven specifiek op HAVO niveau (de gedeeltes van de ankertoets die voorkomen in de toetsen met nummers 1, 2, 3, 6, 7 en 8) en een gedeelte van de opgaven kan beschouwd worden als opgaven die geschikt zijn voor zowel HAVO als VWO (de gedeeltes van de ankertoets die voorkomen in de toetsen met nummers 4, 5, 9, 10, 11, 12, 16 en 17). Deze opgaven staan in het gedeelte met de licht grijze achtergrond.

Een gedeelte van de opgaven uit de ankertoets zijn opgaven specifiek op VWO niveau. Deze opgaven zijn opgenomen in de toetsen met nummers 13, 14, 15, 18, 19 en 20 in Figuur 3. Een ander gedeelte van de ankertoets bestaat uit opgaven specifiek op HAVO niveau. Deze opgaven zijn opgenomen in de toetsen die in Figuur 3 worden weergegeven met nummers 1, 2, 3, 6, 7 en 8. Het derde gedeelte van de opgaven kan beschouwd worden als opgaven die geschikt zijn voor zowel HAVO als VWO. Deze opgaven zijn opgenomen in de aanvullende toetsen die in Figuur 3 worden weergegeven met nummers 4, 5, 9, 10, 11, 12, 16 en 17.

De praktijk van het posttestdesign met een ankertoets

In Appendix A staan voor elk van de HAVO en VWO examenvakken MVT over de periode 1998- 2002 de gegevens over het posttestdesign met een ankertoets. In de tabellen in Appendix A zijn per examenvak en per jaar de opgaven uit de ankertoets en het nieuwe examen in elk van de aanvullende toetsen gespecificeerd. Ook geeft deze tabel de steekproefgrootte en het aantal opgaven dat uit de ankertoets en uit het nieuwe examen komt. In Tabel 1 wordt voor de opgaven uit het nieuwe examen aangegeven hoeveel leerlingen deze opgave in de aanvullende toetsen hebben beantwoord. Opvallend hierbij is dat in een aantal gevallen een groot deel van de examenopgaven niet in de aanvullende toetsen was opgenomen ondanks dat de open vragen nog niet waren ingevoerd. Bij de aanvullende toetsen waarbij opgaven uit het referentie-examen en de ankertoets werden afgenomen ontbreekt bij alle vakken een gedeelte van de opgaven uit het referentie-examen. Bij Engels Havo zijn bijvoorbeeld 22 opgaven van het referentie-examen niet afgenomen in de aanvullende toetsen (dit is 44% van het totaal aantal opgaven). Zolang de ontbrekende opgaven dezelfde

vaardigheid meten als het totale examen hoeft dit niet problematisch te zijn, maar als het ontbrekende deel van de opgave afwijkende vaardigheden meet dan kan dit effect hebben op de resultaten van de normhandhavingprocedure. Verder valt op, dat de steekproeven vaak te klein zijn. In het verleden is opgemerkt dat 200 waarnemingen de ondergrens is waarbij de normhandhaving met het posttestdesign acceptabele betrouwbaarheidsintervallen oplevert.

Tabel 1 het aantal leerlingen per opgave in de aanvullende toetsen

Aantal LL:	0	61-80	81-100	101-120	121-140	141-160	161-180	181-200	>200
engels HAVO									
referentie	22	0	0	0	0	0	0	0	28
1999	0	0	0	0	0	50	0	0	0
2000	8	0	0	0	0	0	37	0	0
2001	17	0	0	10	0	0	0	10	9
2002	25	0	0	25	0	0	0	0	0
engels VWO									
referentie	9	0	0	0	0	0	0	0	41
1999	0	0	0	0	0	0	50	0	0
2000	0	0	0	0	0	0	38	12	0
2001	15	0	34	0	0	0	0	0	0
2002	21	0	0	0	0	0	0	26	0
duits HAVO									
referentie	9	0	0	29	12	0	0	0	0
1999	0	10	40	0	0	0	0	0	0
2000	17	0	30	0	0	0	2	0	0
2001	15	0	0	0	31	0	0	0	0
2002	18	0	8	21	0	0	0	0	3
duits VWO									
referentie	12	0	0	0	38	0	0	0	0
1999	0	0	0	42	8	0	0	0	0
2000	0	0	24	26	0	0	0	0	0
2001	10	27	8	0	0	0	0	0	0
2002	22	0	0	0	0	0	22	0	0
frans HAVO									
referentie	10	0	40	0	0	0	0	0	0
1999	11	29	0	0	10	0	0	0	0
2000	12	0	29	0	0	0	0	4	0
2001	14	14	18	0	0	0	0	0	0
2002	18	14	13	0	0	0	0	0	0
frans VWO									
referentie	9	0	41	0	0	0	0	0	0
1999	0	0	29	21	0	0	0	0	0
2000	0	0	50	0	0	0	0	0	0
2001	14	0	30	0	0	0	0	0	0
2002	16	0	28	0	0	0	0	0	0

Invloed van veranderingen: de theorie

Een aantal recente veranderingen in de examenpraktijk (de introductie van woordenboeken, eindterm 1, open vragen, deelvak en totaalvak en open vragen) kan mogelijk van invloed zijn op de kwaliteit van de posttest. Hieronder wordt de mogelijke invloed van elk van deze veranderingen besproken.

Woordenboekgebruik

Vanaf de 2000 populatie HAVO en de 2001 populatie VWO is het bij de examens moderne vreemde talen toegestaan om een woordenboek te gebruiken. De aanvullende gegevens die nodig zijn om de normhandhavingprocedure uit te kunnen voeren worden echter verzameld in een conditie waarbij het niet toegestaan is om een woordenboek te gebruiken. Hier onder zal

op theoretische gronden kort ingegaan worden op het effect dat woordenboekgebruik mogelijk op de resultaten van de normhandhavingprocedure kan hebben.

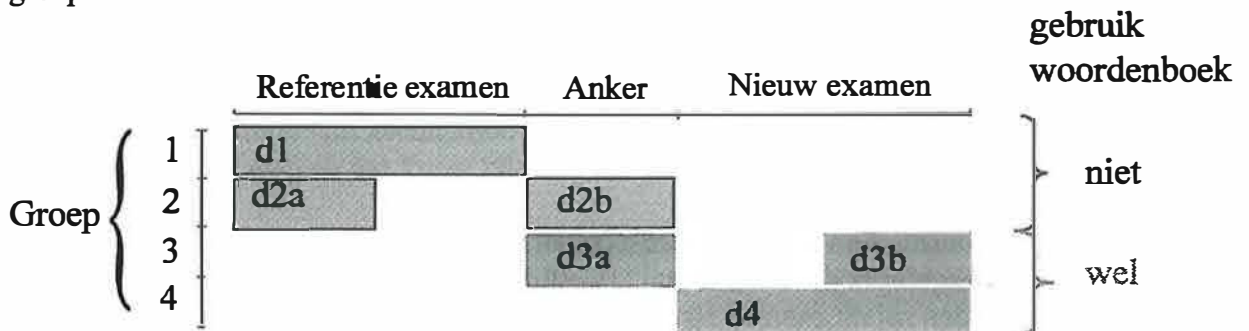
1. Als ze bij de nieuwe vragen geen woordenboek mogen gebruiken zijn de leerlingen op deze vragen in het nadeel t.o.v de examenconditie. Ook bij de oude vragen hebben zij mogelijk problemen met het idioom (als door woordenboekgebruik minder nadruk bestaat op het verwerven van idioomkennis). Wanneer beide effecten even groot zijn is het effect op de moeilijkheidsschatting mogelijk te verwaarlozen (misschien moet dit onderzocht worden).
Waarschijnlijker is echter om te veronderstellen dat woordenboekgebruik effect zal hebben op sommige opgaven en niet op andere opgaven. In een design waarin dezelfde opgaven zowel met woordenboek als zonder woordenboek afgenomen worden zal het Item Response Theorie (IRT) model, dat gebruikt wordt om de data te beschrijven, minder goed bij de data passen.
2. Als bij de constructie van het nieuwe examen rekening gehouden wordt met woordenboekgebruik door een aantal 'moeilijke' woorden op te nemen die opgezocht moeten worden zal dit examen als relatief moeilijk worden geschat wanneer geen woordenboek mag worden gebruikt. Dit leidt ertoe dat de vaardigheid van de leerlingen die het nieuwe examen maken wordt overschat. Immers het nieuwe examen wordt als relatief moeilijker ingeschat dan het oude examen. In de examenconditie met woordenboek is het nieuwe examen echter niet moeilijker dan het oude examen. Het gevolg hiervan is dat de prestatie op het nieuwe examen als beter wordt beoordeeld dan die eigenlijk is.
3. Mogelijk leidt een vergrote moeilijkheid zoals gerapporteerd bij 1) tot het gokken van het juiste antwoord. Als dit effect zich op grote schaal voordoet kan dat tot problemen leiden bij de normhandhavingprocedure. In zo'n geval bevatten de data verzameld met de aanvullende toetsen te weinig informatie om de examendata te koppelen.

Een alternatieve situatie is denkbaar waarbij in de aanvullende toetsen bij zowel de opgaven uit de ankertest als de opgaven uit het nieuwe examens een woordenboek mag worden gebruikt. In dit geval ontstaat een verschil in afnameconditie tussen de opgaven uit het ankerexamen in de posttest bij de huidige afname waarbij een woordenboek gebruikt wordt en eerdere afnamen waarbij geen woordenboek mocht worden gebruikt.

1. Onduidelijk is of dit tot betere resultaten zal leiden omdat het ook mogelijk is dat tijdnood ontstaat door het woordenboekgebruik. In zo'n geval is er enerzijds een verbeterde prestatie op de vragen uit de ankertoets door woordenboekgebruik en anderzijds een verslechterde prestatie op de opgaven aan het einde van de posttest door tijdnood.
2. Als we een eventueel optreden van tijdnood verwaarlozen wordt de vaardigheid van de nieuwe populatie overschat. Dit kan worden geïllustreerd aan de hand van een voorbeeldesign met 4 groepen leerlingen die een gelijke vaardigheid hebben. De eerste groep leerlingen maakt een oud examen. De tweede groep leerlingen maakt zonder woordenboek opgaven uit een oud examen en opgaven uit een ankertoets. De derde groep leerlingen maakt met woordenboek dezelfde opgaven uit de ankertoets en opgaven uit een nieuw examen. De vierde groep maakt het nieuwe examen. Het complete design is weergegeven in Figuur 4.

Zoals we hierboven hebben verondersteld zijn alle groepen even vaardig. De prestaties van groepen 1 en 2 op de gemeenschappelijk gemaakte opgaven zijn daarom vergelijkbaar met elkaar en ook de prestaties van groepen 3 en 4 op de gemeenschappelijk gemaakte opgaven zullen vergelijkbaar zijn. Het gevolg hiervan is dat onder het model de vaardigheden van groep 1 en 2 gelijk zullen zijn en dat de vaardigheden van groep 3 en 4 gelijk zullen zijn. Om het voorbeeld eenvoudig te houden veronderstellen we dat de ankeropgaven en de examenopgaven van beide examens even moeilijk zijn. Dus de gemiddelde moeilijkheid van de opgaven uit het referentie examen die gemaakt worden door groep 2 (d2a) zijn even moeilijk als de opgaven uit het anker (d2b) en de opgaven uit het nieuwe examen die gemaakt worden door groep 3 (d3a) zijn even moeilijk als het

anker (d3b). Omdat groep 3 echter een woordenboek mag gebruiken scoren zij beter op de ankeropgaven dan groep 2. Logischerwijs moet de moeilijkheid van het anker in groep 2 en 3 gelijk ($d2b=d3b$). Om toch het betere resultaat van groep 3 te kunnen verklaren wordt groep 3 als vaardiger dan groep 2 ingeschat. Eerder zagen we dat de vaardigheid van groep 3 gelijk is aan die van groep 4 en dat de vaardigheid van groep 2 gelijk is aan de vaardigheid van groep 1. Het gevolg is dat groep 4 als vaardiger wordt ingeschat dan groep 1.

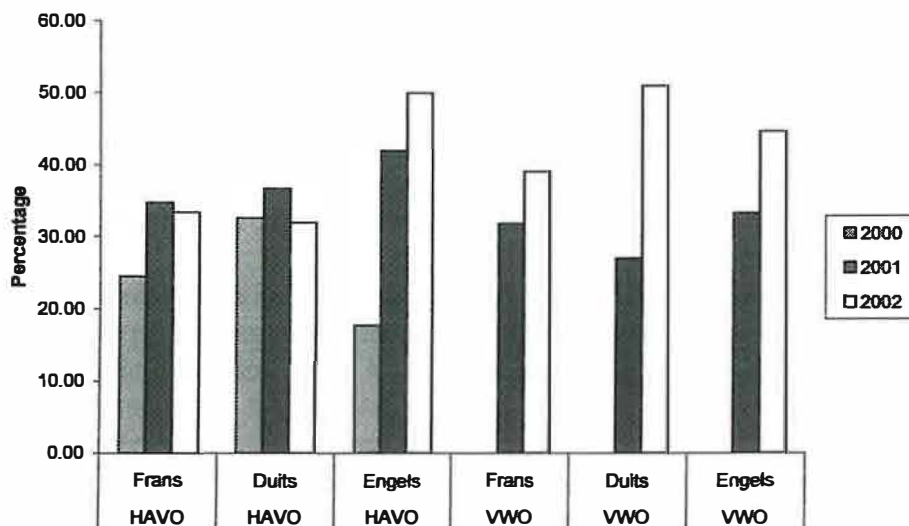


Figuur 4 Schematische weergave van het voorbeeld

Een derde mogelijkheid is dat wel een woordenboek mag worden gebruikt bij de opgaven uit het nieuwe examen en niet bij de opgaven uit de ankertoets. In dit geval zullen de opgaven uit de ankertoets (door gebrek aan idioomkennis) slechter gemaakt worden. Dit leidt tot de inschatting (op basis van de post-test) dat het nieuwe examen veel makkelijker is. De echte prestatie op dit examen valt dan relatief tegen en de vaardigheid van de nieuwe populatie wordt onderschat.

De introductie van open vragen

Sinds 2000 zijn een aantal van de meerkeuzevragen van de HAVO examens vervangen door open vragen en vanaf 2001 gebeurt hetzelfde bij examens VWO. Om te bepalen of de invoering van open vragen invloed heeft op de kwaliteit van de normhandhavingprocedure zijn twee aspecten van belang. Ten eerste welk gedeelte van de te behalen scorepunten worden gebaseerd op de open vragen. Ten tweede in welke mate wijkt de vaardigheid die gemeten wordt met de open vragen af van de vaardigheid die gemeten wordt met de andere opgaven. Wat betreft het eerste aspect kan opgemerkt worden dat het streven is om een derde van de te behalen scorepunten te baseren op open vragen. In Figuur 5 wordt voor de MVT



Figuur 5 Percentage van de te behalen scorepunten dat wordt gebaseerd op open vragen

examens HAVO en VWO het gerealiseerde percentage van de mogelijk te behalen scorepunten gegeven dat is gebaseerd op open vragen.

In welke mate de open vragen een andere vaardigheid meten dan de gesloten vragen is aan de hand van de examens 2001 d.m.v. factor-analyse onderzocht. Bij het merendeel van de examens 2001 werd gevonden dat de open vragen geen noemenswaardig andere vaardigheid meten dan de gesloten vragen. Bij enkele examens lijken er kleine verschillen tussen de gemeten vaardigheden (met open dan wel gesloten vragen) te bestaan. Het is echter niet gegarandeerd dat in nieuwe examens de open vragen wederom dezelfde vaardigheid zullen meten. Daarom kunnen deze bevindingen niet zomaar gegeneraliseerd worden naar nieuwe examens.

De mate van invloed van bovenstaande aspecten op de resultaten van normhandhaving kan nagegaan worden op basis van simulatiestudies. Hierbij wordt voor verschillende afwijkingen in de gemeten vaardigheden nagegaan hoe groot de invloed zou zijn op de resultaten van normhandhaving. Op basis van een simulatiestudie (Béguin, 2002) lijkt de normhandhavingprocedure redelijk robuust voor afwijkingen in de gemeten vaardigheid in een relatief klein deel van het examen. De resultaten van de simulatiestudie zijn niet generaliseerbaar naar alle MVT examens omdat het maximale percentage van de te behalen scorepunten op de afwijkende vaardigheidsdimensie in de simulatiestudie 30 procent betrof. In werkelijkheid is dit percentage voor een aantal examens beduidend groter. Zo wordt bijvoorbeeld 50% van de scorepunten Engels HAVO en Duits VWO bepaald op basis van open vragen. Logischerwijs resulteert een groter percentage opgaven in een ander vaardigheidsdomein tot grotere invloed op de resultaten van de normhandhavingprocedure.

Introductie van het studiehuis en daarmee samenhangend het onderscheid tussen deelvak en totaalvak

Sinds de invoering van het studiehuis wordt voor leerlingen een onderscheid gemaakt tussen examenvakken die deel uitmaken van hun profiel (totaalvak) en vakken waarbij dit niet het geval is (deelvak). Dit leidt tot een verandering in vaardigheid bij de populatie die deelnemen aan de aanvullende toetsen. Voor 2000 bestond deze populatie uit 5 VWO leerlingen die de taal in hun vakkenpakket gekozen hadden. Vanaf 2000 doen alle leerlingen eindexamen in tenminste twee MVT vakken. Hiervan is tenminste één van twee een totaalvak. De andere taal kan op het deelvakniveau worden afgesloten. Leerlingen die vroeger de taal niet in hun vakkenpakket hadden gekozen deden niet mee aan de aanvullende toetsen. Nu echter doen zij examen op deelvakniveau en doen zij wel mee aan de aanvullende toetsen. Het is realistisch om te veronderstellen dat deze leerlingen minder vaardig zijn dan de heelvak populatie en dus is het logisch om te veronderstellen dat de deelnemers aan de aanvullende toetsen minder vaardig geworden zijn (en dat de vaardigheidsverdeling een grotere spreiding heeft). Verder is het onduidelijk in welke mate docenten de leerlingen selecteren voor de aanvullende toetsen. Het is namelijk mogelijk dat zij besluiten de aanvullende HAVO toetsen aan bijvoorbeeld de deelvakleerlingen te geven terwijl zij de VWO-toetsen aan de totaalvak leerlingen voorleggen. We hebben hierover echter geen gegevens omdat dit niet wordt gecontroleerd of wordt voorgeschreven aan de docent.

De introductie van eindterm 1

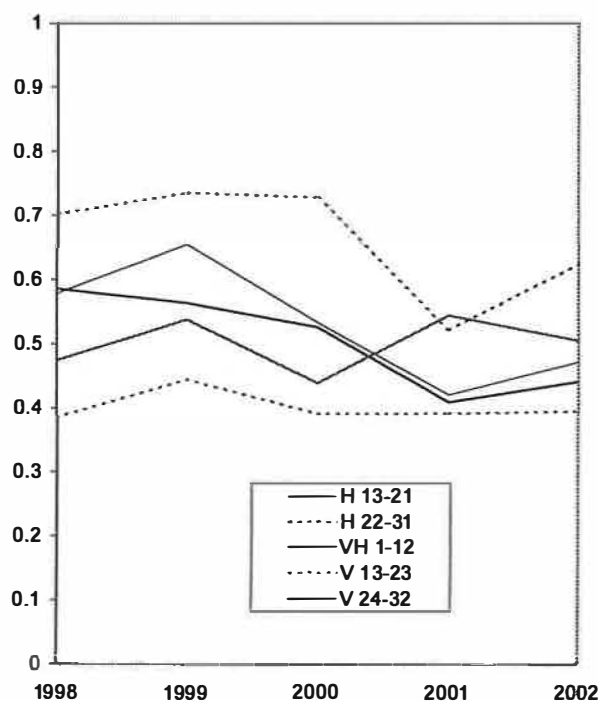
De doelstelling van eindterm 1 is dat leerlingen uit een relatief grote hoeveelheid tekst de relevante informatie kunnen vinden. Sinds 2001 wordt tien procent van de te behalen scorepunten op een examen besteed aan het meten van eindterm 1. Om te bepalen of de invoering van eindterm 1 invloed heeft op de kwaliteit van de normhandhavingprocedure zijn twee aspecten van belang. Ten eerste welk gedeelte van de te behalen scorepunten worden gebaseerd op de nieuwe eindterm (antwoord = 10% van de te behalen scorepunten). Ten tweede in welke mate wijkt de vaardigheid die gemeten wordt met de eindterm 1 opgaven af

van de vaardigheid die gemeten wordt met de andere opgaven. Opgemerkt dient te worden dat eindterm 1 een deelvaardigheid beoogt te meten die vroeger niet expliciet in het examen voorkwam, maar dat deze deelvaardigheid wel een onderdeel is van leesvaardigheid, de totale vaardigheid die de MVT examens beogen te meten. Waarschijnlijk zullen de resultaten op de eindterm 1 opgaven behoorlijk hoog correleren met de rest van het examen. Om te controleren of dit klopt kan bij de examens die al afgenomen zijn onderzocht worden hoe hoog de geschatte correlatie is tussen de vaardigheid die gemeten wordt met eindterm 1 en de correlatie die gemeten wordt met de rest van het examen. Dit geeft een beeld over de mate waarin de opgaven die eindterm 1 meten iets anders meten dan de rest van het examen. De mate van invloed van bovenstaande aspecten op de resultaten van normhandhaving kan nagegaan worden op basis van simulatiestudies. Hierbij wordt voor verschillende afwijkingen in de gemeten vaardigheden nagegaan hoe groot de invloed zou zijn op de resultaten van normhandhaving. Op basis van de eerder genoemde simulatiestudie (Béguin, 2002) lijkt de normhandhavingprocedure redelijk robuust voor afwijkingen in de gemeten vaardigheid in een relatief klein deel van het examen.

Invloed van veranderingen: de praktijk

Veranderingen in gemiddelde P-waarde op de ankertoetsopgaven in de posttest.

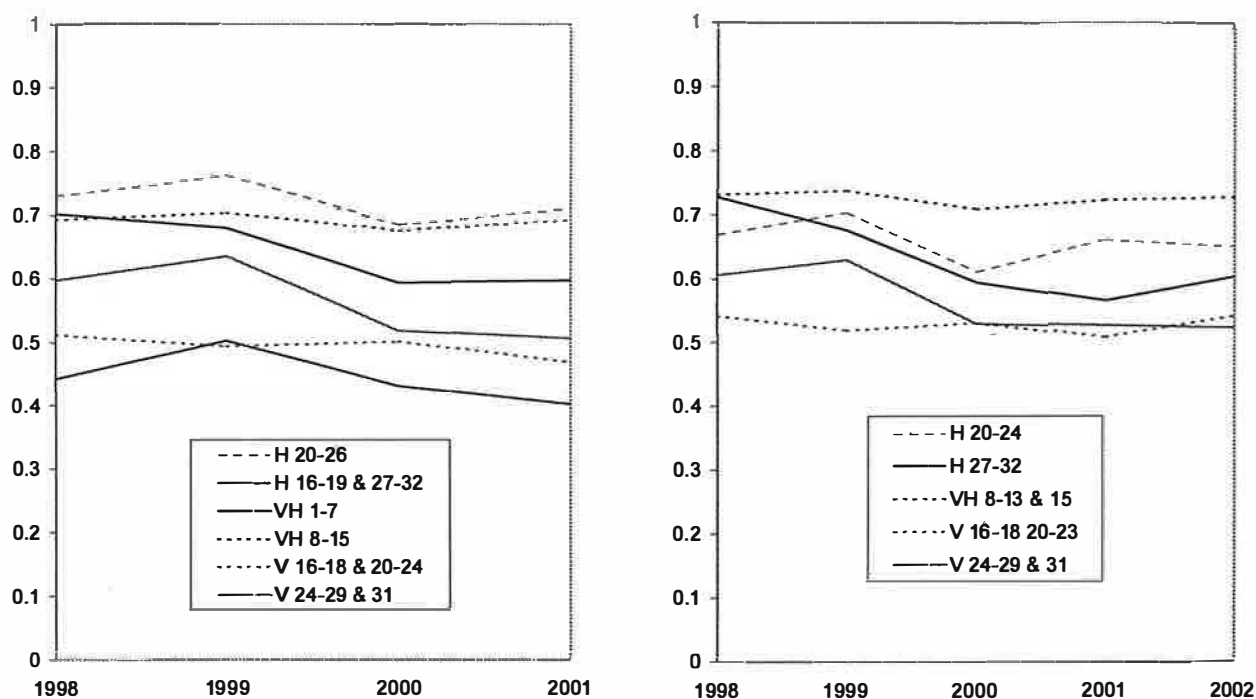
Een indicatie of de opgaven van de ankertoets bekend geworden zijn of dat de opgaven te moeilijk geworden zijn kunnen we krijgen door te kijken naar de gemiddelde P-waarde op de opgaven in de ankertoets. Als de opgaven bekend geworden zijn verwachten we dat meer leerlingen de opgaven goed beantwoorden, en dat de gemiddelde P-waarde in de loop van de tijd zal stijgen. Als de opgaven in verhouding tot het niveau van de leerlingen moeilijker worden verwachten we lagere gemiddelde P-waarden. In Figuur 6, 7 en 8 zijn de gemiddelde P-waarden van de opgaven uit de linkgroepen voor verschillende jaren in één grafiek weergegeven. In Figuur 6 is dit gedaan voor Frans HAVO en VWO. In Figuur 7 staan de gemiddelde P-waarden voor Duits HAVO en VWO. Waarbij Figuur 7A de jaren 1998 tot 2001 betreft en Figuur 7B de jaren 1998 tot 2002. In 2002 zijn de groepen opgaven die



Figuur 6 Gemiddelde P-waarde van de ankertoets opgaven bij de posttest Frans HAVO en VWO. De eerste 12 opgaven worden zowel in combinatie met HAVO als met VWO opgaven afgenomen. Deze opgaven worden aangeduid met 'VH'. De opgaven die worden weergegeven met 'H' worden alleen afgenomen in combinatie met HAVO opgaven. De opgaven die worden weergegeven met 'V' worden alleen afgenomen in combinatie met VWO opgaven.

worden aangeboden in de linkgroepen veranderd, doordat enkele opgaven niet meer worden aangeboden. In Figuur 7B worden de gemiddelde P-waarden over de in 2002 aangeboden opgaven uit de ankertoets weergegeven. In Figuur 8 staan de gemiddelde P-waarden voor Engels HAVO en VWO.

In Figuur 6 zien we dat de gemiddelde P-waarden op de VWO ankertoetsopgaven (de V opgaven) ongeveer gelijk blijven. De verschillen van jaar tot jaar in geobserveerde gemiddelde P-waarde bij de andere groepen ankertoetsopgaven zijn groot. Zo is het verschil in gemiddelde P-waarde tussen 1999 en 2001 van opgaven AH 13-21 bijvoorbeeld 0.23. Verder valt op dat de AH en AVH opgaven in 2001 en 2002 allemaal een gemiddelde P-waarde laten zien die lager is dan de P-waarde in 1998. Verder valt op dat in 2001 en 2002 de gemiddelde P-waarde van de aanvullende toetsen klein zijn. Omdat het meerkeuzevragen betreft, waarbij een leerling ook het goede antwoord zou kunnen gokken, suggereert dit dat er een aanzienlijk deel van de opgaven zeer moeilijk zijn voor de populatie leerlingen die de aanvullende toetsen maken.



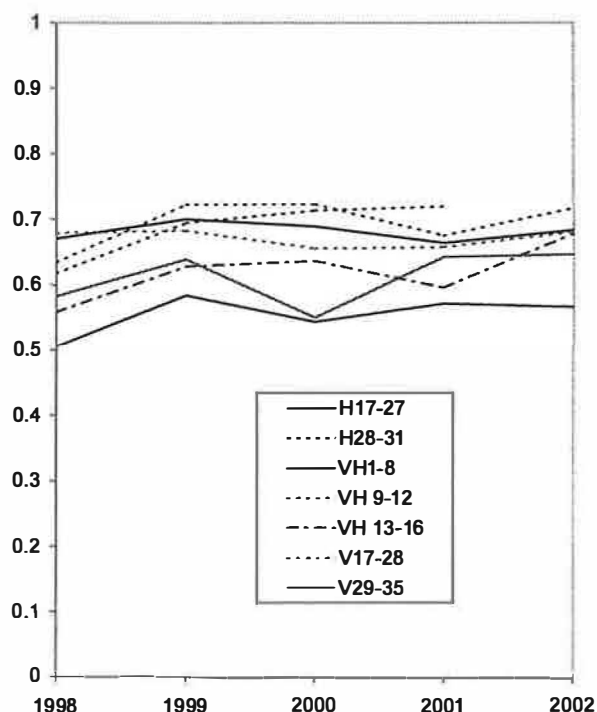
A) Gemiddelde P-waarde Duits HAVO-VWO 1998-2001 B) Gemiddelde P-waarde Duits HAVO-VWO 1998-2002

Figuur 7 Gemiddelde P-waarde van de ankertoets opgaven bij de posttest Duits HAVO en VWO. In Figuur 6A is het gemiddelde genomen over de opgaven die van 1998 tot en met 2001 in de posttest werden gebruikt. In Figuur 6B werd het gemiddelde berekend over de opgaven die werden gebruikt in de posttest 2002

In Figuur 7B zien we dat de gemiddelde P-waarden op de ankertoets Duits redelijk stabiel zijn. Hierbij dient opgemerkt te worden dat in 2002 de opgaven verwijderd zijn uit de ankertoets die minder goed leken te functioneren. Maar ook als we kijken naar de gemiddelde P-waarden op de toets zoals ze tot 2001 zijn afgenomen zien we in vergelijking tot de resultaten bij Frans kleinere verschillen in gemiddelde P-waarde over de jaren heen. Verder lijkt de gemiddelde P-waarde over het algemeen wat groter te zijn dan de P-waarden bij de ankertoets Frans. Op basis hiervan kunnen we concluderen dat de opgaven beter bij het vaardigheidsniveau van de leerlingen lijken te passen.

In Figuur 8 zien we dat de gemiddelde P-waarden op de ankertoets Engels redelijk stabiel zijn en niet lijken te dalen. Net als bij Duits zien we bij Engels in vergelijking tot de resultaten bij Frans kleinere verschillen in gemiddelde P-waarde over de jaren heen. Verder zijn de

gemiddelde P-waarde over het algemeen wat groter dan de P-waarden bij de ankertoetsen Frans en Duits.



Figuur 8 Gemiddelde P-waarde van de ankertoets opgaven bij de posttest Engels HAVO en VWO.

Te moeilijke opgaven

In de vorige paragraaf is opgemerkt dat de gemiddelde P-waarden in de aanvullende toetsen in een aantal gevallen erg laag is. Het is logisch om te veronderstellen dat dan individuele opgaven zeer slecht gemaakt zullen worden. In het geval dat de P-waarde van een meerkeuze opgave in de buurt komt van de gokkans is het meestal realistisch om te veronderstellen dat de antwoorden van een groot deel van de leerlingen in de steekproef gebaseerd zullen zijn op gokken en niet op de vaardigheid die we wensen te meten. Om een indruk te krijgen van het percentage opgaven dat (te) moeilijk is hebben we het percentage opgaven bepaald dat een geobserveerde P-waarde heeft kleiner dan de gokkans + 15%. Hierbij is de gokkans 100 gedeeld door het aantal alternatieven. De resultaten hiervan staan in de tabellen in Appendix B. Uit deze tabellen kunnen we opmaken dat bij alle VWO examens aanvullende toetsen voorkomen waarbij 50 % van de opgaven P-waarden heeft lager dan het criterium. Vooral bij Frans VWO en in mindere mate Frans Havo en Duits VWO is dit percentage opgaven soms zeer groot (tot 77%). Een van de oorzaken hiervan zou kunnen zijn dat de leerlingen in tijdnood komen bij het maken van de aanvullende opgaven (zie ook de volgende paragraaf). Wanneer zij dan geen antwoord invullen wordt dit als een fout antwoord gescoord. Om te onderzoeken of dit de hoge percentages opgaven die niet aan het criterium voldoen verklaart hebben we hetzelfde criterium toegepast op de P-waarde gecorrigeerd voor missende waarnemingen (P_{corr}). Deze P-waarde is gedefinieerd als:

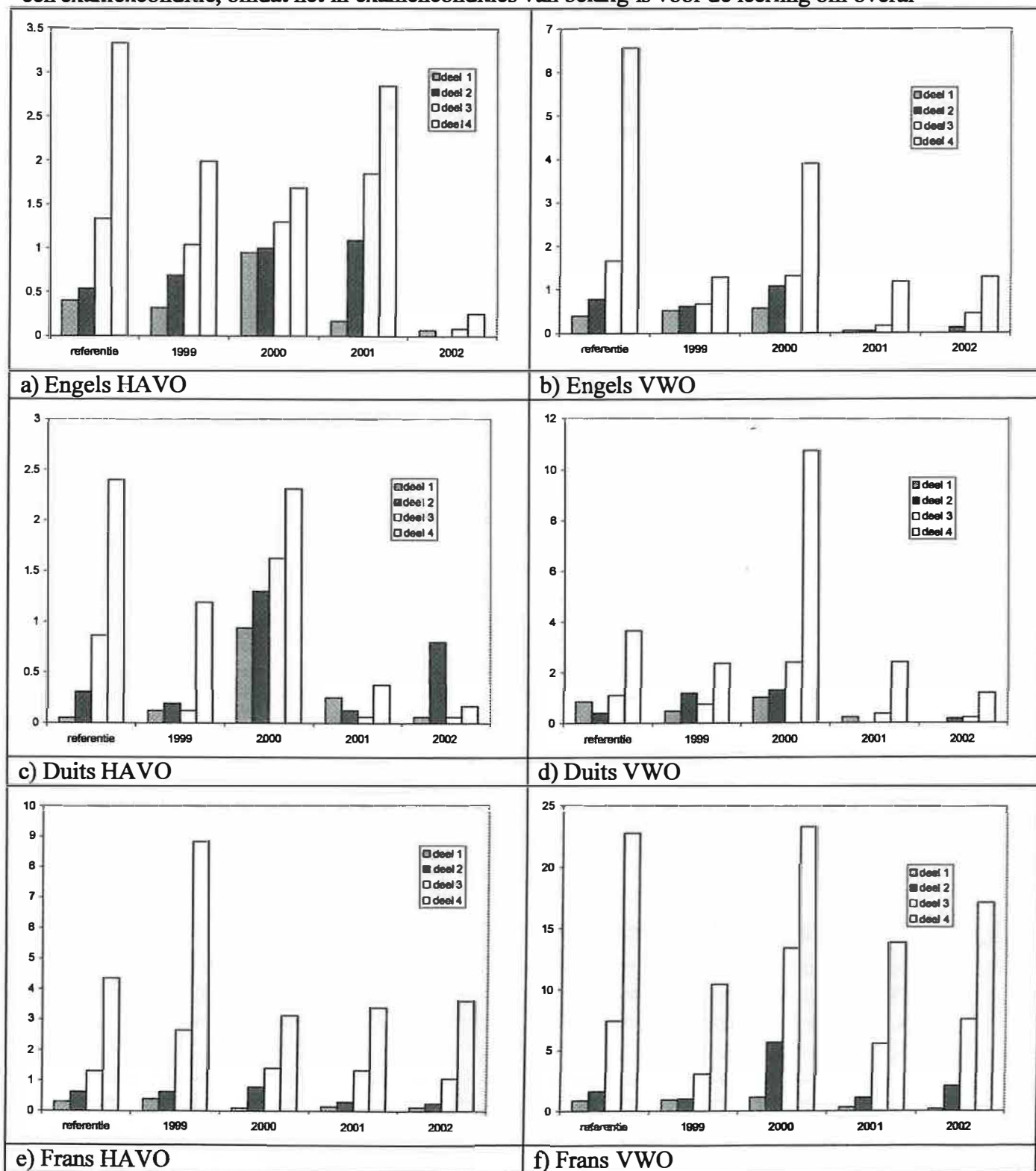
$$P_{corr} = P * 100 / (100 - m),$$

waarbij P de geobserveerde P-waarde van een opgave is en m het percentage leerlingen dat deze opgave niet heeft beantwoord. Deze correctie is logischerwijs te progressief omdat niet alleen gecorrigeerd wordt voor ontbrekende waarnemingen door tijdnood maar ook voor ontbrekende waarnemingen die het gevolg zijn van de moeilijkheid van de opgave. Uit de tabellen in Appendix B blijkt dat bij een aantal examens een groot percentage van de opgaven volgens het criterium te lage gecorrigeerde P-waarden heeft. De problemen doen zich voor bij de aanvullende toetsen van de volgende examens: alle jaren Frans VWO, met als extreem

2001, Frans HAVO in de examenopgaven van 2002, Duits VWO bij het referentiejaar en in 2001 en 2002, Duits HAVO in de ankeropgaven tot en met 2001 en Engels vwo 2001. (De lezer die geïnteresseerd is in welke specifieke opgaven in de aanvullende toetsen P-waarden hebben lager dan het criterium wordt verwezen naar Appendix C.)

Tijdnood

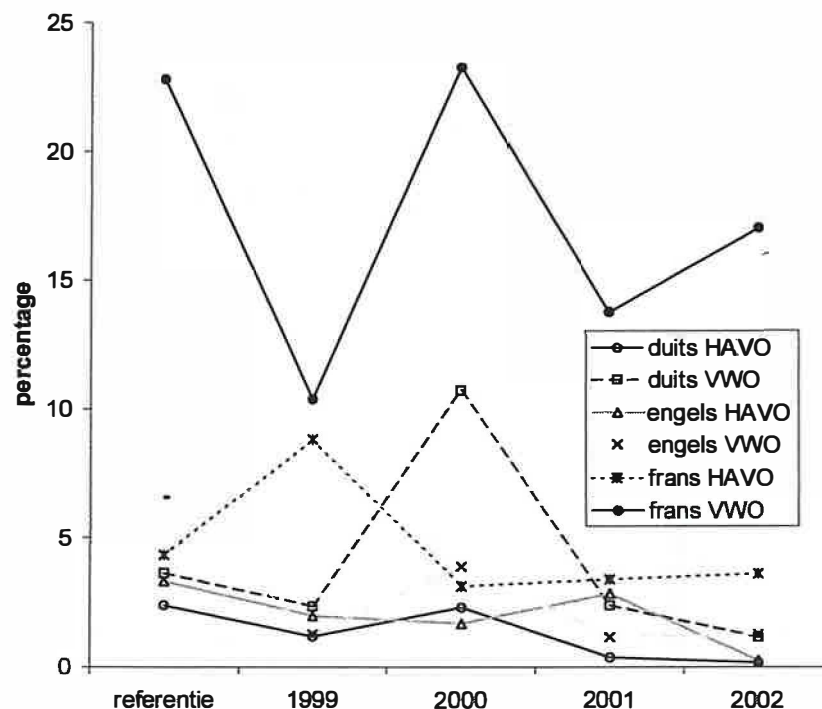
Bij de aanvullende toetsen worden vaak vragen op het eind van de toets overgeslagen. Dit is een aanwijzing voor tijdnood (maar wijst ook op afnamecondities die onvergelijkbaar zijn met een examenconditie, omdat het in examencondities van belang is voor de leerling om overall



Figuur 9 Gemiddeld percentage ontbrekende waarnemingen in elk van vier delen van de aanvullende toetsen.

een antwoord in te vullen). Om zicht te krijgen op de grootte van dit probleem zijn de gemiddelde percentages ontbrekende waarnemingen bepaald in vier delen van de aanvullende toetsen. Hiervoor zijn de opgaven uit de proeftoetsen op afnamevolgorde in vier zoveel mogelijk gelijke delen opgedeeld. Elk deel bevat ongeveer 25% van de opgaven, waarbij deel 1 het eerste kwart van de toets bevat, deel 2 het tweede, deel 3 het derde en deel 4 het laatste kwart van de toets. In Figuur 9 staan de resultaten van de proeftoetsen van elk van de HAVO en VWO examens MVT.

In Figuur 9 zijn de verschillende examenvakken slecht met elkaar te vergelijken, omdat per examen de schalen verschillen, waarbij het maximum bij Engels HAVO 3.5% is en bij Frans VWO 25 %. Verder stijgt het percentage ontbrekende waarnemingen aan het eind van de toetsen. Dit suggereert dat het om tijdnoed gaat en niet om overgeslagen vragen die te moeilijk werden gevonden. In Figuur 10 worden de gemiddelde percentages ontbrekende waarnemingen in het laatste deel van de proeftoetsen weergegeven. Duidelijk is dat het percentage ontbrekende waarnemingen bij Frans VWO erg hoog is en dat er verder relatief grote hoeveelheden ontbrekende waarnemingen voorkwamen bij Duits VWO 2000 en Engels VWO 1999.



Figuur 10 de gemiddelde percentages ontbrekende waarnemingen in het laatste deel van de proeftoetsen over de jaren heen voor elk van de vakken.

Observaties & Conclusies

Hierboven is een aantal problemen met betrekking tot de normhandhaving bij de examens MVT besproken. De invloed van elk van deze punten op zich is misschien verwaarloosbaar, maar gezamenlijk vormen ze een ernstige bedreiging voor de effectiviteit van de normhandhavingprocedure.

1. Een behoorlijk deel van de opgaven van het referentie-examen is niet opgenomen in de aanvullende toetsen. Het gevolg is een minder goede koppeling tussen het referentie-examen en de ankertoets. Met name bij Engels HAVO waar slechts 56% van de opgaven uit het referentie-examen is opgenomen in het design bedreigt dit de validiteit van de normhandhaving. Alleen wanneer het ontbrekende deel van de opgaven uit het referentie-

examen equivalent zijn met de in het design opgenomen opgaven kan namelijk worden waargemaakt dat de norm van het referentie-examen wordt overgebracht naar het nieuwe examen.

2. De steekproefgrootte van de aanvullende toetsen is vaak te klein.
3. De opgaven in de aanvullende toets worden te slecht gemaakt. Dit fenomeen treedt vooral op bij de opgaven Frans HAVO en VWO en Duits VWO. Ook na correctie voor ontbrekende waarnemingen is het aantal opgaven dat zeer slecht wordt gemaakt bij deze examens zorgwekkend groot.
4. Er treedt bij afname van de aanvullende toetsen regelmatig een grote mate van tijdnoed op. Dit komt het sterkst voor bij Frans VWO.

De data zoals die tot nu toe verzameld zijn bij Frans VWO zijn niet geschikt om een normhandhavingprocedure op te baseren. Hetzelfde geldt in mindere mate voor de data bij Duits VWO. Hier lijken twee van de vier aanvullende toetsen met opgaven uit het referentie-examen redelijk. Na aanpassing van de ankertoets in 2002 lijken de prestaties op deze opgaven goed, de prestaties op de eigenlijke examen opgaven zijn echter nog steeds verdacht. Verder zijn de data bij Frans HAVO in 2001 en 2002 qua geobserveerde moeilijkheid vergelijkbaar met de data bij Frans VWO en daardoor waarschijnlijk niet geschikt voor normhandhaving. Wel treed bij Frans HAVO minder tijdnoed op. Bij alle andere aanvullende toetsen komen is het noodzakelijk dat de ankeropgaven waarop slecht wordt gepresteerd onderzocht worden en dat eventueel aanpassingen worden doorgevoerd. Verder moet het tijdnoedprobleem worden opgelost.

Voorstellen

- Verhogen van het niveau van de steekproef die de aanvullende toetsen maakt door selectie van uitsluitend totaalvak leerlingen. Verder kan overwogen worden om woordenboek gebruik toe te staan (in combinatie met een verkorting van de toetsen). Hiervoor moet echter ook inhoudelijk advies ingewonnen worden om te kijken of dit geen consequenties heeft voor de moeilijkheidsgraad van de ankertoetsopgaven. Ik denk dat deze maatregel afdoende zou moeten zijn voor de HAVO examens. Bij de VWO-examens blijft het de vraag of het niveau van de 5-VWO-leerlingen voldoende is. Vooral bij Frans en Duits is dit mogelijk een onoplosbaar probleem (zie volgende punt).
- Als de normhandhavingprocedure op basis van het posttestdesign met een ankertest bij Frans VWO gehandhaafd blijft is het noodzakelijk om tussentijds extra toetsen af te nemen bij groepen leerlingen die in staat zijn om de opgaven te beantwoorden. De op dit moment verzamelde gegevens zijn kwalitatief zo twijfelachtig dat ze volgens ons ongeschikt zijn om te gebruiken in een normhandhavingprocedure. Waarschijnlijk is het beter te veronderstellen dat er geen verschillen optreden in de vaardigheid van de populatie over de jaren heen dan de verschillen op basis van deze gegevens serieus te nemen.
- Bij aanhoudende problemen met de posttest-data moeten de VWO examens Frans en Duits en het HAVO examen Frans opgenomen worden in een vorm van pre-testing.
- Waar mogelijk oplossen van tijdnoed door een inkorting van de toetsen. Hierbij moeten de ankertoetsen voldoende groot blijven om redelijk betrouwbaar de vaardigheid van de leerling te kunnen bepalen en moet de hoeveelheid nieuwe opgaven zo klein gehouden worden dat de toets binnen een lesuur gemaakt kan worden. Moeilijkheid hierbij zal zijn dat de vragen samenhangen met een tekst waardoor ze als groep moeten worden geselecteerd. Dit laatste beperkt de flexibiliteit.
- In de analyses kunnen overgeslagen opgaven aan het einde van de toets beschouwd worden alsof ze niet aangeboden zijn. Nu worden deze opgaven nog fout gerekend. Nadeel is dat je alleen op het eind van de toets corrigeert en dat je eigenlijk niet weet of de

leerling de opgaven onbeantwoord heeft gelaten omdat hij/zij er niet aan toe kwam of dat hij/zij de opgaven heeft geprobeerd te maken maar daar niet in is geslaagd.

- Vergroten van de steekproef. Mede naar aanleiding van het vorige punt is het noodzakelijk om de proeftoetsen aan meer leerlingen voor te leggen.
- Verzamelen van extra aanvullende gegevens om de ankertoets en het referentie-examen beter te verbinden. Bij de examens waarbij niet alle opgaven uit het referentie-examen en samen met de ankertoets zijn afgenomen moet onderzocht worden of de ontbrekende opgaven uit het referentie-examen statistisch vergelijkbaar zijn met de wel opgenomen opgaven. Zo niet dan moeten extra gegevens worden verzameld die het referentie-examen koppelen aan de ankertoets.
- Op inhoudelijke gronden bepalen of de probleemopgaven (zie Appendix C) deel moeten blijven uitmaken van de ankertoetsen.
- In het vervolg moet bij statistische analyses rekening gehouden worden met de veranderingen in de opgaven bij Engels. Opgaven waarbij de tekst van de opgaven of woorden in de te lezen tekst gewijzigd zijn moeten als niet identieke opgaven behandeld worden.

Referenties

Béguin, A.A. (2002). *Robustness of IRT Test Equating to violations of the Representativeness of the Common Items in a Nonequivalent Groups Design*. Paper presented at the 2002 annual meeting of the National Council of Measurement in Education, New Orleans, IL.

Béguin, A.A., Alberts, R.V.J., & Kremers, E.J.J. (2002). *Normhandhaving bij examens: hoe worden toetsen vergelijkbaar gemaakt?* Toetswijzer (beschikbaar via: <http://toetswijzer.kennisnet.nl/html/normering/home.htm>).

Appendix A

Engels HAVO

jaartal	boekje	Anker	Examen	kA	kE	N
referentie	1	32,1-8	21-31	9	11	291
	2	17-27	32-40	11	9	294
	3	9-16,28-31	2-9	12	8	291
1999	1	17-27	49-50,32-36	11	7	142
	2	1-8	2-11	8	10	141
	3	9-16	1,12-20	8	10	141
	4	9-12,28-31	21-31	8	11	149
	5	13-16,28-31	37-48	8	12	141
2000	1	1-8	45,4-9,42-44	8	10	177
	2	28-31,9-12	13-22	8	10	171
	3	17-27	23,34-35,37,39-41	11	7	176
	4	13-16,28-31	25-27,29-33,2-3	8	10	170
2001	1	1-8	1,10-19	8	11	108
	2	9-16	3,5-6,8-9,25,27-28,30-31	8	10	104
	3	17-27	32-35,37-40	11	8	106
	4	28-31,13-16	32-35,37-40,1	8	9	106
	5	9-16	3,5-6,8-9,25,27-28,30-31	8	10	96
2002	1	1-8	2-5,7-8,10-12	8	9	108
	2	9-16	13-21	8	9	106
	3	17-27	22,36,23,25-26,28-29	11	7	106

kA: aantal opgaven uit de ankertoets

kE: aantal opgaven uit het nieuwe examen

N: aantal personen in de afname

Engels VWO

jaartal	boekje	Anker	Examen	kA	kE	N
referentie	1	1-8	5-13,22-23	8	11	226
	2	17-28	14-21	12	8	217
	3	29-36,9-12	24-31	12	8	217
	4	13-16	39-49,2-4	4	14	217
1999	1	29-35	1,36-42	7	8	176
	2	17-28	2-8	12	7	174
	3	1-8	27-35	8	9	173
	4	13-16	9-16,43-49	4	15	174
	5	9-12	17-26,17	4	11	176
2000	1	29-35	1-2,35-42	7	10	172
	2	17-28	43-48	12	6	176
	3	1-8	13-22	8	10	173
	4	13-16	23-34	4	12	181
2001	5	9-12	3-12,49-50	4	12	174
	1	1-8	1-2,35,38,41-44	8	8	88
	2	9-12	16,18-22,24,24	4	8	86
	3	13-16	6-15	4	10	83
	4	17-28	45-46	12	2	83
2002	5	29-35	28-32,34	7	6	84
	1	29-35	4-12	7	9	193
	2	17-28	22-23,26,26	12	4	187
	3	1-8	16-17,19-21,31	8	6	188
	4	9-16	13,34-39	8	7	193

kA: aantal opgaven uit de ankertoets

kE: aantal opgaven uit het nieuwe examen

N: aantal personen in de afname

Duits HAVO

jaartal	boekje	Anker	Examen	kA	kE	N
referentie	1	16-19,27-32	1-9	10	9	119
	2	1-5,33,6,7	10-21	8	12	121
	3	8-15	22-29	8	8	118
	4	20-26	30-41	7	12	119
1999	1	16-19,27-32	1,8-16	10	10	80
	2	1-7	2-7,40-42	7	9	82
	3	8-15	17-18,43-50	8	10	85
	4	20-26	30-39	7	10	82
	5	1-7	19-29	7	11	81
2000	1	20-26	2-3,6-8,11,13-15	7	9	92
	2	1-7	16-25	7	10	92
	3	8-15	26-30,32,43-44	8	8	89
	4	16-19,27-32	35-36,38-40,43-44	10	7	88
2001	1	16-19,27-32	35-37,39-41	10	6	134
	2	1-7	19-27	7	9	129
	3	8-15	2-3,6,8-9,30-31,33	8	8	136
	4	20-26	10-13,15-17,15	7	8	129
2002	1	8-13,15	1-2,4-6,37-39	7	8	104
	2	8-13,15	8-13,15-16	7	8	100
	3	27-32	18-28	6	11	106
	4	20-24 (& 26 fout)	29,32-34,36-39	6	8	103

kA: aantal opgaven uit de ankertoets

kE: aantal opgaven uit het nieuwe examen

N: aantal personen in de afname

Duits VWO

jaartal	boekje	Anker	Examen	kA	kE	N
referentie	1	16-22,32,23	1-8	9	8	125
	2	24-31	20-27	8	8	129
	3	1-5,33,6,7	9-18	8	10	124
	4	8-15	37-40,42-49	8	12	124
1999	1	16-23	1-4,18-24	8	11	117
	2	24-31	5-12	8	8	121
	3	1-7	13-17,46-50	7	10	115
	4	8-15	25-35	8	11	120
	5	8-15	36-45	8	10	116
2000	1	1-7	1-6,44-50	7	13	101
	2	16-23	7-19	8	13	105
	3	24-31	20-31	8	12	94
	4	8-15	32-43	8	12	96
2001	1	16-23	9-13,15-17	8	8	81
	2	24-31	19-29	8	11	79
	3	1-7	31-34,36-38,40-41	7	9	71
	4	8-15	1,5,7,30,42-43,42	8	7	71
2002	1	8-13,15	9,12-13,17,27-28	7	6	176
	2	24-29,31	19-26	7	8	173
	3	16-18,20-23	29-31,33-34,36,38,40	7	8	168

kA: aantal opgaven uit de ankertoets

kE: aantal opgaven uit het nieuwe examen

N: aantal personen in de afname

Frans HAVO

jaartal	boekje	Anker	Examen	kA	kE	N
referentie	1	13-21	4-13	10	10	94
	2	1-12	34-42	12	9	91
	3	22-31	24-33	10	10	99
	4	22-31	1-3, 43-50	10	11	100
1999	1	22-31	1-10	10	10	72
	2	1-12	42-50	12	9	70
	3	22-31	11-20	10	10	69
	4	13-21	11-20	9	10	67
2000	5	1-12	32-41	12	10	67
	1	22-31	33-34, 36-38, 40-42, 44-45	10	10	100
	2	1-12	1-5, 7, 30-31	12	8	98
	3	13-21	18-27, 31	9	11	97
2001	4	22-31	10-11, 14-15, 17, 4-5, 4	10	8	84
	1	22-31	5-8, 27	10	5	79
	2	1-12	1, 3, 15-16	12	4	93
	3	13-21	19-26	9	8	85
2002	4	22-31	29-32, 28, 28	10	6	82
	5	13-21	10, 12-14, 38-39, 41-43	9	9	78
	1	22-31	7-8, 10-13, 22	10	7	77
	2	1-12	32-33, 35-37	12	5	81
	3	13-21	23-30	9	8	82
	4	22-31	16-21, 16	10	7	72

kA: aantal opgaven uit de ankertoets

kE: aantal opgaven uit het nieuwe examen

N: aantal personen in de afname

Frans VWO

jaartal	boekje	Anker	Examen	kA	kE	N
referentie	1	1-12	15-25	12	11	99
	2	13-23	1-9	11	9	90
	3	13-23	26-35	11	10	89
	4	24-32	10-14, 36-41	9	11	97
1999	1	1-12	9-17	12	9	103
	2	13-23	41-50	11	10	99
	3	24-32	1-8, 37-40	9	12	101
	4	1-12	18-26	12	9	100
2000	5	13-23	27-36	11	10	92
	1	1-12	1-9	12	9	99
	2	13-23	33-37, 47-50	11	9	97
	3	24-32	23-32, 22	9	11	97
2001	4	1-12	38-46	12	9	96
	5	13-23	10-21	11	12	98
	1	1-12	1-2, 4-6, 8-9	12	7	98
	2	13-23	22-27, 22	11	7	88
2002	3	13-23	34-37, 39, 21	11	6	88
	4	24-32	11-20	9	10	89
	1	1-12	17-19, 21-22	12	5	89
	2	13-23	7, 25-26, 29, 31-32	11	6	100
	3	24-32	8-16	9	9	92
	4	24-32	33, 35, 37-38, 41, 2-4	9	8	95

kA: aantal opgaven uit de ankertoets

kE: aantal opgaven uit het nieuwe examen

N: aantal personen in de afname

Appendix B

Engels HAVO

Ankertoets

jaartal	boekje	kA	kE	N	%mis A	%mis E	%cmis A	%cmis E
referentie	1	9	11	291	11	0	11	0
	2	11	9	294	9	11	9	11
	3	12	8	291	8	0	8	0
1999	1	11	7	142	9	0	9	0
	2	8	10	141	12	0	12	0
	3	8	10	141	0	10	0	10
	4	8	11	149	0	9	0	9
	5	8	12	141	0	0	0	0
2000	1	8	10	177	12	10	12	10
	2	8	10	171	0	0	0	0
	3	11	7	176	9	14	9	14
	4	8	10	170	0	0	0	0
2001	1	8	11	108	12	9	12	9
	2	8	10	104	0	0	0	0
	3	11	8	106	9	0	9	0
	4	8	9	106	0	0	0	0
	5	8	10	96	0	0	0	0
2002	1	8	9	108	12	0	12	0
	2	8	9	106	0	0	0	0
	3	11	7	106	9	14	9	14

kA:aantal opgaven uit de ankertoets, kE:aantal opgaven uit het nieuwe examen
 N:aantal personen in de afname, %mis A:percentage ankertoetsopgaven met te kleine p-waarde
 %mis E:percentage examenopgave met te kleine p-waarde, %cmis A: %mis A gecorrigeerd voor missing, %cmis E:%mis E gecorrigeerd voor missing

Examen

jaartal	K	N	%mis E	%cmis E
Referentie	50	2039	4	4
1999	50	4000	4	4
2000	37	1801	8	8
2001	29	1885	6	6
2002	25	2019	16	16

K:aantal opgaven in het nieuwe examen, N:aantal personen in de afname, %mis E:percentage examenopgave met te kleine p-waarde, %cmis E:%mis E gecorrigeerd voor missing

Engels VWO

Ankertoets

jaartal	boekje	kA	kE	N	%mis A	%mis E	%cmis A	%cmis E
referentie	1	8	11	226	12	36	12	27
	2	12	8	217	8	0	8	0
	3	12	8	217	25	25	8	25
	4	4	14	217	25	28	25	28
1999	1	7	8	176	0	12	0	12
	2	12	7	174	16	0	16	0
	3	8	9	173	12	11	12	11
	4	4	15	174	0	26	0	26
	5	4	11	176	0	18	0	18
2000	1	7	10	172	14	30	0	30
	2	12	6	176	16	16	16	16
	3	8	10	173	12	0	12	0
	4	4	12	181	0	50	0	50
	5	4	12	174	0	0	0	0
2001	1	8	8	88	12	25	12	25
	2	4	8	86	0	12	0	12
	3	4	10	83	0	20	0	20
	4	12	2	83	16	50	16	50
	5	7	6	84	14	16	14	16
2002	1	7	9	193	14	22	14	22
	2	12	4	187	16	25	16	25
	3	8	6	188	12	16	12	16
	4	8	7	193	0	0	0	0

KA:aantal opgaven uit de ankertoets, kE:aantal opgaven uit het nieuwe examen

N:aantal personen in de afname, %mis A:percentage ankertoetsopgaven met te kleine p-waarde

%mis E:percentage examenopgave met te kleine p-waarde, %cmis A: %mis A gecorrigeerd voor missing, %cmis E:%mis E gecorrigeerd voor missing

Examen

Jaartal	K	N	%mis E	%cmis E
referentie	50	2018	2	2
1999	50	4000	2	2
2000	50	4000	12	12
2001	34	1407	2	2
2002	26	2000	3	3

K:aantal opgaven in het nieuwe examen, N:aantal personen in de afname, %mis E:percentage examenopgave met te kleine p-waarde, %cmis E:%mis E gecorrigeerd voor missing

Duits HAVO

Ankertoets

Jaartal	boekje	kA	kE	N	%mis A	%mis E	%cmis A	%cmis E
referentie	1	10	9	119	0	0	0	0
	2	8	12	121	37	16	37	16
	3	8	8	118	0	12	0	12
	4	7	12	119	14	16	14	16
1999	1	10	10	80	0	0	0	0
	2	7	9	82	42	0	42	0
	3	8	10	85	0	0	0	0
	4	7	10	82	14	10	14	10
	5	7	11	81	28	9	28	9
2000	1	7	9	92	14	11	14	11
	2	7	10	92	42	20	42	20
	3	8	8	89	12	0	12	0
	4	10	7	88	10	0	10	0
2001	1	10	6	134	0	16	0	16
	2	7	9	129	42	0	42	0
	3	8	8	136	0	0	0	0
	4	7	8	129	0	25	0	25
2002	1	7	8	104	0	12	0	12
	2	7	8	100	0	0	0	0
	3	6	11	106	16	9	16	9
	4	6	8	103	16	12	16	12

KA:aantal opgaven uit de ankertoets, kE:aantal opgaven uit het nieuwe examen

N:aantal personen in de afname, %mis A:percentage ankertoetsopgaven met te kleine p-waarde

%mis E:percentage examenopgave met te kleine p-waarde, %cmis A: %mis A gecorrigeerd voor missing, %cmis E:%mis E gecorrigeerd voor missing

Examen

Jaartal	K	N	%mis E	%cmis E
referentie	50	2129	10	10
1999	50	4000	4	4
2000	33	1699	12	12
2001	31	1916	16	16
2002	34	1866	14	14

K:aantal opgaven in het nieuwe examen, N:aantal personen in de afname, %mis E:percentage examenopgave met te kleine p-waarde, %cmis E:%mis E gecorrigeerd voor missing

Duits VWO

Ankertoets

jaartal	boekje	kA	kE	N	%mis A	%mis E	%cmis A	%cmis E
referentie	1	9	8	125	44	12	44	12
	2	8	8	129	0	25	0	25
	3	8	10	124	50	30	50	30
	4	8	12	124	12	16	12	16
1999	1	8	11	117	37	18	37	18
	2	8	8	121	12	12	12	12
	3	7	10	115	42	20	42	20
	4	8	11	120	0	27	0	27
	5	8	10	116	0	40	0	40
2000	1	7	13	101	57	23	28	23
	2	8	13	105	25	69	25	53
	3	8	12	94	25	33	12	33
	4	8	12	96	0	41	0	33
2001	1	8	8	81	50	50	50	50
	2	8	11	79	37	45	37	36
	3	7	9	71	71	22	57	22
	4	8	7	71	0	14	0	14
2002	1	7	6	176	0	16	0	16
	2	7	8	173	14	50	14	50
	3	7	8	168	28	25	28	25

KA:aantal opgaven uit de ankertoets, kE:aantal opgaven uit het nieuwe examen
 N:aantal personen in de afname, %mis A:percentage ankertoetsopgaven met te kleine p-waarde
 %mis E:percentage examenopgave met te kleine p-waarde, %cmis A: %mis A gecorrigeerd voor
 missing, %cmis E:%mis E gecorrigeerd voor missing

Examen

Jaartal	K	N	%mis E	%cmis E
referentie	50	2012	8	8
1999	50	4000	20	20
2000	50	4000	14	14
2001	35	1570	14	14
2002	24	1788	25	25

K:aantal opgaven in het nieuwe examen, N:aantal personen in de afname, %mis E:percentage
 examenopgave met te kleine p-waarde, %cmis E:%mis E gecorrigeerd voor missing

Frans HAVO

Ankertoets

Jaartal	boekje	kA	kE	N	%mis A	%mis E	%cmis A	%cmis E
referentie	1	10	10	94	10	0	10	0
	2	12	9	91	16	0	8	0
	3	10	10	99	0	0	0	0
	4	10	11	100	0	0	0	0
1999	1	10	10	72	0	0	0	0
	2	12	9	70	0	22	0	11
	3	10	10	69	0	10	0	10
	4	9	10	67	0	10	0	10
	5	12	10	67	25	20	8	20
2000	1	10	10	100	0	10	0	10
	2	12	8	98	0	12	0	12
	3	9	11	97	11	9	11	9
	4	10	8	84	0	12	0	12
2001	1	10	5	79	40	20	30	20
	2	12	4	93	33	25	33	25
	3	9	8	85	22	25	22	25
	4	10	6	82	30	33	30	33
	5	9	9	78	55	22	55	22
2002	1	10	7	77	20	28	10	28
	2	12	5	81	25	60	25	60
	3	9	8	82	44	37	22	37
	4	10	7	72	20	42	20	42

KA:aantal opgaven uit de ankertoets, kE:aantal opgaven uit het nieuwe examen

N:aantal personen in de afname, %mis A:percentage ankertoetsopgaven met te kleine p-waarde

%mis E:percentage examenopgave met te kleine p-waarde, %cmis A: %mis A gecorrigeerd voor missing, %cmis E:%mis E gecorrigeerd voor missing

Examen

Jaartal	K	N	%mis E	%cmis E
referentie	50	2144	4	4
1999	50	4000	10	10
2000	34	1722	0	0
2001	32	1790	31	31
2002	32	1910	40	40

K:aantal opgaven in het nieuwe examen, N:aantal personen in de afname, %mis E:percentage examenopgave met te kleine p-waarde, %cmis E:%mis E gecorrigeerd voor missing

Frans VWO

Ankertoets

jaartal	boekje	kA	kE	N	%mis A	%mis E	%cmis A	%cmis E
referentie	1	12	11	99	0	72	0	45
	2	11	9	90	72	11	63	11
	3	11	10	89	63	80	63	30
	4	9	11	97	44	27	11	27
1999	1	12	9	103	0	22	0	22
	2	11	10	99	54	40	54	20
	3	9	12	101	11	33	11	33
	4	12	9	100	0	66	0	44
	5	11	10	92	72	30	63	30
2000	1	12	9	99	25	33	0	22
	2	11	9	97	63	77	54	44
	3	9	11	97	44	9	22	0
	4	12	9	96	8	33	0	22
	5	11	12	98	72	41	54	41
2001	1	12	7	98	41	28	25	28
	2	11	7	88	72	42	63	42
	3	11	6	88	72	66	72	66
	4	9	10	89	33	70	33	70
2002	1	12	5	89	25	20	0	20
	2	11	6	100	54	50	54	50
	3	9	9	92	44	11	22	11
	4	9	8	95	22	37	22	37

KA:aantal opgaven uit de ankertoets, kE:aantal opgaven uit het nieuwe examen
 N:aantal personen in de afname, %mis A:percentage ankertoetsopgaven met te kleine p-waarde
 %mis E:percentage examenopgave met te kleine p-waarde, %cmis A: %mis A gecorrigeerd voor
 missing, %cmis E:%mis E gecorrigeerd voor missing

Examen

Jaartal	K	N	%mis E	%cmis E
referentie	50	2154	10	10
1999	50	4000	12	12
2000	50	4000	10	10
2001	32	1522	9	9
2002	28	1849	14	14

K:aantal opgaven in het nieuwe examen, N:aantal personen in de afname, %mis E:percentage
 examenopgave met te kleine p-waarde, %cmis E:%mis E gecorrigeerd voor missing

Appendix C

Ankeropgaven met p-waarde dicht bij de gokkans

Engels HAVO

De volgende opgaven scoren in alle boekjes dicht bij de gokkans:

label	P	aantal alternatieven
Anker.04	0.36	3
Anker.20	0.39	4

Bijzonderheden:

Een aantal opgaven is veranderd tussen de originele afname in 1998 en de latere afnamen.

Boekje 1:

Ankeritem 1: in de latere afnamen heeft dit item 4 i.p.v. 5 categorieën

Ankeritem 4: de tekst (regel 60) is aangepast

Boekje 3:

Ankeritem 10: de tekst van alternatief c is gewijzigd

Ankeritem 14: de tekst is aangepast

Engels VWO

de volgende opgaven scoren in alle boekjes dicht bij de gokkans:

label	P	aantal alternatieven
Anker.04	0.41	3
Anker.17	0.35	3
Anker.24	0.44	3
Anker.35	0.44	3

Bijzonderheden:

Een aantal opgaven is veranderd tussen de originele afname in 1998 en de latere afnamen.

Boekje 1:

Ankeritem 1: in de latere afnamen heeft dit item 4 ipv 5 categorieën

Ankeritem 4: de tekst (regel 60) is aangepast

Boekje 3:

Ankeritem 29: in de latere afnamen heeft dit item 3 ipv 4 categorieën

Boekje 4:

Ankeritem 10: de tekst van alternatief c is gewijzigd

Ankeritem 14: de tekst is aangepast

Duits HAVO

De volgende opgaven scoren in alle boekjes dicht bij de gokkans:

Label	P	aantal alternatieven
Anker.02	0.24	3
Anker.05	0.46	3
Anker.06	0.33	4

Verder scoren de volgende opgaven dicht bij de gokkans in de volgende afnamen:

Label	P	aantal alternatieven	afnamejaar
Anker.04	0.39	4	1999
Anker.14	0.37	4	2000
Anker.23	0.39	4	1998, 1999, 2000 en 2002
Anker.31	0.44	3	2000 en 2002

Bijzonderheden:

In boekje 4 van de 2002 afname was het laatste item (ankeritem 26) verkeerd opgenomen en dit item vervalt daardoor

Duits VWO

De volgende opgaven scoren in alle boekjes dicht bij de gokkans:

label	p	aantal alternatieven
Anker.02	0.15	3
Anker.05	0.27	3
Anker.06	0.23	4
Anker.18	0.42	3
Anker.19	0.19	5
Anker.22	0.44	3
Anker.26	0.35	4

Verder scoren de volgende opgaven dicht bij de gokkans in de volgende afnamen:

label	P	aantal alternatieven	afnamejaar
Anker.03	0.31	4	2001
Anker.07	0.39	4	2000 en 2001 T
Anker.15	0.36	4	1998
Anker.20	0.33	4	2001
Anker.24	0.44	3	2001
Anker.25	0.37	3	2001
Anker.31	0.38	4	2000 T

Een T achter het jaartal geeft aan dat bij deze opgave in dit jaar sprake was van tijdnood. Gecorrigeerd voor tijdnood ligt de p-waarde NIET dicht bij de gokkans

Bijzonderheden:

In de 1998 afname boekje 1 is **Ankeritem 32** na de originele afname niet meer afgenomen en ook bij de originele afname verwijderd uit de datafile

In de 1998 afname boekje 3 is **Ankeritem 33** na de originele afname niet meer afgenomen en ook bij de originele afname verwijderd uit de datafile

In de 2002 afname boekje 3 is een stukje tekst weggelaten dit heeft waarschijnlijk weinig gevolgen voor de opgaven

Frans HAVO

De volgende opgaven scoren dicht bij de gokkans in de volgende afnamen:

label	P	aantal alternatieven	afnamejaar
Anker.04	0.31	4	2001 en 2002
Anker.08	0.39	4	1999 en 2001
Anker.10	0.32	4	2002
Anker.11	0.38	4	1998,1999,2001 en 2002
Anker.12	0.29	4	1998,1999 en 2001
Anker.16	0.36	4	2001
Anker.17	0.37	4	2001 en 2002
Anker.19	0.38	4	2001 en 2002T
Anker.20	0.19	4	2000, 2001 en 2002
Anker.21	0.31	4	2001 en 2002T
Anker.27	0.33	4	2001 en 2002T
Anker.29	0.35	4	2001
Anker.30	0.30	4	2001 en 2002
Anker.31	0.34	5	2001T en 2002

Een T achter het jaartal geeft aan dat bij deze opgave in dit jaar sprake was van tijdnoed. Gecorrigeerd voor tijdnoed ligt de p-waarde NIET dicht bij de gokkans

Bijzonderheden:

In de 1998 afname boekje 1 is **Ankeritem 33** na de originele afname niet meer afgenomen en ook bij de originele afname verwijderd uit de datafile

Frans VWO

De volgende opgaven scoren in alle boekjes dicht bij de gokkans:

label	P	aantal alternatieven
Anker.13	0.33	4
Anker.14	0.16	4
Anker.18	0.36	3
Anker.20	0.25	3
Anker.21	0.32	3
Anker.22	0.29	3
Anker.26	0.30	4

Verder scoren de volgende opgaven dicht bij de gokkans in de volgende afnamen:

label	P	aantal alternatieven	afnamejaar
Anker.07	0.38	4	2001T
Anker.08	0.26	4	200T en 2001
Anker.10	0.37	4	2001T en 2002T
Anker.11	0.30	4	2000T,2001 en 2002T
Anker.12	0.36	4	2000T,2001 en 2002T
Anker.15	0.36	4	1998T,1999,2000 en 2001
Anker.23	0.27	4	1998T, 1999T, 2000 en 2001
Anker.25	0.39	4	2002
Anker.30	0.36	4	1998T,2000,2001 en 2002
Anker.31	0.44	3	1998T en 2000T
Anker.32	0.32	4	2000T 2001 en 2002

Een T achter het jaartal geeft aan dat bij deze opgave in dit jaar sprake was van tijdnoed. Gecorrigeerd voor tijdnoed ligt de p-waarde NIET dicht bij de gokkans

