Gesegmenteerde online beoordeling, beoordelaarseffecten en feedback aan beoordelaars

Pilot Spreken

Mei 2007

Henk Kuijper
Anne-Marie Anthonissen
Timo Bechger
Gunter Maris

Inhoud

1 Inleiding

Bij het *Staatsexamen voor Nederlands als tweede taal* wordt bij de productieve vaardigheid Spreken gebruik gemaakt van open vragen. Kandidaten geven antwoord op deze vragen en de gegeven antwoorden worden beoordeeld door twee beoordelaars. Ieder antwoord wordt beoordeeld op verschillende deelaspecten. Aan sommige aspecten zijn dichotoom, andere zijn polytoom.
In eerdere onderzoeksrapporten (Bechger, Maris 2004 en 2006) is uiteengezet dat het beoordelen van open vragen door menselijke beoordelaars een aantal complicaties met zich meebrengt:

– Beoordelaarovereenstemming:
  Bij de vaardigheid Spreken is de mate van beoordelaarovereenstemming gering. Dit wordt onder andere veroorzaakt door:
  - Onvoldoende goede beoordelingsvoorschriften:
    - de onderscheiden beoordelingsaspecten behoeven relatief veel toelichting
    - tussen verschillende scores van één beoordelingsaspect bestaat verwarring.
    - de beoordelingsvoorschriften zijn relatief complex
      (Bechger, Maris, 2004, p.9)
    - Slechte toepassing van het beoordelingsvoorschrift
  - Bij Spreken is voor beide evidentie gevonden (Bechger, Maris, 2004, p.3) . Door een herziening van de beoordelingsvoorschriften (dichotomiseren), het hertrainen van de beoordelaars kan de beoordelingsovereenstemming vergroot worden.

– Feedback aan beoordelaars:
  Het is belangrijk om beoordelaars te controleren, te informeren over de kwaliteit van hun oordelen en waar mogelijk te stimuleren om correcte beoordelingen uit te voeren (Bechger, Maris, 204, p. 10).
  Het is hiervoor noodzakelijk dat beoordelaars at random aan kandidaten worden toegekend.
  De feedbackbrief kan, naast informatie over het functioneren van de betreffende beoordelaar ten opzichte van de gemiddelde beoordelaar ook informatie bevatten over de contra-beoordelaar (zie Bechger, Maris 2006, p.18 ev). Daarnaast kan er feedback worden gegeven over hoe de beoordelaar logistiek gezien functioneert. Binnen het onderzoeksbudget van 2006 is onderzocht hoe kandidaten random kunnen worden toegewezen aan beoordelaars. Ook is de bestaande feedbackbrief uitgebreid met informatie over de contrabeoordelaar. Met ingang van 2006 krijgen beoordelaars informatie over hun logistieke functioneren.

– Psychometrische theorie:
  Er moet verder gewerkt worden aan een uitgewerkte en algemeen aanvaarde psychometrische theorie voor het scoren van open vragen, waarbij expliciet rekening gehouden wordt met enerzijds beoordeelbaarheid, moeilijkheid en het discriminerend vermogen van items en anderzijds met beoordeelbaarheid en vaardigheid van kandidaten.
  Hierbij is een random toewijzing van beoordelaars aan opgave/kandidaat-combinaties een voorwaarde.

– Beoordelaarseffecten:
  Het meten van de productieve vaardigheden (schrijven en spreken) met menselijke beoordelaars brengt onvermijdelijk subjectiviteit met zich mee en introduceert oneerlijkheid voor kandidaten, doordat de score van een kandidaat niet alleen wordt bepaald door zijn of haar prestatie, maar ook door de eigenschappen van de beoordelaar (zie bijlage 1).
  De huidige beoordelingswijze bestaat eruit dat elke beoordelaar een heel examen van een kandidaat beoordeelt. Bij een beoordelingswijze waarbij de reacties van één kandidaat door meerdere beoordelaars worden beoordeeld (= segmentbeoordeling), nemen beoordelaarseffecten af of worden ze geneutraliseerd. Zo zal het halo-effect per definitie verdwijnen, omdat een beoordelaar niet meer het gehele examen van een kandidaat beoordeelt. Andere beoordelaarseffecten zoals bijvoorbeeld strengheid en het sequentie-effect zullen geneutraliseerd worden, omdat elke kandidaat door veel beoordelaars beoordeeld wordt.

Gezien de logistieke complexiteit van het random toekennen van kandidaatuitingen aan meerdere beoordelaars is een volledig digitale afname van de spreekexamens vereist. Met ingang van 2007 worden alle spreekexamens digitaal afgenomen. Segmentbeoordeling is daarmee in principe mogelijk geworden. In dit rapport wordt antwoord gegeven op de vraag of een gesegmenteerde online beoordeling met random toewijzing ook praktisch realiseerbaar is.

2 Doel van de pilot

De pilot met betrekking tot het gesegmenteerd beoordelen is opgezet om te onderzoeken:
– hoe digitaal afgenomen spreekexamens kunnen worden verknipt (hoofdstuk 4) om ze
– at random aan een groep beoordelaars van het Staatsexamen NT2 toe te kennen (hoofdstuk 4) en
– de uitingen online te laten beoordelen (hoofdstuk 5)
Daarnaast onderzoeken we
– de aanwezigheid en gevolgen van halo-effecten (7.1).
– of gesegmenteerde beoordeling van invloed is op de beoordelaarsovereenstemming (7.2) en
– de samenhang tussen de examenscores tot stand gekomen met behulp van de reguliere beoordelingswijze en de examenscores bij het gesegmenteerd beoordelen (7.1).

.
Deze pilot is uitgevoerd om te onderzoeken of de volgende aanbevelingen (Bechger, Maris, 2006, pagina 17) uitvoerbaar zijn.

> *"Willekeurige Toewijzing van Kandidaten aan Beoordelaars*
> *Het is duidelijk dat de systematische wijze van toewijzing de evaluatie van beoordelaars bemoeilijkt. Voordat er verder onderzoek kan worden gedaan naar manieren om afwijkende beoordelaars op te sporen moet er sprake zijn van willekeurige toewijzing. Om halo-effecten te vermijden moeten we nog iets verder gaan en beoordelaars toewijzen aan combinaties van opgaven en kandidaten. Daarnaast is willekeurige toewijzing van beoordelaars aan opgave/kandidaat combinaties een voorwaarde voor het verantwoord schalen (of kalibreren) van de data. Hierover hebben we eerder bericht in een interne notitie (Maris en Bechger, 2004).*
>
> *Verbeteren van de Criteria Voor het Opsporen van Afwijkende Beoordelaars*
> *In Maris en Bechger (2004) wordt betoogd dat willekeurige toewijzing ervoor zorgt dat de identiteit van individuele beoordelaars geen rol speelt bij het schalen. Dit lijkt in tegenspraak met de conclusie dat willekeurige toewijzing van essentieel belang is om afwijkende beoordelaars te kunnen opsporen. De tegenspraak verdwijnt wanneer we ons realiseren dat het doel van kalibratie anders is dan het doel van de beobrief. Bij het bepalen van de examencijfers willen we de invloed van afwijkende beoordelaars zo klein mogelijk maken. Verder onderzoek is erop gericht om hetzelfde kalibratiemodel te gebruiken om afwijkende beoordelaars op te sporen.*
>
> *Het Online Verzamelen van Beoordelingen*
> *Beoordelaars werken doorgaans in de anonimiteit van hun eigen huis. Dit impliceert dat we niet kunnen meekijken om te controleren wat een beoordelaar feitelijk doet. Dit zou verbeteren wanneer de beoordelingen elektronisch verzameld zouden worden. Bij de huidige stand van de techniek en de beschikbaarheid van computers ligt het voor de hand om gebruik te maken van het Internet.*
>
> *Een elektronisch beoordelingsvoorschrift is in feite een computerprogramma dat een spraak- of schrijffragment presenteert, er vragen over stelt op het scherm, en de reacties van beoordelaars registreert. Met behulp van een dergelijk programma kunnen de beoordelaars, bijvoorbeeld, op de volgende wijze worden gecontroleerd:*
>
> - *Beoordelaars kunnen worden gedwongen alle vragen te beantwoorden in een vaste volgorde*
> - *We kunnen vaststellen hoeveel tijd een beoordelaar bezig is.*
> - *We kunnen vaststellen of en hoe vaak een beoordelaar een spraakfragment heeft beluisterd.*
> - *Door gebruik te maken van een digitaal antwoordblad, zullen fouten bij het aanstrepen en inlezen verdwijnen.*
>
> *Andere voordelen van het elektronisch verzamelen zijn:*
> - *De logistieke verwerking is eenvoudiger, beter te controleren, en minder arbeidsintensief.*
> - *De beoordelingen zijn meteen ingevoerd en hoeven niet achteraf te worden ingelezen.*
> - *Het is voor beoordelaars gemakkelijker om spraakfragmenten opnieuw te beluisteren*
> - *Willekeurige toewijzing is relatief eenvoudig te realiseren.*
>
> *In het bijzonder voor het toewijzen van verschillende opgaven aan verschillende beoordelaars geld dat dit eigenlijk alleen maar gerealiseerd kan worden wanneer spraak en schrijffragmenten elektronisch worden opgeslagen. Het ligt dan voor de hand om deze ook elektronisch aan beoordelaars voor te leggen."*

3 Opzet van de pilot

Selectie van kandidaten
Van 50 willekeurige examenkandidaten van Spreken Programma II, middagafname juli 2006 zijn de prestaties opnieuw, maar nu gesegmenteerd, beoordeeld door 10 beoordelaars met behulp van een experimentele versie van een online beoordelingssysteem.

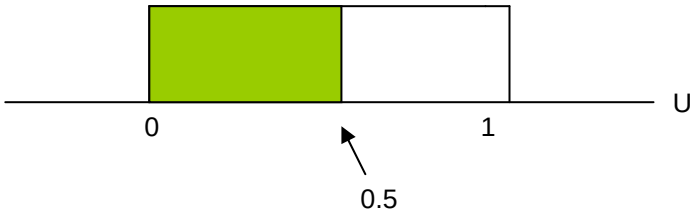
Segmenteren examenuitingen en random toekenning aan beoordelaren
De examenuitingen van elk van de kandidaten zijn per spreekopgave opgeknipt. Vervolgens werd elke uiting via een algoritme random toegekend aan beoordelaars. Bij deze toewijzing werd ervoor gezorgd dat de verhouding korte, middellange en lange opgave voor elke beoordelaar in dezelfde proportie voorkwam als in een compleet examen.

4 Toewijzing op basis van toeval

Als we spreken over *random* of *willekeurig* toewijzen bedoelen we dat de toewijzing gebeurt door middel van een toevalsexperiment. Een *toevalsexperiment* is een experiment waarbij de uitkomst niet met zekerheid kan worden voorspeld voordat het experiment heeft plaats gevonden. Bij een toevalsexperiment hoort een uitkomstruimte. De *uitkomstruimte* beschrijft alle mogelijke uitkomsten van het experiment. Bijvoorbeeld, als we een munt opwerpen dat zijn er twee mogelijke uitkomsten: KOP of MUNT.

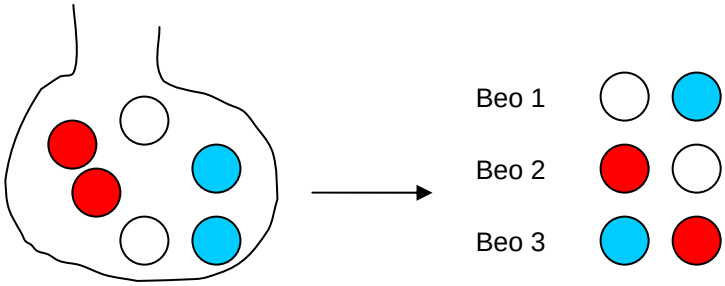
In het kader van het huidige onderzoek is een toevalsexperiment uitgevoerd waarbij kandidaat-opgave combinaties dusdanig over beoordelaars zijn verdeeld dat de kans op elke verdeling even waarschijnlijk is.

In de praktijk wordt de computer gebruikt om een toevalsexperiment te simuleren. Hiertoe vraagt men de computer om een willekeurig nummer tussen 0 en 1 zodanig dat elke waarde van 0 en 1 even waarschijnlijk is. We noemen dit een *toevalsgetal* en duiden dit getal aan met U.



Om een worp met een zuivere munt te simuleren observeren we of U kleiner is dan ½. Als dat zo is noemen we de uitkomst KOP, anders MUNT. In bovenstaande figuur wordt geïllustreerd dat de kans dat U<0,5 (KOP) gelijk is aan 0,5: het donkere gebied is even groot als het lichte gebied.

De Pot
Het experiment dat is uitgevoerd is gelijk aan het random trekken van *2 keer n* balletjes uit een pot met daarin even zoveel balletjes. Elk balletje correspondeert met een combinatie van een kandidaat en een opgave.



Er zijn *n* verschillende combinaties van kandidaten en opgaven; *n* is dus het product van het aantal kandidaten en het aantal opgaven. Het aantal balletjes is gelijk aan 2 keer *n* omdat elke combinatie van een kandidaat en een opgave twee keer wordt beoordeeld. In bovenstaande figuur zien we een pot met 6 balletjes corresponderend met combinaties van drie kandidaten en één opgave (n=3).

Elke van de beschikbare beoordelaars heeft aangegeven hoeveel examens/kandidaten hij of zij wil beoordelen. We noemen dit de *werklast*. De werklast van beoordelaar b wordt aangeduid met w_b.
Het experiment bestaat uit het toekennen van w_b balletjes aan elk van de beoordelaars. We beginnen met één beoordelaar om het principe te illustreren.

Elk van de *2n* balletjes in de pot moet worden getrokken en toegewezen aan onze beoordelaar. We trekken *zonder teruglegging*. Als een balletje is getrokken, wordt het niet meer in de pot teruggestopt zodat het niet meer mogelijk is om hetzelfde balletje opnieuw te trekken.

Om de trekking volgens toeval te doen verlopen gebruiken we w_b toevalsgetallen. Voor iedere trekking is een nieuw toevalsgetal nodig. Om het eerste balletje te kiezen doen we het volgende
1. De balletjes worden genummerd van 1 tot *2n*.
2. we genereren een toevalsgetal tussen 1 en *2n*.
3. We kiezen het corresponderende balletje.
4. We verwijderen het gekozen balletje.

Het gekozen balletje wordt uit de pot verwijderd zodat er nog maar *2n-1* balletjes in de pot zitten.
Om het tweede balletje te kiezen doen we het volgende
1. De balletjes worden genummerd van 1 tot *2n-1*.
2. We genereren een toevalsgetal tussen 1 en *2n-1*.
3. We kiezen het bijbehorende balletje.
4. We verwijderen het tweede balletje.

Als we zo verder gaan dan zien we dat we het volgende moeten doen om w_b balletjes te trekken.
1. *t=1*
2. Nummer de overige balletjes van 1 tot *2n-t+1*
3. Genereer een toevalsgetal tussen 1 en *2n-t+1*
4. Kies het bijbehorende balletje.
5. Verwijder het gekozen balletje
6. Als $t=w_b$ dan stoppen we. Als $t<w_b$, wordt *t* gelijk gezet aan *t+1* en gaan we terug naar stap 2.

Bij meerdere beoordelaars verandert er niets wezenlijk. Nadat we een beoordelaar voldoende balletjes hebben toegekend gaan we verder met de volgende. Het aantal balletjes in de pot is dan verminderd met het aantal toegewezen balletjes.

Het afwijsprincipe
Er zijn echter uitkomsten die we niet willen. Het kan bijvoorbeeld voorkomen dat een beoordelaar twee keer dezelfde kandidaat-opgave combinatie beoordeeld. Bij het schema zoals we dit hierboven hebben geschetst kan dit wel gebeuren. We hebben met andere woorden een *uitkomstruimte gekozen die eigenlijk te groot is*.

De oplossing is eenvoudig. Als een ongewenste uitkomst zich voordoet dan negeren we die en voeren het experiment opnieuw uit. Dit staat bekent als *het afwijsprincipe*. Het gevolg is dat we de uitkomstruimte hebben beperkt tot die uitkomsten die we willen en wel zodanig dat elke uitkomst nog steeds even waarschijnlijk is.

Het afwijsprincipe werkt altijd maar de efficiëntie ervan hangt af van het percentage niet toegestane uitkomsten. Als dit percentage hoog is dan duurt het lang voordat we een uitkomst hebben die voldoet. In dit geval hangt de efficiëntie af van het aantal kandidaten, opgaven en beoordelaars. In de praktijk zijn deze aantallen groot genoeg zodat de kans op een niet toegestane oplossing relatief klein is.

We moeten wel definiëren hoe de uitkomstruimte er dan uit ziet:
1. Beoordelaars kunnen niet twee keer dezelfde combinatie van een kandidaat en een opgave beoordelen.
2. Beoordelaars moeten nooit twee keer dezelfde kandidaat beoordelen.
3. etc.

In de pilot hebben we alleen de eerste eis opgelegd. We staan toe dat een beoordelaar incidenteel *verschillende* opgaven van *dezelfde* kandidaat beoordeeld.

Meerdere opgaven
Bij de pilot is het uitgangspunt dat beoordelaars steeds een geheel examen beoordelen. Bij het examen *Spreken II* wil dat zeggen: 6 korte opgaven, 7 middellange opgaven en 1 lange opgave. We kunnen twee dingen doen:
1. Doe het experiment voor elke opgave
2. Doe het experiment voor elke soort opgave

Kiezen we voor het eerste dan is gegarandeerd dat een beoordelaar een identiek en volledig examen krijgt te beoordelen, hoewel de opgaven door andere kandidaten zijn gemaakt. Kiezen we voor de tweede optie dan krijgt elke beoordelaar het juiste aantal van elke soort opgaven voorgelegd maar deze opgaven zijn niet allemaal verschillend.

We hebben gekozen voor de tweede optie, omdat daarmee het aantal mogelijke uitkomsten groter is.

5 Het online beoordelingssysteem

Gezien het beperkte budget is ervoor gekozen om niet te investeren in een software programma. Er is gebruik gemaakt van een bestaand programma.
Het Cito beschikte over het programma CVE (commentaar verzamel engine), waarmee online oordelen van beoordelaars konden worden verzameld. Deze tool is aangepast om te voldoen aan de eisen in deze pilot.

– Er is een nieuw introductiescherm gemaakt. (zie bijlage 2).Vanuit dit scherm kunnen beoordelaars:
  - Het examenboekje downloaden
  - De examenopdrachten beluisteren
  - De te beoordelen kandidaatreacties selecteren
– De beoordelaarsschermen zijn aangepast aan het correctievoorschrift van de Staatsexamens
  - Na selectie van een kandidaatreactie in het introductiescherm verschijnt het beoordelingsscherm. In dit scherm kunnen de beoordelaars:
    - De kandidaatuiting afluisteren (door te klikken op ‘bijlage bij dit concept’)
    - De kandidaat beoordelen volgens het beoordelingsvoorschrift. (zie bijlage 3 voor een voorbeeld van een beoordelingsscherm voor een opdracht uit deel 1, korte opdrachten en bijlage 4 voor een voorbeeld van een opdracht uit deel 3, Lange opdrachten)

6 De beoordelaarstaak

De beoordelaarstaak van een beoordelaar in deze gesegmenteerde online beoordeling was – afgezien van het medium en het feit dat men geen gehele examens meer beoordeelde – gelijk aan de beoordelaarstaak in een regulier examen. De beoordelaar beoordeelde elk spreekfragment aan de hand van een beoordelingsscherm dat letterlijk de tekst bevatte van het correctievoorschrift van het reguliere examen.

7 Resultaten: beoordelaarseffecten en de relatie tussen scores in de pilot en de oorspronkelijke examenscores

In de huidige opzet van de STEX beoordeelt dezelfde beoordelaar alle uitingen van een kandidaat. Deze opzet is gekozen om logistieke redenen maar brengt het risico met zich mee dat uitingen bij verschillende opdrachten niet onafhankelijk worden beoordeeld. Dit staat in de literatuur bekend als het *halo-effect*.

Een halo-effect kan optreden om verschillende redenen. Een beoordelaar kan zich op basis van de eerste uitingen een indruk vormen van de vaardigheid van een kandidaat om op basis van die indruk de overige beoordelingen te doen zonder nog verder te luisteren naar wat de kandidaat zegt. Beoordelaars kunnen, bijvoorbeeld, een voorkeur hebben voor bepaalde accenten of hun oordelen laten bepalen door de behoefte (bewust of onbewust) om consistent te zijn.

Halo-effecten zijn niet noodzakelijk nadelig voor een kandidaat. Het betekend echter altijd dat de eigenschappen van welbepaalde beoordelaars invloed hebben op de uitslag. Daarmee is het een bedreiging voor de validiteit van het examen.

Uit onderzoek blijkt dat halo-effecten optreden ook al zijn beoordelaars door training bewust gemaakt van de gevaren. Het gesegmenteerd beoordelen sluit halo-effecten uit en is daarmee effectiever dan training, simpelweg doordat verschillende uitingen door verschillende beoordelaars worden beoordeeld.

Hoewel de huidige studie vooral is opgezet om “proef te draaien” met de logistiek van de nieuwe toewijzing van beoordelaars willen we de gelegenheid niet voorbij laten gaan om de hier verzamelde oordelen (zonder halo-effecten) te vergelijken met de oordelen die zijn gegeven bij de oorspronkelijke examen. Hierover wordt verslag gedaan in het vervolg van deze sectie.

In paragraaf 7.1 onderzoeken we de aanwezigheid en gevolgen van halo-effecten. In de paragrafen 7.2 doen we verslag van de overeenstemming tussen beoordelaars. De samenhang tussen de reguliere examenscores en de scores in de gesegmenteerde beoordeling worden in paragraaf 7.3 besproken.

7.1 Beoordelaarseffecten

Halo-effecten
Om halo-effecten zichtbaar te maken gebruiken we data van het normexamen van Spreken II (ochtend maart 2006). Hiertoe kijken we binnen en tussen beoordelaars naar de relatie tussen de scores van kandidaten op de drie onderdelen van het examen:

- Korte opdrachten
- Middellange opdrachten
- Lang opdracht

Wanneer er geen sprake is van een halo-effect maakt het niet uit of we dezelfde beoordelaar vragen om twee afzonderlijke delen van het examen te beoordelen of dat verschillende beoordelaars afzonderlijke delen beoordelen. Als er een halo-effect optreedt, zijn de correlaties tussen de opeenvolgende examenonderdelen bij dezelfde beoordelaar hoger dan de correlaties tussen opeenvolgende delen die beoordeeld zijn door verschillende beoordelaars.

Om te kunnen beoordelen of een halo-effect optreedt, is daarom de correlatie berekend tussen beoordelaars en binnen beoordelaars bij de beoordeling van de korte, middellange en lange opdrachten.

Tabel 1 Correlaties tussen examenonderdelen binnen beoordelaars en tussen beoordelaars op het Normexamen Spreken II volgens de reguliere beoordelingswijze (2 beoordelaars die het hele examen beoordelen)

		R1			R2		
		K	ML	L	K	ML	L
R1	K	1.000					
	ML	0.630	1.000				
	L	0.624	0.823	1.000			
R2	K	**0.622**	0.574	0.548	1.000		
	ML	0.534	**0.756**	0.662	0.667	1.000	
	L	0.549	0.666	**0.710**	0.612	0.795	1.000

Deze correlatietabel bevat twee verschillende soorten correlaties: *Binnen-Beoordelaar (BB) correlaties* en *Tussen-Beoordelaar (TB)correlaties*. De BB-correlaties zijn correlaties tussen oordelen van dezelfde beoordelaar. In de linker bovenhoek staan de BB-correlaties van de eerste beoordelaar (R1). In de rechter benedenhoek staan BB-correlaties van de tweede beoordelaar (R2). TB-correlaties zijn correlaties tussen oordelen van de eerste en de tweede beoordelaar. Deze staan in de linker benedenhoek.

In de correlatietabel is het volgende te zien:
1. De BB-correlaties voor de eerste en de tweede beoordelaar (gearceerde cellen) lijken erg op elkaar. De reden hiervoor is dat eerste en tweede beoordelaar gekozen zijn uit dezelfde verzameling getrainde beoordelaars.
2. De correlaties tussen middellange en lange opgave zijn zowel bij BB als bij TB groter dan die tussen korte en middellange opgaven of korte en lange opgaven.
3. De BB-correlaties (gearceerde cellen) zijn altijd hoger dan de overeenkomstige TB-correlaties (niet-gearceerde cellen).

Dat laatste betekent dat oordelen van dezelfde beoordelaar over opeenvolgende toetsonderdelen samenhangen sterker dan die van verschillende beoordelaars. Dit is een aanwijzing voor het optreden van halo-effecten.
We hebben dezelfde correlaties berekend op de data van de pilot. Zoals reeds eerder vermeld treedt bij deze manier van beoordelen het halo-effect niet op, omdat een beoordelaar bij het beoordelen van een uiting geen voorkennis van de kandidaat heeft op grond van eerdere uitingen. Dit moet tot uiting komen in het feit dat de BB- en TB-correlaties niet langer verschillen.

Tabel 2 Correlaties tussen examenonderdelen binnen beoordelaars en tussen beoordelaars op het juli-examen Spreken II 2006 (gesegmenteerde beoordeling)

		R1			R2		
		K	ML	L	K	ML	L
R1	K	1					
	ML	0,598	1				
	L	0,565	0,656	1			
R2	K	**0,814**	0,745	0,577	1		
	ML	0,602	**0,915**	0,636	0,678	1	
	L	0,489	0,585	**0,621**	0,555	0,581	1

Uit tabel 2 blijkt dat deze verschillen inderdaad verdwenen zijn.
Bij gesegmenteerd beoordelen is de BB-correlatie niet langer altijd hoger dan de TB-correlatie. In de helft van de gevallen is de BB- correlatie hoger dan de TB-correlatie, in de andere helft is hij lager. Dat niet alle verschillen nul zijn, is het gevolg van steekproefvariaties in een steekproef van 50 kandidaten.
Het gemiddeld verschil tussen de BB-correlaties en de TB-correlaties bij het gesegmenteerd beoordelen is volgens verwachting nul: -0,001. In de reguliere beoordeling zijn deze verschillen hoger: 0,104 voor beoordelaars 1 en 0,103 voor beoordelaars 2. In bijlage 5 is een tabel opgenomen waarin een overzicht wordt gegeven van de verschillen tussen de BB-correlaties en de TB-correlaties bij de reguliere beoordeling en de gesegmenteerde beoordeling.

Te zien is dat ML en L niet langer de hoogste correlatie vertonen.
De correlatietabellen en de tabel met verschillen tussen de BB en TB-correlaties tonen aan dat bij het gesegmenteerd beoordelen het halo-effect dat wordt gevonden bij de reguliere beoordeling verdwenen is.

Overige beoordelaarseffecten
In bijlage 1 worden nog andere beoordelaarseffecten beschreven. Gesegmenteerd beoordelen heft deze weliswaar niet op, maar het is duidelijk dat door een random toewijzing en gesegmenteerde beoordeling deze effecten verdeeld worden over een kandidaat. In het oude systeem heeft bijvoorbeeld een strenge beoordelaar een sterk effect op de score van een kandidaat. In het gesegmenteerd beoordelen wordt dit effect teniet gedaan, omdat elke kandidaat door meerdere beoordelaars beoordeeld wordt.

7.2 Beoordelaarsovereenstemming

De beoordelaarsovereenstemming is in de pilot hetzelfde als bij de reguliere beoordeling. Ook wanneer we kijken naar de beoordelaarsovereenstemming per aspect, is er grote overeenkomst tussen de examen data en de pilot. Aspecten die slecht (goed) te beoordelen waren bij de reguliere manier van beoordelen waren dit ook bij het gesegmenteerd beoordelen.

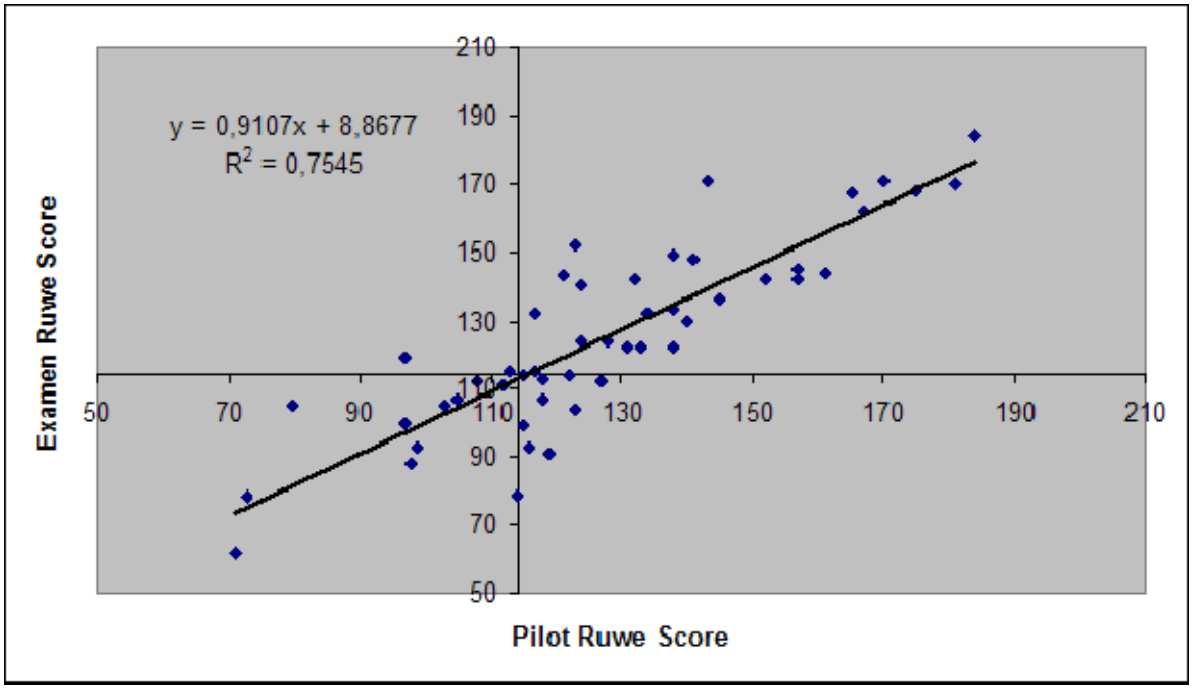
Tabel 3 Gemiddelde beoordelaarsovereenstemming en overeenstemming per aspect bij segment beoordeling en reguliere beoordeling

		Gesegmenteerd beoordelen		Reguliere beoordeling	
	Aantal	Gemiddeld % overeenstemming	Gem. kappa	Gemiddeld % overeenstemming	Gem. kappa
IH	14	0,69	0,37	0,70	0,33
WZ	11	0,59	0,29	0,61	0,33
WS	6	0,61	0,33	0,57	0,27
US	4	0,58	0,28	0,52	0,17
TP	3	0,89	0,29	0,91	0,22
CH	1	0,50	0,23	0,49	0,19
Gemiddeld		0,65	0,32	0,65	0,30

7.3 Relatie tussen scores in de pilot en de oorspronkelijke examenscores

De score die kandidaten kregen op het oorspronkelijke examen en na herbeoordeling bij de pilot zijn sterk gerelateerd. De correlatie bleek gelijk te zijn aan 0,87. Een spreidingsplot is te zien in Figuur 1.
De gemiddelde scores in de pilot en het reguliere examen verschilden nauwelijks en bedroegen respectievelijk 127,58 en 125,6. Van de 50 kandidaten slaagden er 8 kandidaten volgens de pilot, terwijl ze gezakt waren voor het examen. Bij 2 kandidaten was het omgekeerde het geval.

Figuur 1 Samenhang tussen de scores op het reguliere examen en de pilotscores



7.4 Conclusies

Op grond van de analyse van de data van de pilot kan worden geconcludeerd dat gesegmenteerd beoordelen het halo-effect tenietdoet.
Het gesegmenteerd beoordelen heeft geen effect op de beoordelaarsovereenstemming.
Tussen beide manieren van beoordelen bestaat een hoge samenhang. Deze bevindingen geven geen aanleiding van gesegmenteerde beoordeling af te zien.

8 Resultaten: de werking van het online beoordelingssysteem

Het online beoordelingssysteem is een aanpassing van een reeds bestaand systeem. Dit houdt in dat de vormgeving en het gebruiksgemak van het programma nog niet optimaal is voor toepassing bij het Staatsexamen NT2. Hiervan hebben enkele beoordelaars melding gemaakt. Het systeem kon echter wel de noodzakelijke functies voor het beoordelen uitvoeren (selecteren van kandidaten, afluisteren van spreekuitingen, invullen van het correctievoorschrift) uitvoeren. De meeste beoordelaars konden hun taak dan ook naar behoren uitvoeren. Twee beoordelaars hebben hun taak afgebroken vanwege problemen met het beoordelingssysteem. Zij zijn vervangen door twee nieuwe beoordelaars.

De pilot heeft duidelijk gemaakt dat gesegmenteerd online beoordelen mogelijk is.
Hoewel het duidelijk moge zijn dat het in de pilot gebruikte systeem niet als definitief bedoeld is kunnen deze gebruikservaringen nuttig zijn bij het ontwerp van een definitief systeem.

9 Samenvatting en aanbevelingen

In deze pilot is onderzocht of een beoordelingssysteem waarin online segmentbeoordeling met random toekenning van beoordelaars in de praktijk uitvoerbaar zou zijn.
Zoals in de inleiding is beschreven is deze wijze van beoordelen aanbevolen in eerdere onderzoeksrapporten (Bechger en Maris, 2004, 2006).

Een dergelijke beoordeling heeft de volgende voordelen.
- Verbetering van de feedback aan beoordelaars
- Verbetering van de psychometrische onderbouwing van Spreken en Schrijven
- Het doen verdwijnen of neutraliseren van ongewenste beoordelaarseffecten

Om de praktische uitvoerbaarheid te testen is een bestaand online systeem aangepast en een programma voor random toewijzing geschreven. Tien beoordelaars hebben de prestaties van 50 Staatsexamen-kandidaten Programma II met bekende scores opnieuw gesegmenteerd beoordeeld met gebruikmaking van dit online systeem.

De pilot heeft uitgewezen dat deze gesegmenteerde online beoordeling met random toewijzing praktisch realiseerbaar is.

Uit de analyse van de data in deze pilot is gebleken dat de scores verkregen door het gesegmenteerd beoordelen zeer sterk samenhangen met de oorspronkelijke scores van de kandidaten in deze steekproef. De gemiddelde scores van de segmentbeoordeling weken nauwelijks af van de scores volgens de reguliere beoordeling. De segmentbeoordeling leidde tot een hoger percentage geslaagde kandidaten.
De beoordelaarsovereenstemming was in beide beoordelingswijze gelijk. De data wezen tevens uit dat halo-effecten waren verdwenen.

Aanbevelingen
1. Op grond van deze pilot is het vanuit psychometrisch oogpunt gezien aan te bevelen een verdere uitwerking te geven aan segmentbeoordeling bij Spreken. Vanuit logistiek oogpunt gezien moet het traject nog worden uitgetest.
2. Invoering van deze beoordelingssytematiek bij Schrijven is eveneens aan te bevelen. Voorwaarde hiervoor is wel dat ook deze beoordeling gedigitaliseerd kan worden. Dit wordt in 2007 nader onderzocht
3. Indien overgegaan wordt op gesegmenteerde digitale beoordeling, dient hiervoor een systeem ontworpen te worden dat aangepast is aan de eisen van de Staatsexamens. Het in de pilot toegepaste systeem is weliswaar uitvoerbaar, maar niet voldoende gebruiksvriendelijk.
4. Daarnaast is het aan te bevelen om verdere maatregelen te nemen ter verhoging van de beoordelaarsovereenstemming van de beoordeling van Spreken en Schrijven door verduidelijking en – indien nodig – een verdere dichotomisering van de beoordelingsvoorschriften. Verdere dichotomisering wordt in 2007 nader onderzocht.

Referenties
Bechger, T.M., Maris, G. (2006). *Het scoren van open vragen: het beoordelen van beoordelaars.* POK memorandum 2006-2. Cito, Arnhem
Maris G., Bechger, T. M. (2004). *Het scoren van open vragen: Theorie en Praktijk.* POK memorandum 2004-2. Cito, Arnhem

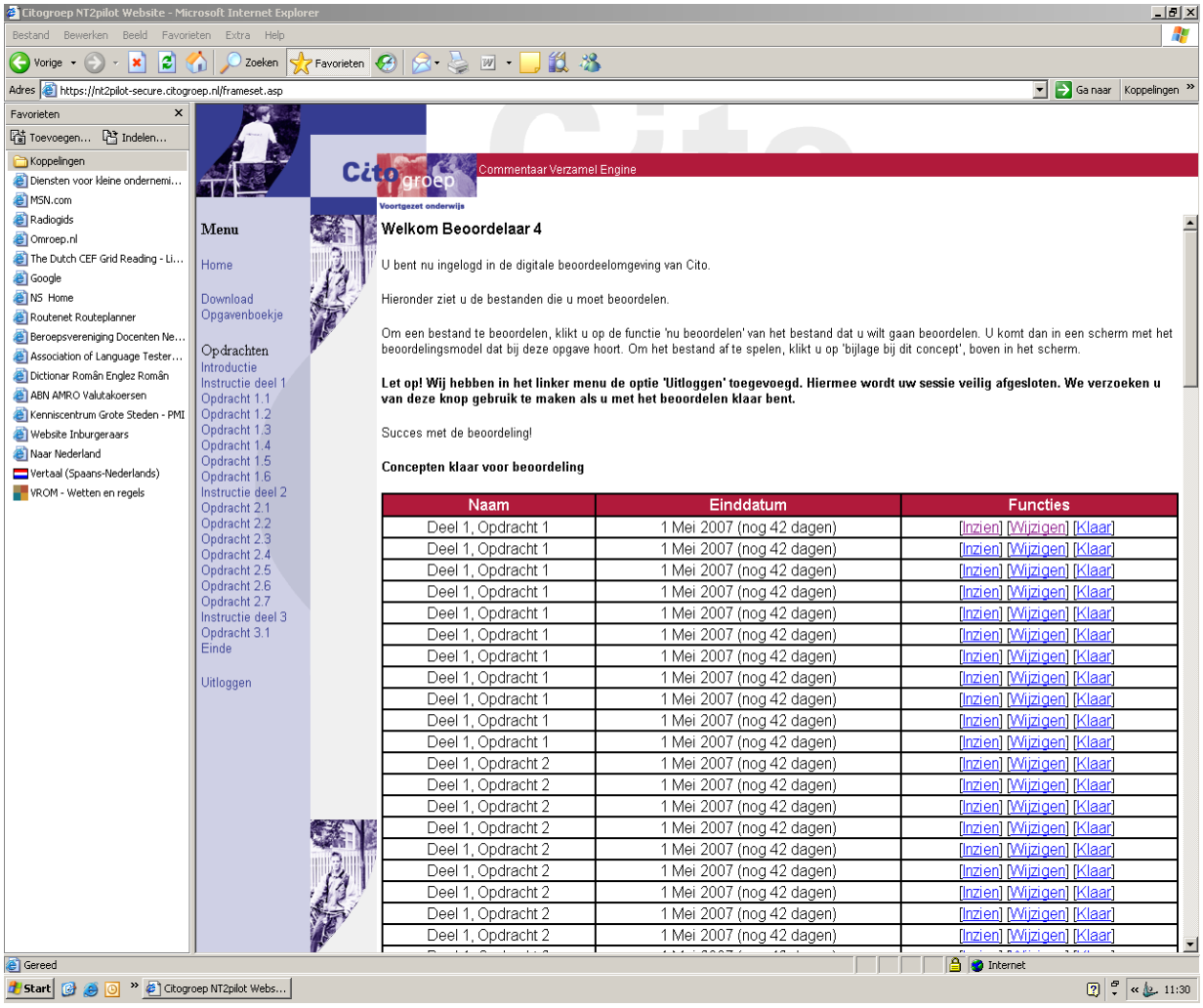
Bijlagen

Bijlage 1 Beoordelaarseffecten

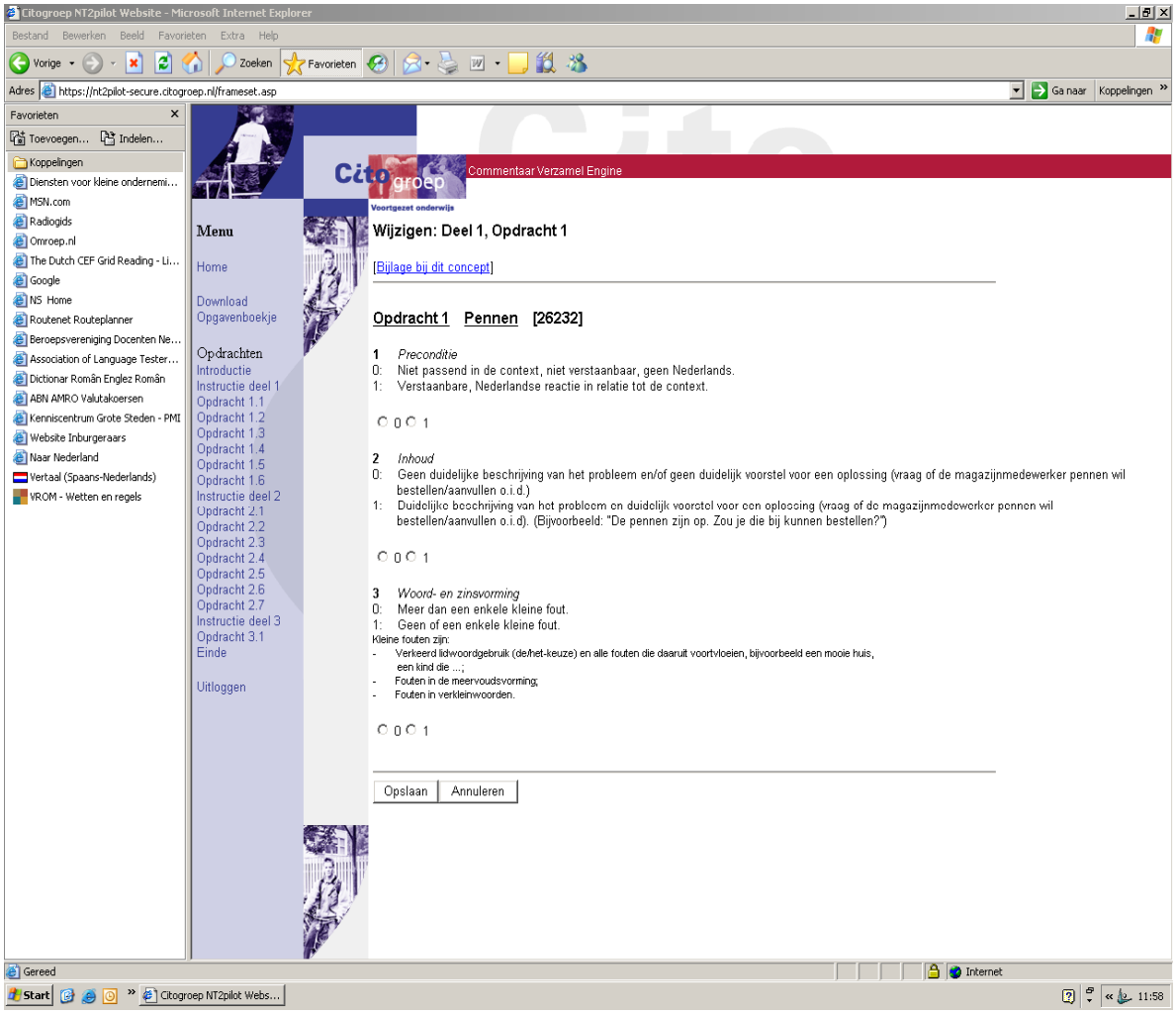
Beoordelaars staan al dan niet bewust bloot aan invloeden die een eerlijke beoordeling van uitingen in de weg staan. Enkele beoordelaarseffecten zijn:

- contaminatie-effect
  De beoordelaar kent aan de beoordeling van het werk nog andere dan de bedoelde functie(s) toe.
- halo-effect
  De storende 'uitstraling' van niet ter beoordeling staande kwaliteiten van een kandidaat op de beoordeling van een geleverde prestatie.
- normverschuiving
  De neiging van een beoordelaar zich in de strengheid van zijn beoordelingen aan te passen aan het gemiddelde prestatieniveau van een groep kandidaten.
  Onder het begrip normverschuiving wordt ook wel verstaan het beperkt gebruik maken van de beoordelingsschaal; sommige beoordelaars geven overwegend hoge punten, andere lage en weer andere neigen naar het midden van een schaal.
- sequentie-effect
  De nawerking van voorafgaande beoordelingen van het werk van andere leerlingen bij het beoordelen van een leerlingprestatie. Een beoordelaar zal bijvoorbeeld na enkele zwakke prestaties de neiging hebben om aan de volgende redelijk goede prestatie een relatief hogere waardering te geven, dan wanneer dezelfde prestatie zou volgen op enkele zeer goede.
- signifisch effect
  De effecten die optreden indien de beoordelingstaak verschillend wordt opgevat door twee of meer beoordelaars.

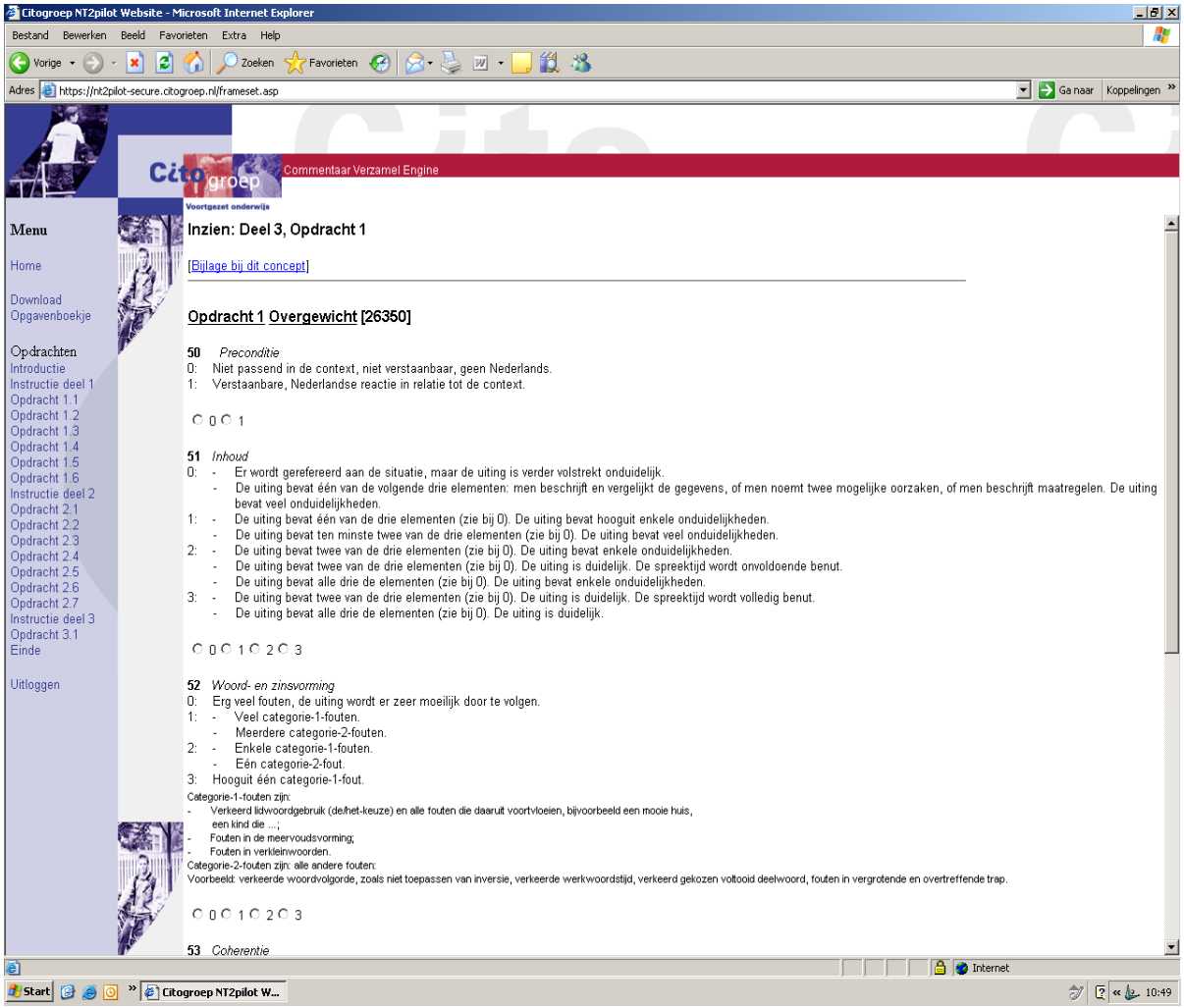
Bijlage 2 Het introductiescherm van het online beoordelingssysteem



Bijlage 3 Het beoordelingsscherm voor opdrachten deel 1



Bijlage 4 Het beoordelingsscherm voor opdrachten deel 3



Bijlage 5 Verschillen binnenbeoordelaarscorrelaties (BB) en tussenbeoordelaarscorrelaties (TB) tussen de examenonderdelen

Reguliere beoordeling			
Boordelaars 1	BB-correlatie	TB-correlatie	BB-TB
kort middellang a	0,63	0,534	0,096
kort middellang b	0,63	0,574	0,056
kort lang a	0,624	0,549	0,075
kort lang b	0,624	0,548	0,076
middellang lang a	0,823	0,666	0,157
middellang lang b	0,823	0,662	0,161
gemiddeld verschil BB-TB			0,104
Beoordelaars 2	BB-correlatie	TB-correlatie	BB-TB
kort middellang a	0,667	0,534	0,133
kort middellang b	0,667	0,574	0,093
kort lang a	0,612	0,549	0,063
kort lang b	0,612	0,548	0,064
middellang lang a	0,795	0,666	0,129
middellang lang b	0,795	0,662	0,133
gemiddeld verschil BB-TB			0,103
totaalgemiddelde verschillen			0,103
Gesegmenteerde beoordeling			
Boordelaars 1	BB-correlatie	TB-correlatie	BB-TB
kort middellang a	0,598	0,602	-0,004
kort middellang b	0,598	0,745	-0,147
kort lang a	0,565	0,489	0,076
kort lang b	0,565	0,577	-0,012
middellang lang a	0,656	0,585	0,071
middellang lang b	0,656	0,636	0,02
gemiddeld verschil BB-TB			0,001
Boordelaars 2	BB-correlatie	TB-correlatie	BB-TB
kort middellang a	0,678	0,602	0,076
kort middellang b	0,678	0,745	-0,067
kort lang a	0,555	0,489	0,066
kort lang b	0,555	0,577	-0,022
middellang lang a	0,581	0,585	-0,004
middellang lang b	0,581	0,636	-0,055
gemiddeld verschil BB-TB			-0,001
totaalgemiddelde verschillen			-0,001