

Het scoren van open vragen: het beoordelen van beoordelaars

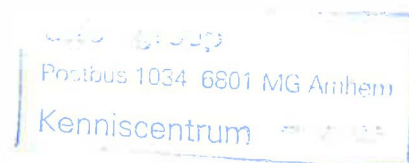
Timo M. Bechger
Gunter Maris

**Het scoren van open vragen:
het beoordelen van beoordelaars**

Timo M. Bechger

Gunter Maris

Cito
Arnhem, november 2006



8501 007 0607



Uit deze uitgave mogen zonder toestemming van de auteur geen aanhalingen worden overgenomen.

Inhoudsopgave

Aanleiding	3
Het geval van beoordelaar 749	4
De toewijzing van kandidaten aan beoordelaars	6
Welke informatie wordt er verzameld?	9
De geschiedenis van beoordelaars	10
De relatie tussen het nummer en de presentaties van kandidaten.....	13
Hoe komt het oordeel in de beobrieven tot stand?	14
Het effect van de beobrief	15
Aanbevelingen	17
Conclusie: Veroordelen of stimuleren?	22
Referenties	23
Bijlage:.....	24

Aanleiding

Bij het *Staatsexamen voor Nederlands als tweede taal*, worden voor het meten van de productieve vaardigheden (schrijven en spreken) menselijke beoordelaars ingezet:

- Iedere kandidaat wordt beoordeeld door twee beoordelaars. De oordelen worden gecombineerd bij het bepalen van het cijfer (zie Maris en Bechger, 2004).
- Er wordt onderzoek naar beoordelaarovereenstemming verricht bij het pretesten van nieuwe opgaven om zo slecht beoordeelbare opgaven te kunnen verwijderen, en inzicht te krijgen in het soort opgaven dat tot lage overeenstemming aanleiding geeft.
- Er wordt informatie verzameld over individuele beoordelaars.

Dit memorandum heeft betrekking op het derde punt. Het Staatsexamen dient als casus.

Informatie over individuele beoordelaars wordt verzameld om drie redenen: Ten eerste wordt ze gebruikt om tegemoet te komen aan de wens van beoordelaars om informatie te zien over de eigen oordelen in vergelijking met die van anderen. Iedere beoordelaar krijgt na een examen een brief (de “de beobrief”) met daarin informatie over zijn functioneren. De tweede reden is dat men slecht functionerende beoordelaars wil kunnen identificeren. Deze beoordelaars zouden dan kunnen worden uitgenodigd, bijvoorbeeld voor hertraining, of in het uiterste geval zelfs kunnen worden ontslagen. Een derde reden is de veronderstelling dat de beobrief een positief effect zou kunnen hebben op de motivatie van beoordelaars en zo op de kwaliteit van de oordelen. Deze veronderstelling kwam voort uit de observatie dat beoordelaars in het kader van een onderzoek tot grotere overeenstemming kwamen dan gebruikelijk tijdens examens (zie Lam, 2003).

De beobrief bevat, naast informatie, ook een *oordeel* over het gedrag van iedere beoordelaar. De vraag die we hier aan de orde zullen stellen is of dit oordeel gefundeerd is, en hoe het beter kan. We zullen aantonen dat de systematische wijze waarop kandidaten worden toegewezen aan beoordelaars de onderbouwing van het oordeel over een beoordelaar bemoeilijkt. Daarnaast bespreken we een klein onderzoek naar het effect van de beobrieven op de kwaliteit van de examens. We eindigen met een lijst van werkzaamheden en onderwerpen voor toekomstig onderzoek.

De belangrijkste conclusie is dat het, voor het *beoordelen van beoordelaars*, noodzakelijk is om beoordelaars op basis van toeval toe te wijzen aan kandidaten. Op dit moment geschiedt de toewijzing van beoordelaars aan kandidaten volgens een logistiek handige systematiek.

We beschouwen dit rapport als een vervolg op een eerder rapport (Maris en Bechger, 2004) over het verzamelen en schalen van de oordelen.¹ Net als bij het eerdere rapport dient het staatsexamen als concreet voorbeeld maar hebben de argumenten betrekking op een grotere klasse van soortgelijke examens.

Ter inleiding bespreken we het geval van de onfortuinlijke beoordelaar 749.

¹ Een bewerkte, Engelstalige versie zal verschijnen als een hoofdstuk in “The Handbook of Statistics, 27”. Zie Maris en Bechger (2006).

Het geval van beoordelaar 749

In juli 2005 werd beoordelaar 749 voor de eerste keer ingezet bij het Schrijfexamen van juli 2005. Na afloop kreeg ze een brief met daarin de vermelding dat zij een strenge beoordelaar was geweest (zie Bijlage). Beoordelaar 749 deed voor het eerst mee. We hebben daarom niet kunnen vaststellen hoe beoordelaar 749 heeft gefunctioneerd in eerdere examens. Toch werd er in de brief een negatief oordeel uitgesproken.

De brief liet er weinig twijfel over bestaan. Zo vermeldt de brief dat slechts 14% van de kandidaten van beoordelaar 749 zijn geslaagd, terwijl 56% van alle kandidaten is geslaagd voor het examen. Laten we beoordelaar 749 toch nog eens nader beschouwen en ons afvragen of het oordeel terecht was.

Beoordelaar 749 heeft 21 kandidaten beoordeeld. Dezelfde kandidaten zijn een tweede keer beoordeeld door één of meer *contrabeoordelaars*. We zullen in het vervolg spreken van *de* contrabeoordelaar. De scores die zijn toegekend door beoordelaar 749 en die toegekend door de contrabeoordelaar van dezelfde kandidaten staan in onderstaande tabel.

Tabel 1. Kandidaten die beoordeeld zijn door beoordelaars 749 en 445 en de cijfers die horen bij de score die elke beoordelaar toekende aan ieder van de kandidaten.

Kandidaat	scoreBeo1	ScoreBeo2	Beo1	Beo2
942595	503,66	482,36	436	749
942598	473,93	473,93	436	749
942610	469,69	473,93	445	749
942620	473,93	478,13	445	749
942627	499,36	490,79	445	749
942631	478,13	469,69	445	749
942632	452,53	490,79	445	749
942634	473,93	495,06	445	749
942636	499,36	486,56	445	749
942608	495,06	517,02	445	749
942605	541,29	531,17	445	749
942637	478,13	490,79	445	749
942604	478,13	473,93	445	749
942638	443,6	465,46	445	749
942623	461,19	469,69	445	749
942617	456,89	434,27	445	749
942614	419,03	429,38	445	749
942607	490,79	495,06	445	749
942619	473,93	486,56	445	749
942635	536,13	558,13	445	749
942615	452,53	448,1	445	749
942645	443,6	469,69	445	750
942642	564,44	587,5	445	750

Uit Tabel 1 blijkt dat beoordelaar 749 vrijwel altijd gekoppeld is geweest aan beoordelaar 445. Er waren slechts twee kandidaten waar de contrabeoordelaar een andere was. In het vervolg van dit stuk wordt betoogd dat dit niet toevallig is maar een typisch gevolg van de systematiek van toewijzing van kandidaten aan beoordelaars.

Vergelijken we beide beoordelaars dan krijgen we de volgende tabel. Op de diagonaal staan de aantallen kandidaten die bij beide geslaagd dan wel gezakt zijn.

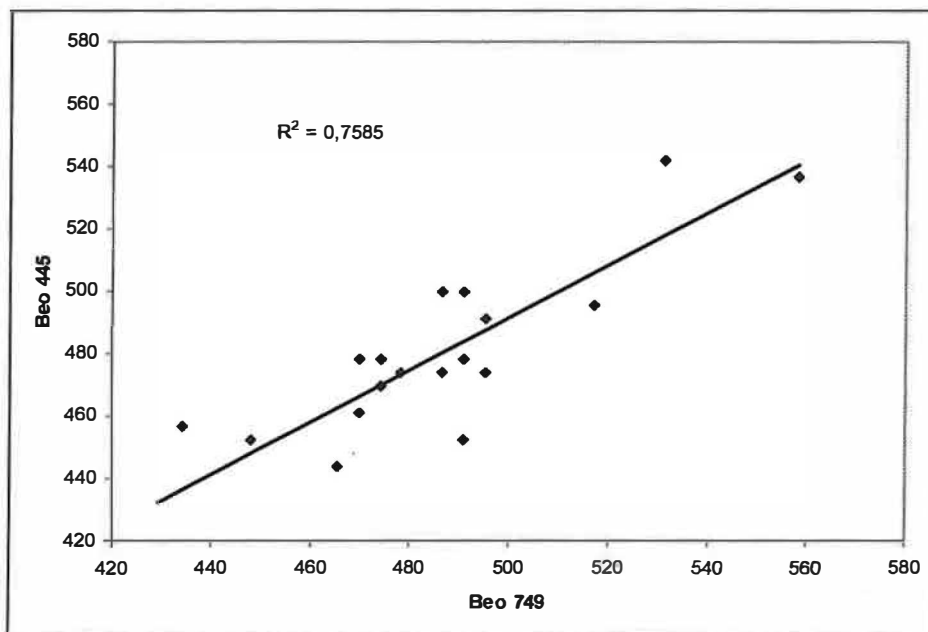
Tabel 2. Tabel met aantallen geslaagd/gezakt voor beoordelaars 749 en 445 (Schijven programma II, juli 2005).

		Beo 749		
		Geslaagd	gezakt	
Beo 445	Geslaagd	2	0	2
	Gezakt	1	16	17
		3	16	19

We zien dat beide beoordelaars het vrijwel altijd eens zijn (Cohen's kappa is gelijk aan 0.97). We zien echter ook dat de meerderheid van de kandidaten bij beide beoordelaars gezakt zijn. Inderdaad kreeg ook beoordelaar 445 een brief met daarin de boodschap dat hij of zij te streng was geweest (zie Bijlage).

We concluderen dat we wat voorbarig waren in ons het oordeel. We kunnen namelijk niet bepalen of beide beoordelaars zijn toegewezen aan relatief slechte kandidaten of dat *beide* te streng waren ten opzichte van een doorsnee beoordelaar.

Figuur 1. Een plot van scores van kandidaten op basis van beoordelingen door beoordelaar nummer 445 en beoordelaar 749 (gemaakt met de data in Tabel 2). De zak-slaag grens ligt door constructie bij een score van 500.



In de volgende paragrafen wordt duidelijk dat het geval van beoordelaar 749 niet op zichzelf staat. We zullen betogen dat de systematische toewijzing van kandidaten aan beoordelaars hier de oorzaak van is.

De toewijzing van kandidaten aan beoordelaars

Iedere kandidaat krijgt twee beoordelaars toegewezen. De toewijzing gebeurt volgens een logistiek handige systematiek:

- Men begint kandidaten toe te wijzen aan de beoordelaar met het laagste nummer. Daarna worden kandidaten toegewezen aan de tweede beoordelaar, de derde, etc.
- Wanneer alle kandidaten zijn toegewezen begint men weer bij de eerste kandidaat en wordt een contrabeoordelaar toegewezen.

Dit levert in alle gevallen data op zoals schematisch in Tabel 3 weergegeven.

Tabel 3. Een overzicht van de beoordelaars toegewezen aan kandidaten. Het betreft hier echte examendata. Waar een XX staat zijn beoordelingen verzameld. Waar een 0 staat waren er geen data.

	14	24	44	51	52	71	...	382	332	337	352	360
930001	XX	0	0	0	0	0	...	0	XX	0	0	0
930002	XX	0	0	0	0	0	...	0	XX	0	0	0
930004	XX	0	0	0	0	0	...	0	XX	0	0	0
930005	XX	0	0	0	0	0	...	0	XX	0	0	0
930006	XX	0	0	0	0	0	...	0	XX	0	0	0
930009	XX	0	0	0	0	0	...	0	XX	0	0	0
930010	XX	0	0	0	0	0	...	0	XX	0	0	0
930012	XX	0	0	0	0	0	...	0	XX	0	0	0
	*	*	*	*	*	*	...	*	*	*	*	*
	*	*	*	*	*	*	...	*	*	*	*	*
	*	*	*	*	*	*	...	*	*	*	*	*
930029	XX	0	0	0	0	0	...	0	XX	0	0	0
930030	XX	0	0	0	0	0	...	0	XX	0	0	0
930032	XX	0	0	0	0	0	...	0	XX	0	0	0
930035	XX	0	0	0	0	0	...	0	XX	0	0	0
930036	XX	0	0	0	0	0	...	0	XX	0	0	0
930037	0	XX	0	0	0	0	...	0	XX	0	0	0
930038	0	XX	0	0	0	0	...	0	XX	0	0	0
930039	0	XX	0	0	0	0	...	0	XX	0	0	0
930041	0	XX	0	0	0	0	...	0	XX	0	0	0
930042	0	XX	0	0	0	0	...	0	XX	0	0	0
	*											
	*											
	*											
930195	0	0	0	0	XX	0	...	0	0	0	0	XX
930196	0	0	0	0	XX	0	...	0	0	0	0	XX
930198	0	0	0	0	XX	0	...	0	0	0	0	XX
930199	0	0	0	0	0	XX	...	0	0	0	0	XX
930200	0	0	0	0	0	XX	...	0	0	0	0	XX
930201	0	0	0	0	0	XX	...	0	0	0	0	XX
930202	0	0	0	0	0	XX	...	0	0	0	0	XX

Tabel 3 laat duidelijk zien hoe systematisch de toewijzing is. Wat we zien is dat:

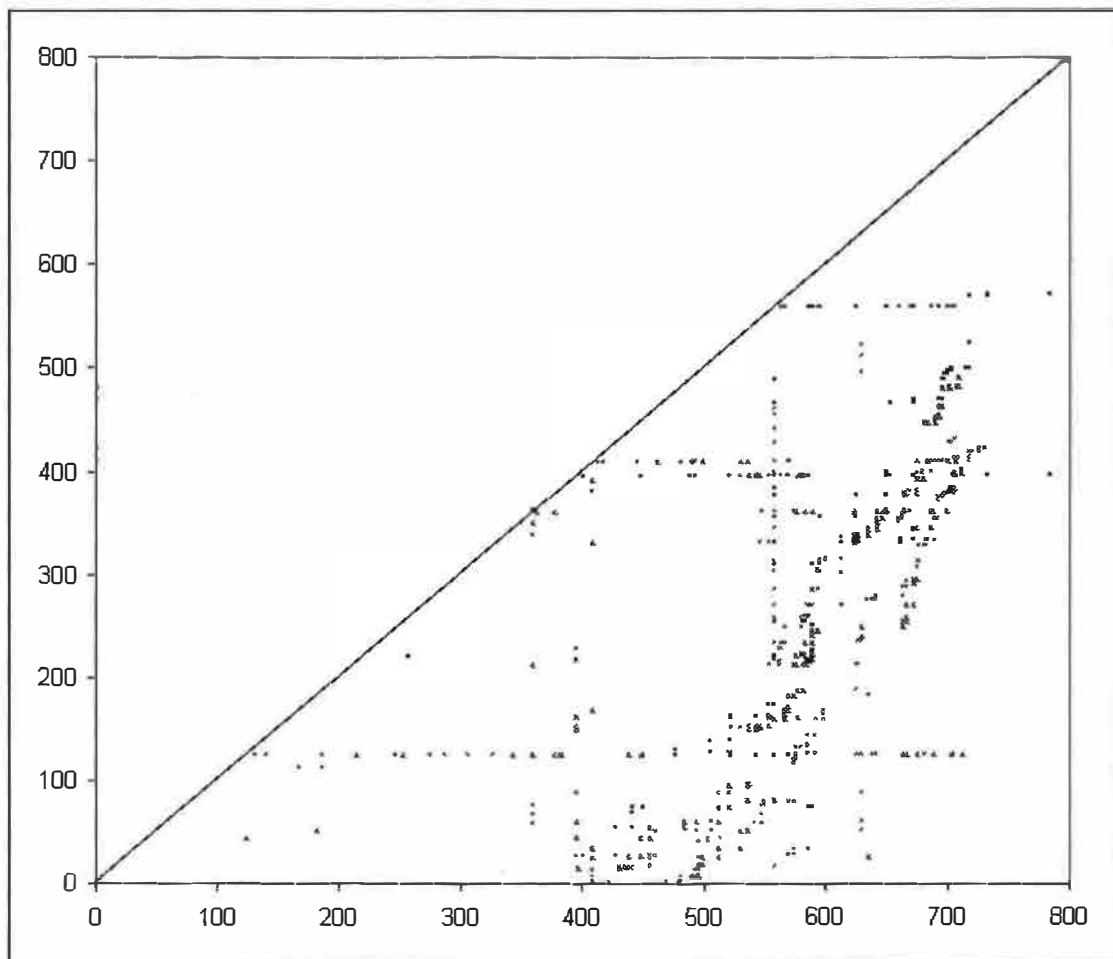
- Hoe zeer de systematische toewijzing er toe leidt dat een beoordelaar voornamelijk gekoppeld wordt aan één welbepaalde contrabeoordelaar blijkt ook uit onderstaande tabel gebaseerd op dezelfde data als de voorgaande tabel.

30 11
 25 1
 23 5
 27
 3 25
 3 3 6
 21 6
 21 6
 52 5
 23 6
 19 14
 21 27 6
 22 6
 29
 23 6
 29
 24 3

De beoordelaarnummers worden in oplopende volgorde toegekend wanneer een beoordelaar in het bestand wordt opgenomen. Dat wil zeggen dat beoordelaars met lagere nummers langer in het bestand zitten dan beoordelaars met hoge nummers. De systematiek in het toewijzen van kandidaten aan beoordelaars leidt er dus toe dat ervaren beoordelaars worden gekoppeld aan minder ervaren beoordelaars.

Deze systematiek wordt toegepast op alle examens. Hierdoor is het mogelijk dat een beoordelaar ook *over examens heen* vaak met dezelfde andere beoordelaar voorkomt. Figuur 2 illustreert dat dit inderdaad het geval is.

Figuur 2. Beoordelaarcombinaties die voorkwamen in drie examens in 2005. De examens zijn onderscheiden door kleuren (of grijs tinten). Op de assen staat het nummer van de beoordelaars.



Elk punt in Figuur 2 correspondeert met een beoordelaarpaar. Deze figuur is gebaseerd op drie opeenvolgende Spreken II examens van 2005 en laat zien dat een beoordelaar hoofdzakelijk gekoppeld wordt aan andere beoordelaars waarvan de nummers kort bij elkaar liggen. Onder bepaalde omstandigheden wordt een *derde oordeel* bepaald. In Figuur 2 zijn de derde beoordelingen herkenbaar als horizontale en verticale lijnen. Dat komt doordat deze extra beoordelingen vaak door dezelfde (als expert beschouwde) beoordelaars worden gedaan.

Tenslotte merken we op dat, eveneens om logistieke redenen, iedere beoordelaar steeds alle uitingen van een kandidaat beoordeelt. Dit brengt het risico met zich mee dat een beoordelaar zich, op basis van een eerste indruk, een impressie vormt van een kandidaat die alle oordelen beïnvloedt. Dit wordt het *halo effect* genoemd. De consequentie is dat verschillende aspecten (inhoud, spelling, etc.) *niet* onafhankelijk van elkaar worden beoordeeld zoals verondersteld bij de interpretatie en de analyse van de data (zie Maris en Bechger, 2004).

Welke informatie wordt er verzameld?

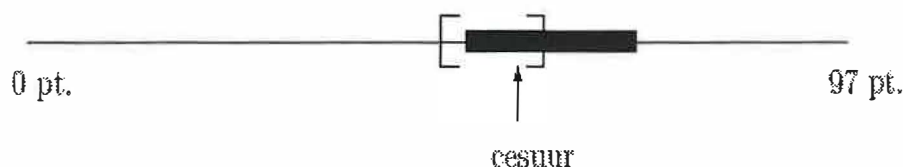
Op dit moment worden, na ieder examen, de volgende gegevens bepaald voor ieder van de betrokken beoordelaars:

1. *Mildheid*: Het verschil tussen de gemiddelde score van een beoordelaar en het gemiddelde berekend over alle beoordelaars.
2. *Scorebereik*: Het interval waarbinnen 75% van de scores van de beoordeelde kandidaten viel.²
3. *Betrouwbaarheid*: De correlatie van de oordelen van een beoordelaar met de contraoordelen.
4. *Slaagpercentage*: Het percentage van de beoordeelde kandidaten, dat slaagt op het examen.
5. *Percentage vervallen beoordelingen*. Het percentage van de beoordelingen van een beoordelaar dat is vervangen door een derde beoordeling. Een derde beoordeling wordt verricht indien:
 - a. De score die door een beoordelaar aan een kandidaat wordt toegekend meer dan 20% van de maximum score afwijkt van de score die door de contrabeoordelaars is toegekend.
 - b. De totaalscore van een beoordelaar tot een andere beslissing leidt dan die van de contrabeoordelaars. Hiervoor wordt een voorlopige cesuur gebruikt die wordt bepaald met behulp van proeftoets gegevens.

Na een derde beoordeling worden de twee oordelen gekozen die het dichtst bij elkaar lagen. Het derde oordeel vervalt.

De vergelijking tussen een beoordelaar en alle anderen wordt verder aanschouwelijk gemaakt met behulp van een plaatje zoals in Figuur 3, waarin verschillende soorten informatie tegelijkertijd zichtbaar zijn. De rechte haken geven het *Scorebereik* aan van de oordelen van de individuele beoordelaar. De zwarte balk doet datzelfde voor alle beoordelaars samen.

Figuur 3. Een plaatje uit de beobrief waarin informatie over de locatie van de beoordelingen wordt samengevat.



Het doel is om een beoordelaar te vergelijken met *een doorsnee beoordelaar die een doorsnede van de kandidaten beoordeeld*. Door de systematiek van toewijzen bereiken we dit doel niet.

- Een beoordelaar wordt steeds vergeleken met een beperkt aantal andere beoordelaars, die niet willekeurig zijn gekozen uit de verzameling van alle beoordelaars.

² Bepaald door eerst de mediaan te nemen en dan van scores aan weerszijde van de mediaan opnieuw.

- Het is mogelijk dat een beoordelaar systematisch wordt toegewezen aan een bepaalde soort kandidaten. Hierdoor is niet uit te sluiten dat een beoordelaar (ook over examens heen) terecht relatief lage of hoge scores toekent.

De veronderstelling dat de contrabeoordelaars een doorsnede vormen van alle beoordelaars is essentieel om de correlatie te kunnen interpreteren als een maat voor de betrouwbaarheid van een beoordelaar. Een aantal modelmatige schattingen van betrouwbaarheid die we hebben geprobeerd gaven resultaten die niet overeenkwamen met de correlaties. Wanneer we, in de volgende paragraaf de geschiedenis van beoordelaars beschouwen blijkt ook dat de correlatie met de andere beoordelaar geen verband vertoont met andere variabelen. Beoordelaars die min of meer willekeurig (met de ogen dicht) invullen rond het midden van de schaal kunnen we op dit moment niet goed detecteren. *Mildheid* levert echter een constant beeld op.

De geschiedenis van beoordelaars

Het verzamelen van informatie over individuele beoordelaars is begonnen op kleine schaal voor onderzoeksdoeleinden. Nadat enkele jaren “proefdraaien” is er een programma geschreven waarmee de gegevens van iedere beoordelaar gemakkelijk kunnen worden opgevraagd.

Ter illustratie geven we de volgende tabel met informatie over beoordelaar nummer 6.³

Tabel 5. Overzicht van beoordelingen verricht door beoordelaar nummer 6. De rijen zijn examens. Bijvoorbeeld, “sp1jan03” betekent: Spreken I examen van januari 2003. Een “m” of een “o” betekent een middag- of ochtendexamen.

Examen	Aantal kandidaten	Mildheid	Betrouwbaarheid	% derde beoordelingen
sp2jun01	38	-1,71	0,8986	0
sp1jun02	27	-7,48	0,6288	4
sp1jan03	26	-10,85	0,8417	8
sp1jun03	27	-9,3	0,7622	0
sp2mrt04	27	2,52	0,8961	4
sv2jul04m	20	-6,2	0,7539	5
sv2jul04o	23	-1,48	0,8435	4
sp1sep04	24	-2,33	0,505	0
sp2jul05m	15	0,07	0,6875	13
sv2mrt05	40	-0,65	0,8442	8

Merk op dat de absolute waarde van Mildheid afhankelijk is van het bereik van de scores op de examens. Daardoor is het tussen examens van verschillende vaardigheden niet te vergelijken. We kijken dan ook vooral naar het teken. Het scorebereik van de examens is wel redelijk constant over de jaren.

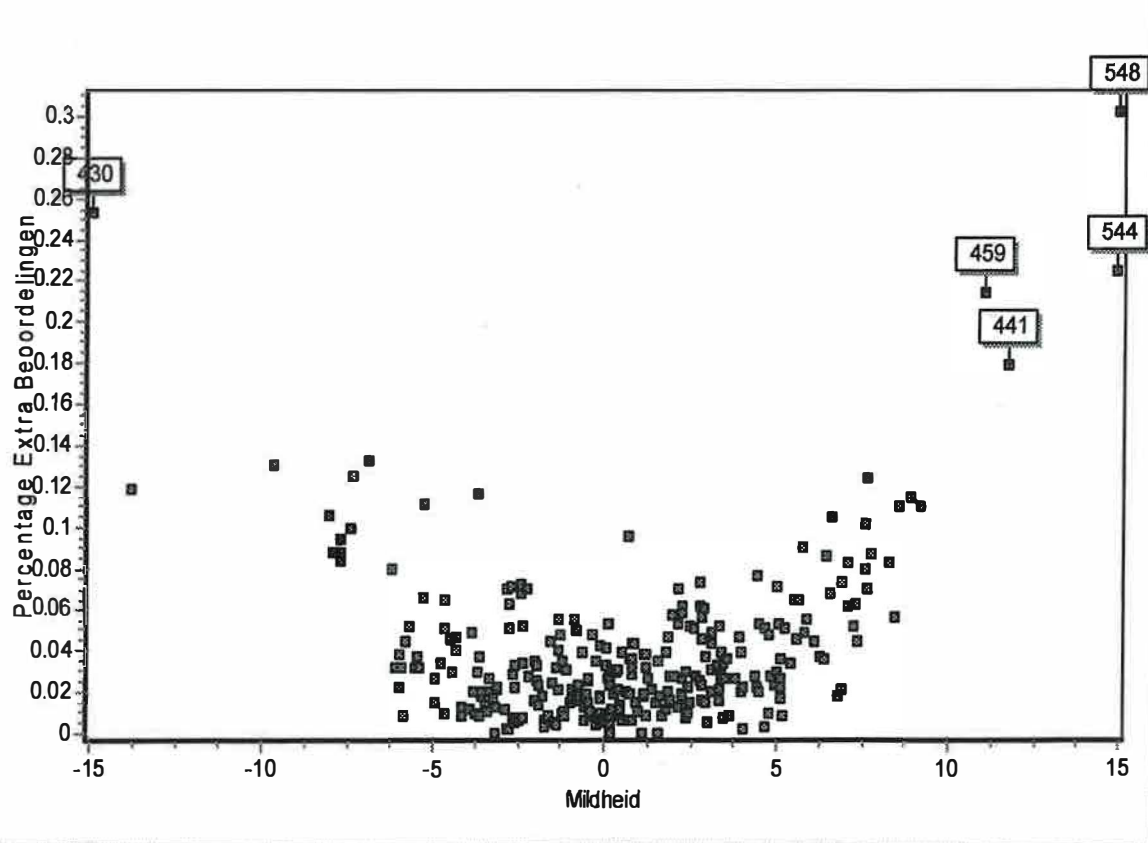
We zien dat beoordelaar nummer 6, over de jaren, lagere scores gaf dan gemiddeld (mildheid is met één uitzondering negatief). Gemiddeld slaagt slechts 48 percent van de kandidaten die door beoordelaar 6 zijn beoordeeld. De gegevens suggereren daarmee dat beoordelaar 6 een

³ Beoordelaar 6 is bewust gekozen omdat hij een lid is van de staatsexamencommissie. Slaagpercentages worden nog niet zo lang verzameld en zijn in de tabel daarom niet vermeld.

streng beoordelaar is. De betrouwbaarheid van beoordelaar 6 is gemiddeld 0,77. Samenvattend concluderen we dat beoordelaar 6 een relatief strenge, matig betrouwbare beoordelaar is. Ook hier is de vraag of we kunnen beargumenteren dat deze conclusie juist is.

Het is duidelijk dat de geschiedenis van een beoordelaar informatie kan verschaffen over consequent afwijkend gedrag. Belangrijk is dat uit de gegevens blijkt dat de meerderheid van de beoordelaars over de jaren heen een constante “stijl” van beoordelen heeft. Daarnaast kunnen we nagaan wat, over jaren heen, de relatie is tussen verschillende aspecten van het functioneren van beoordelaars. Bijvoorbeeld, de relatie tussen Mildheid en het Percentage derde beoordelingen die is afgebeeld in Figuur 4. In Figuur 4 is te zien is dat er meer extra beoordelingen worden verricht voor extreem milde of extreem strenge beoordelaars. Zie ook Figuur 5. Van enkele opvallende beoordelaars is het nummer aangegeven. Onze interpretatie van dit resultaat is dat Mildheid ten minste samenhangt met iets dat duidt op een afwijkende wijze van beoordelen.

Figuur 4. Een plot van het gemiddelde percentage vervallen beoordelingen tegen de gemiddelde mildheid van beoordelaars. De beoordelaars hebben een minimaal aantal kandidaten (15) beoordeeld per examen en een minimaal aantal examens (5).

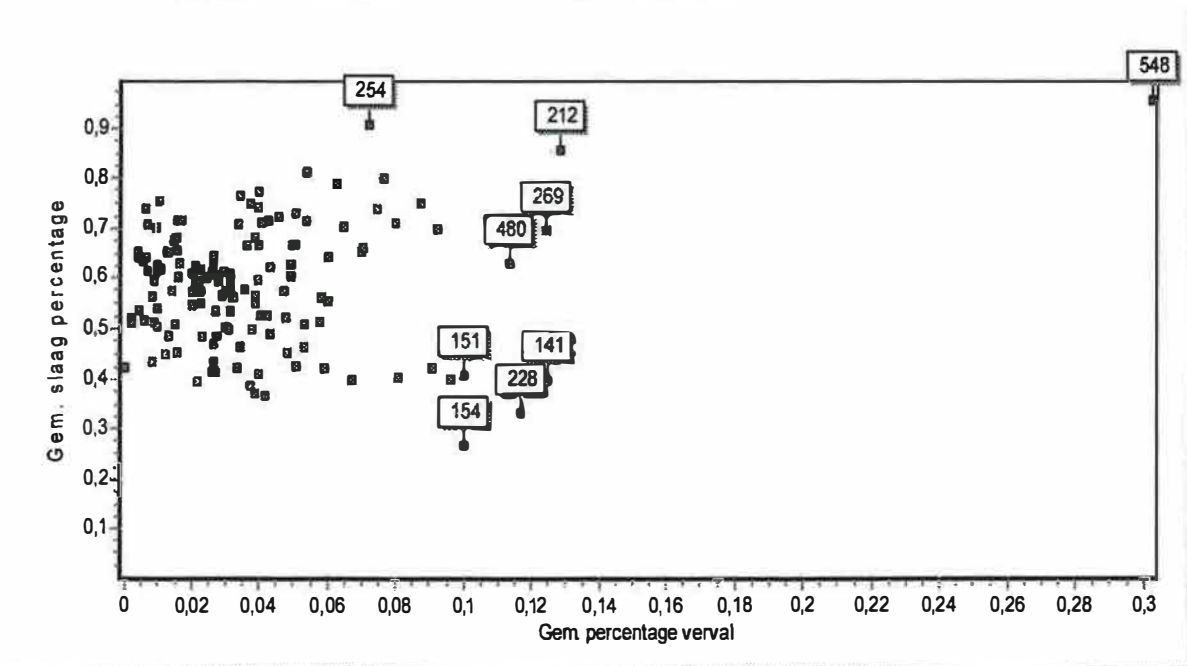


Ter illustratie bevat Tabel 6 enkele gegevens over beoordelaar nummer 430. Gemiddeld was het percentage kandidaten dat bij beoordelaar 430 slaagt 10.5%. Ter vergelijking, bij beoordelaar 548 was dit percentage 95%.

Tabel 6. Overzicht van beoordelingen verricht door beoordelaar nummer 430.

Examen	Aantal kandidaten	Mildheid	Betrouwbaarheid	%derde beoordelingen
sp1jan05	36	-15,58	0,8169	22
sp1jun01	28	-19,36	0,6451	21
sp1sep04	28	-17,64	0,8225	11
Sp1dec01	39	-17,08	0,8183	26
sp1jun02	25	-22,04	0,8372	64
sp2mrt04	30	-15,70	0,8620	37
sp2sep03	39	-6,21	0,8980	5
sp2jun01	84	-5,92	0,8179	17

Door te kijken naar de geschiedenis van beoordelaars kunnen we afwijkende beoordelaars op het spoor komen. Zie bijvoorbeeld ook Figuur 5. De vraag is of op basis van de geschiedenis van een beoordelaar een *valide* oordeel kan worden geveld. Ook hier geldt dat de onderbouwing van het oordeel wordt bemoeilijkt doordat de contrabeoordelaar over examens heen vaak dezelfde is, terwijl de systematiek van toewijzing ook niet uitsluit dat een beoordelaar steeds dezelfde soort kandidaten beoordeeld (zie de volgende paragraaf).

Figuur 5. Een plot van het gemiddelde percentage extra beoordelingen tegen de gemiddelde slaagpercentage van beoordelaars. Ook hier zijn afwijkende beoordelaars goed zichtbaar.

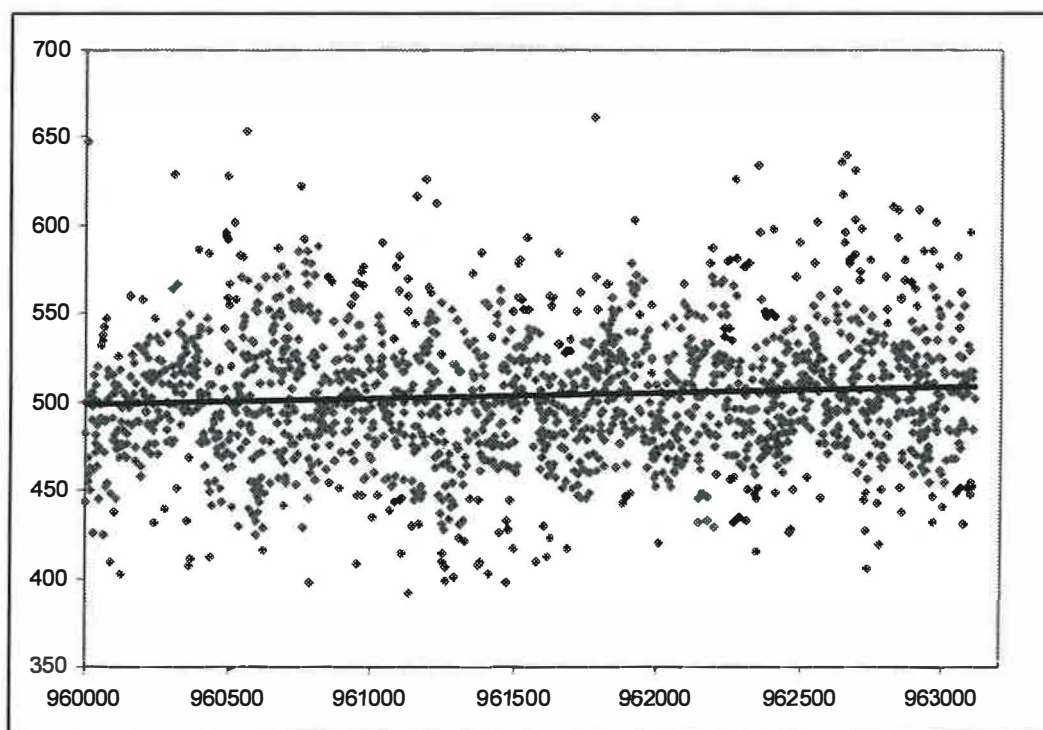
Ter controle hebben we enkele consequent milde, consequent strenge, en enkele willekeurig gekozen beoordelaars uitgezocht. Van iedere beoordelaar zijn enkele kandidaten herbeoordeeld en op basis hiervan geclassificeerd als streng, mild, of goed. Het bleek dat de resultaten overeenstemden met het eerdere oordeel op basis van de geschiedenis van de beoordelaars (zie Bergsma en Bechger, 2003 en Lam, 2004).

De relatie tussen het nummer en de presentaties van kandidaten

De systematiek van toewijzing suggereert de mogelijkheid dat het nummer van de beoordelaar gerelateerd is aan het soort kandidaten dat een beoordelaar krijgt toegewezen.

Het aantonen van een dergelijk relatie is gecompliceerd doordat eigenschappen van beoordelaars en kandidaten moeilijk los van elkaar kunnen worden bekeken. Over de jaren heen zien we een zwak verband tussen het nummer van kandidaten en hun cijfer. Het verband is soms positief, soms negatief en soms niet-lineair. Figuur 6 toont een voorbeeld waarbij het verband positief was.

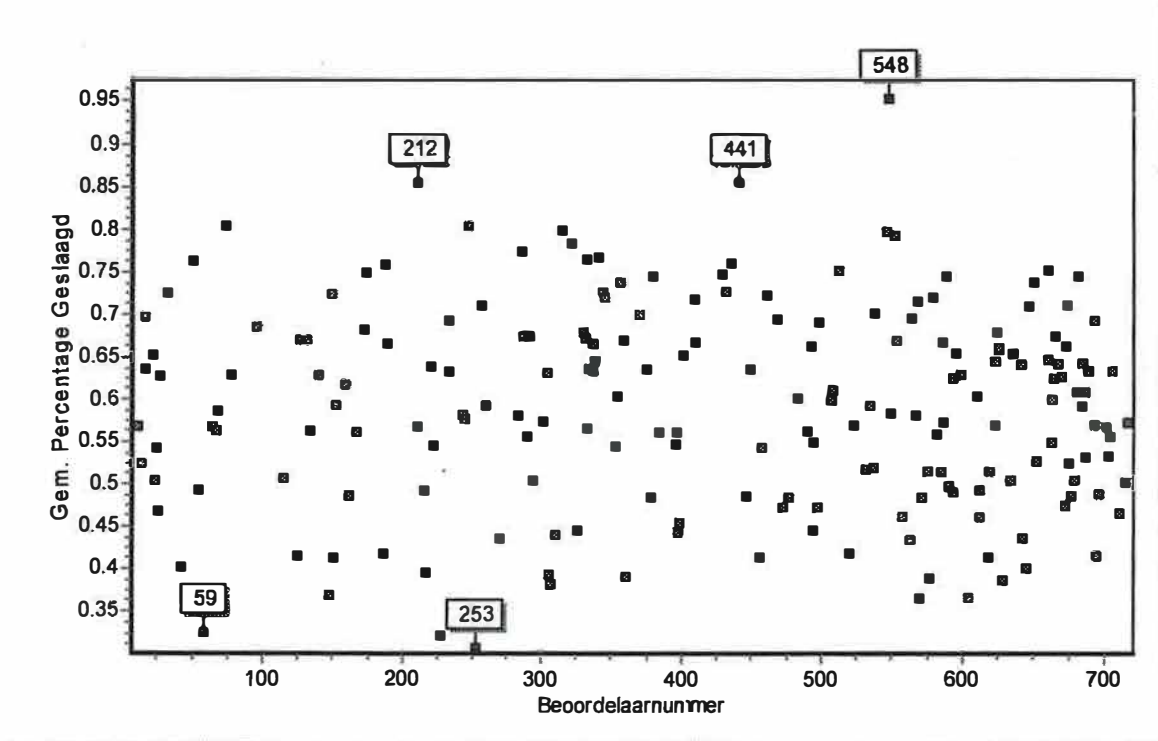
Figuur 6. Relatie tussen kandidaat-nummer en score op de 500-schaal voor het examen Spreken II van maart 2005.



Zelfs als we, bijvoorbeeld, aannemen dat kandidaten met hogere nummers beter presteren, is het moeilijk om te bepalen welk verband dit impliceert tussen de oordelen en de beoordelaarsnummers. Welke nummers volgens de huidige systematiek toegewezen worden aan kandidaten met hogere nummers is afhankelijk van het aantal kandidaten, de beschikbare beoordelaars, en het aantal beoordelingen dat beoordelaars wensen te verrichten.

Kijken we naar spreidingsplots zoals Figuur 7 dan lijkt in de huidige gegevens geen evidentie te zijn voor een verband tussen de beoordelingen en het beoordelaarsnummer.

Figuur 5. Spreidingsplot van gemiddelde slaagpercentages en beoordelaarnummer.



Hoe komt het oordeel in de beoebrieven tot stand?

Het oordeel in de beoebrieven is op dit moment *niet* gebaseerd op de geschiedenis van een beoordelaar maar uitsluitend op gegevens van het betreffende examen.

Allereerst wordt vastgesteld of het percentage geslaagde kandidaten bij een beoordelaar statistisch significant afwijkt van het percentage op het examen. Dat zal dus eerder het geval zijn bij beoordelaars die veel kandidaten hebben beoordeeld. Wanneer het slaagpercentage afwijkt, wordt gekeken of een beoordelaar te streng is of te mild waarbij beoordelaars met veel derde beoordelingen volgens strengere criteria worden beoordeeld. De criteria voor mild(streng)-heid zijn min of meer willekeurig gekozen op ongeveer één standaardafwijking van de mildheden door te kijken waar, bijvoorbeeld in Figuur 4, de grootste groep beoordelaars ligt.

Aangezien de criteria daarmee voor een deel arbitrair zijn is het redelijk om bijvoorbeeld enkele kandidaten opnieuw te beoordelen alvorens een beoordelaar aan te spreken op zijn of haar gedrag.

Het effect van de beobrief

Hoewel het nog te vroeg is om definitieve uitspraken te doen, kunnen we toch al iets zeggen over het effect dat de beobrieven hebben gehad op de kwaliteit van de examens.

Onderstaande tabel geeft een overzicht van:

- Het gemiddelde van het percentage beoordelingen dat van iedere beoordelaar opnieuw moest worden gedaan
- De betrouwbaarheid van het examen. Een schatting van de correlatie tussen twee afnamen onder identieke condities (zie Longford 1995, Hfd. 2). Dus met dezelfde beoordelaars.

Tabel 7. Examen, hun betrouwbaarheid en het gemiddelde percentage vervallen beoordelingen over een aantal jaren.

	Jaar	%vervallen	Betrouwbaarheid
spreken 1	2002	3.2	0,88
	2003	3.3	0,90
	2004	4.6	0,88
	2005	2.2	0,93
brief			
brief			
spreken 2	2002	4.7	0,89
	2003	4.9	0,88
	2004	3.3	0,90
	2005	2.3	0,93
brief			
brief			
Schrijven 2	2002	5.0	0,94
	2003	3.7	0,94
	2004	4.9	0,93
	2005	3.0	0,95
brief			

In Tabel 7 is aangegeven wanneer beobrieven zijn verzonden. Interessant is vooral het verschil tussen jaren *met* - en *zonder* beobrief. We zien dat in de jaren waarin de beobrief werd verstuurd, de betrouwbaarheid is gestegen en het aantal vervallen beoordelingen is gedaald ten opzichte van de voorgaande jaren. Er is één uitzondering; Spreken 1, 2004.

Een andere manier om het effect van de beobrief zichtbaar te maken is door te kijken naar het percentage van de brieven waarin geen negatief oordeel werd gegeven. Tabel 8 geeft een overzicht. Als beoordelaars beter beoordelen, zo is de verwachting, dan zal dit percentage stijgen.

Tabel 8. Percentages beoordelaars die een brief kregen met het oordeel “gemiddeld”. Daar waar er een ochtend en een middag examen was zijn de percentages gemiddeld. Het totale aantal brieven dat bij deze examens is verstuurd is aangegeven met een sterretje. De aantallen zijn relatief groot omdat sommige beoordelaars twee brieven kregen.

Spoken I		
Examen	%	Aantal brieven
sp1jun04	65	96
sp1sep04	56	63
sp1jan05	65	63
sp1jun05	75	99
sp1sep05	70	81

Spoken II		
Examen	%	Aantal Brieven
sp2jul04	62	132
sp2dec04	61	138
sp2maa05	78	75
sp2jun05	82	268*

Schrijven I		
Examen	%	Aantal Brieven
sv1jun04	59	59
sv1sep04	64	44
sv1jan05	59	34
sv1jun05	82	68
sv1sep05	71	56

Schrijven II		
Examen	%	Aantal Brieven
sv2dec04	61	49
sv2maa05	76	90
sv2jun05	81	260*
sv2dec05	68	53

In Tabel 8 zien we dat het percentages gemiddelde beoordelaars in 2005 zijn gestegen ten opzichte van 2004.

Of deze verschillen daadwerkelijk het gevolg zijn van de beobrieven is niet met zekerheid te zeggen, aangezien er over de jaren ook andere aspecten van de examens veranderd zijn. Zoals het type opgaven, de kandidaten, het aantal kandidaten dat beoordelaars beoordelen, etc. Bovendien verwachten we een effect van de beobrief eigenlijk pas enkele jaren nadat de brieven zijn geïntroduceerd. De resultaten suggereren in ieder geval een positief effect van de beobrieven.

Aanbevelingen

Willekeurig 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 periodiek hertrainen van beoordelaars

Er is in de data op dit moment geen relatie tussen beoordelaarnummer en het functioneren van een beoordelaar. Niets, wijst er, met andere woorden, op dat beoordelaars iets bijleren naarmate ze langer in het bestand zitten. Naast feedback aan beoordelaars is het wellicht aangewezen om een systeem van periodieke hertraining in te stellen.

Het on-line 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:

- beoordelaars naar het Cito zouden komen om onder observatie hun werk te doen
- 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.

De eerste optie zou het beoordelen veel duurder maken. De tweede optie echter zou op termijn geld besparen en lijkt daarom de meest aantrekkelijke. Deze optie wordt momenteel ook overwogen in verband met de invoering van permanente examenlocaties in 2007.

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 wanneer een beoordelaar de vragen heeft ingevuld en hoeveel tijd een beoordelaar bezig is.
- We kunnen vaststellen of en hoe vaak een beoordelaar een spraakfragment heeft beluisterd.

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 geldt 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.

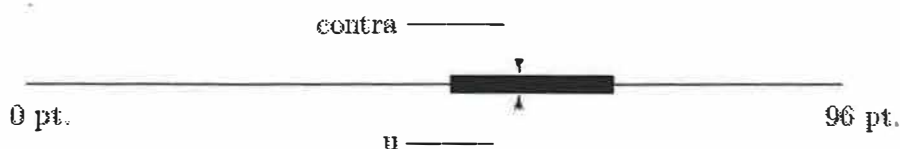
Brief uitbreiden met gegevens over de tweede beoordelaar

Het voorbeeld van beoordelaar 749 heeft laten zien dat de oordelen van andere beoordelaars over dezelfde kandidaten een belangrijke bron van informatie is. We willen deze informatie aan de beoebrieven toevoegen. Bijvoorbeeld,

- een tabel waarbij de oordelen van een beoordelaar worden afgezet tegen de andere oordelen over dezelfde kandidaten (zie Tabel 2),
- een onderscheid maken tussen:
 - mildheid ten opzichte van alle beoordelaars, en
 - mildheid ten opzichte van de beoordelaars van dezelfde kandidaten,
- de andere oordelen in te tekenen in de figuur die in de brief staat (zie Figuur 3 in dit stuk).

Een voorbeeld van een alternatief voor Figuur 3 staat hieronder:

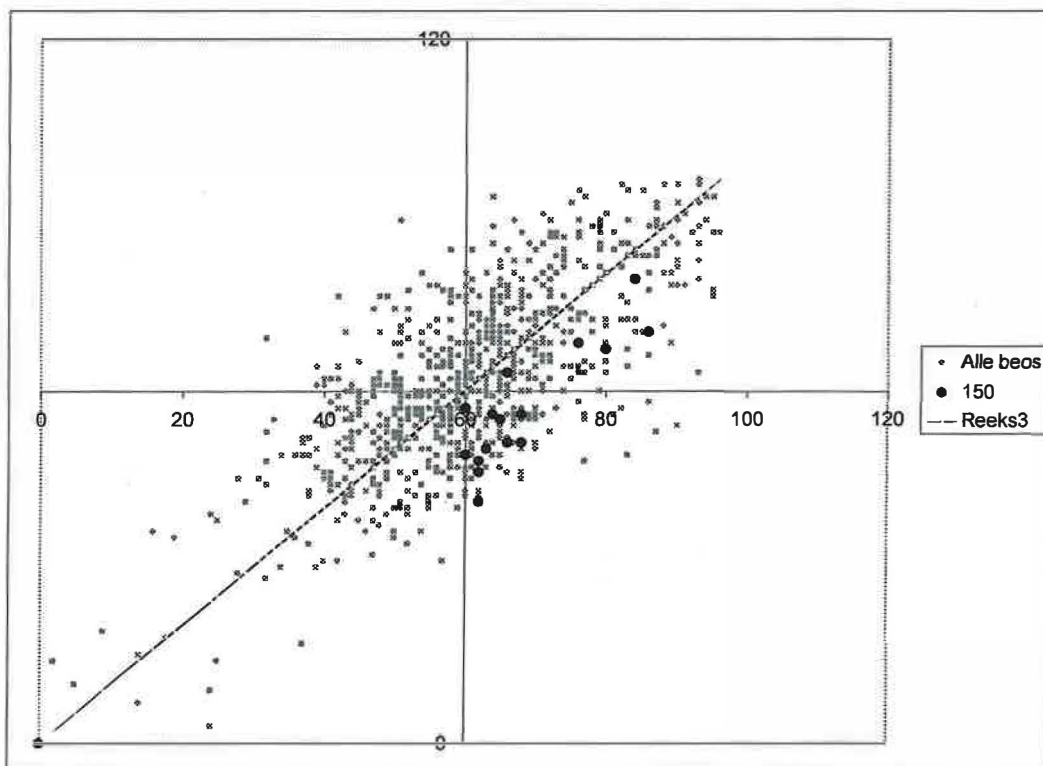
Figuur 8. Een alternatief voor Figuur 3 uitgebreid met informatie over de contrabeoordelingen.



De lijnen geven het scorebereik aan van de individuele beoordelaar en de contrabeoordelaars. Een beoordelaar kan in dit plaatje zien dat hij/zij even streng was als de contrabeoordelaars. Dit suggereert dat de kandidaten die aan deze beoordelaar waren toegewezen minder vaardig waren dan de andere kandidaten.

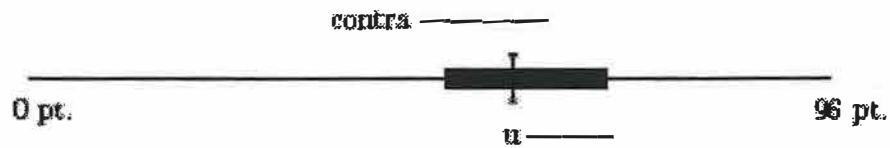
Beter, maar tegelijkertijd moeilijker te verklaren en te visualiseren in een brief, is een spreidingsplot van alle eerste tegen alle contrabeoordelingen waarin de oordelen van een beoordelaar worden uitgelicht.

Figuur 9. Spreidingsplot van eerste en contrabeoordelingen. De oordelen waarbij beoordelaar 150 betrokken zijn uitgelicht. Data komen van het Spreken II examen van maart 2006.



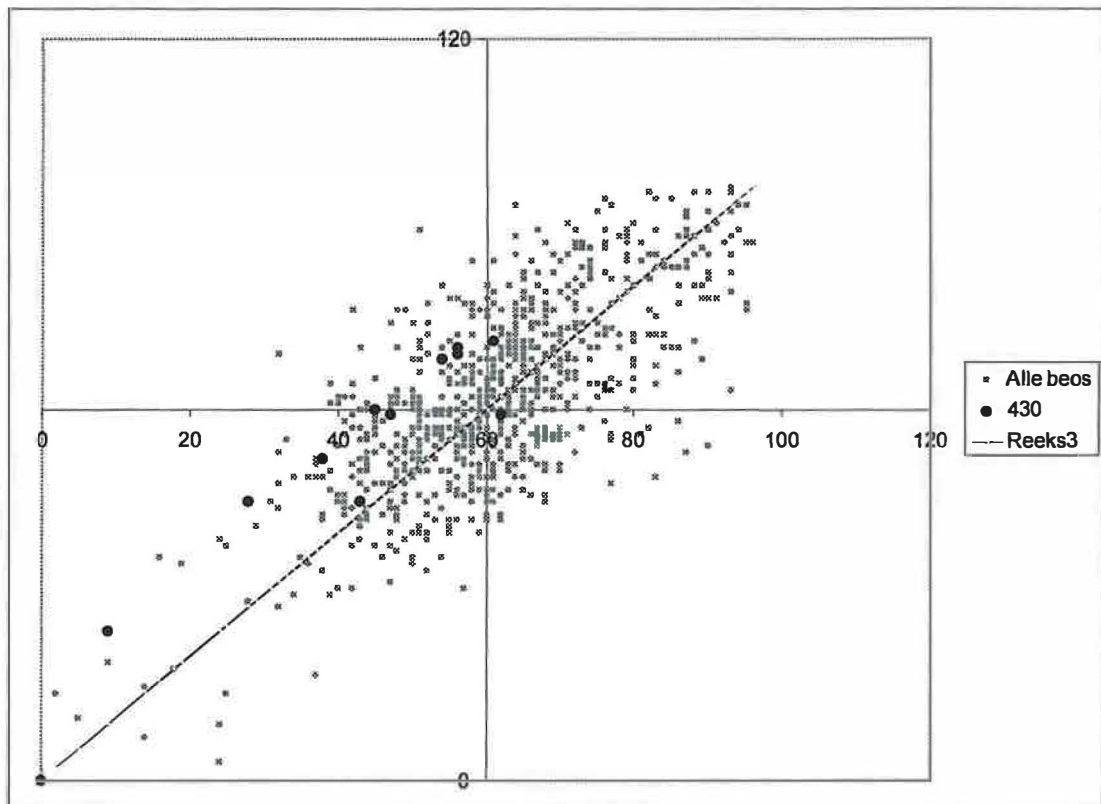
Om de grafiek te kunnen interpreteren zijn de oordelen van beoordelaar 150 altijd geplot als x-waarden (beoordelaar 150 is altijd de eerste beoordelaar). De diagonale is de lijn $x=y$. De puntenwolk geeft aan waar de overige beoordelingen zich bevonden. Doordat de oordelen van beoordelaar 150 altijd onder de lijn $x=y$ liggen zien we dat beoordelaar 150 systematisch hogere scores geeft dan de contrabeoordelaars.

Figuur 10. Alternatieve voorstelling van de oordelen van beoordelaar 150 en zijn of haar contrabeoordelaars ten opzichte van de andere beoordelaars.

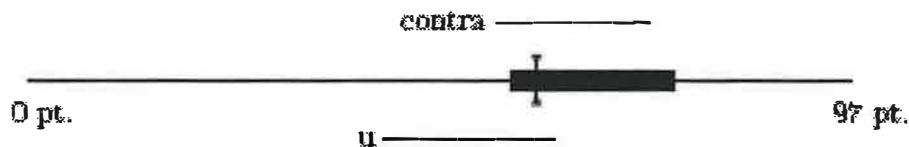


In Figuur 11 en 12 zien we, ten slotte, dat beoordelaar nummer 430, die ons al eerder was opgevallen (zie Figuur 4), duidelijk milder is dan anderen.

Figuur 11. Spreidingsplot van eerste en contrabeoordelingen met de oordelen waarbij beoordelaar 430 betrokken was uitgelicht. Data komen van het Spreken II examen van maart 2006.



Figuur 12. Beoordelaar 430 en zijn of haar contrabeoordelaars ten opzichte van de andere beoordelaars.



Gegevens van logistiek combineren met psychometrische gegevens

Beoordelaars kunnen op meerdere manieren slecht functioneren. Dat een beoordelaar, bijvoorbeeld, formulieren onvolledig of verkeerd invult kan een signaal zijn dat hij of zij niet serieus beoordeelt. Deze informatie zou kunnen worden toegevoegd aan de brief.

Verder kunnen problemen bij de logistieke afhandeling, gecombineerd met de psychometrische gegevens, een sterker argument leveren om beoordelaars te classificeren. Welke informatie we hiervoor kunnen gebruiken en hoe moet nog verder worden onderzocht. Zo melden de logboeken dat beoordelaar nummer 430 zich recentelijk heeft afgemeld wegens overspannenheid (zie Tabel 6) ! Beoordelaar 548 is niet opvallend aanwezig in de logboeken.

Beoordelaarvoorschriften verbeteren

Het is al lang bekend dat de huidige beoordelaarvoorschriften (ook wel *beoordelingsmodellen* genoemd) op een aantal punten verbeterd, en vooral versimpeld, kunnen worden (bijv. Akker, van den, 1995; Berlo, 1997; Berlo, de Groot en de Jong, 1997). In het bijzonder de lage overeenstemming, zowel tussen interne als tussen externe beoordelaars, is zorgwekkend. Een eerdere poging hiertoe was niet succesvol (Lam, 2003). We denken dat een opzet zoals voorgesteld door Maris en Bechger (2004) wel succesvol zou kunnen zijn.

Conclusie: Veroordelen of stimuleren?

Wij denken dat er in de beobrieven beter geen oordeel kan worden gegeven zolang we niet over de goede informatie beschikken. Een voorwaarde is dat de toewijzing van beoordelaars aan kandidaten op basis van toeval plaatsvindt zodat individuele beoordelaars kunnen worden vergeleken met de doorsnee beoordelaar.

Als aandachtspunten voor de toekomst zijn besproken:

- Beoordelaars zouden periodiek bij moeten leren.
- Beoordelingen kunnen beter on-line verzameld worden.
- De beoordelingsvoorschriften moeten worden verbeterd.
- Logistieke gegevens combineren met psychometrische gegevens om afwijkende beoordelaars beter op te kunnen sporen.
- De beobrief uitbreiden met informatie over de tweede beoordelaar.
- Het kalibratiemodel gebruiken om afwijkende beoordelaars te vinden.

Tenslotte. Het versturen van beobrieven zal er in het beste geval toe leiden dat beoordelaars consciëntieuzer en daardoor beter gaan beoordelen. De vraag is of een oordeel daar aan bijdraagt. Een beoordelaar, die ten onrechte een negatief oordeel ontvangt, gaat zich daardoor wellicht anders gedragen dan hij of zij zou moeten doen. Het verdient de aanbeveling het oordeel intern te gebruiken voor het opsporen en corrigeren van slecht functionerende beoordelaars.

Referenties

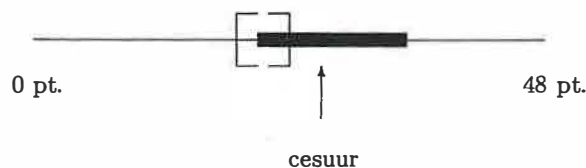
- Akker, A. van den. (1995). *Beoordelingsvoorschriften en beoordelaarsgedrag. Een onderzoek naar de kwaliteit van de beoordelingsvoorschriften van het onderdeel spreken van het Staatsexamen Nederlands als tweede taal*. Arnhem: Cito.
- Bergsma, A., & Bechger, T. M. (2003). *Beoordelaarsgedrag beoordeeld: Een onderzoek naar een instrument waarmee beoordelaarsgedrag geanalyseerd kan worden*. Cito: Arnhem.
- Berlo, H. van. (1997). *Onderzoek naar de mogelijkheden tot vereenvoudiging van de beoordelingsvoorschriften Spreken en Schrijven I en II van het Staatsexamen NT2. Fase 2 inventarisatie functioneren Schrijven I en II*. Cito: Arnhem.
- Berlo, H. van., de Goot, M., en de Jong, J. (1997). *Onderzoek naar de mogelijkheden tot vereenvoudiging van de beoordelingsvoorschriften Spreken en Schrijven I en II van het Staatsexamen NT2. Fase 1 inventarisatie functioneren Spreken I en II*. Cito: Arnhem.
- Lam, Jo Fond. (2003). *Onderzoek naar nieuwe beoordelingsmodellen Staatsexamen Nt2: Spreken en schrijven programmaonderdelen I en II*. Bureau ICE: Lienden.
- Lam, Jo Fond. (2004). *Beoordelaarsgedrag beoordeelt*. Bureau ICE: Lienden.
- Longford, N. I. (1995). *Models for uncertainty in Educational testing*. Springer: New-York.
- Maris G., & Bechger, T. M. (2004). *Het scoren van open vragen: Theorie en Praktijk*. POK memorandum 2004-2.
- Maris, G., & Bechger, T. M. (2006). *Scoring Open Ended Questions*. In "Handbook of Statistics", 27, C. R. Rao & S. Sinharay (Eds.). Amsterdam: North-Holland.

Bijlage:**De brief aan beo 749**

Doorkiesnummer	Ons kenmerk	Uw kenmerk	Datum
026-3521515	05-480-252	749	15 december 2005
E-mail			
aukje.bergsma@citogroep.nl			

Geachte beoordelaar,

U bent betrokken geweest bij de beoordeling van het juli-examen Schrijven van het Staatsexamen NT2. Het ging om het ochtendexamen. Zoals u misschien al weet, zullen wij na elk examen aan de beoordelaars feedback geven op hun beoordelingen. U heeft hierover van ons een brief ontvangen (kenmerk 03-480-276). In het onderstaande schema kunt u zien hoe u het gedaan heeft ten opzichte van de andere beoordelaars.



De [] geven aan waar de meerderheid van uw beoordelingen zich bevindt op de beoordelingsschaal. De zwarte balk geeft aan waar de meeste beoordelaars zich bevinden.

Algemeen geldt: Bevindt u zich links van de zwarte balk, dan beoordeelt u strenger dan de meeste beoordelaars. Bevindt u zich rechts van de zwarte balk dan beoordeelt u milder dan de meeste beoordelaars. Bevindt u zich in de zwarte balk, dan komen uw beoordelingen overeen met die van de meeste beoordelaars. In het belang van de kandidaten streven wij naar zoveel mogelijk overeenstemming tussen beoordelaars. Dan hebben kandidaten de meeste kans om op een vergelijkbare manier beoordeeld te worden.

Uw beoordelingen kunnen we kwalificeren als:

streng

Als u een andere kwalificatie dan *gemiddeld* krijgt, dan verzoeken we u met klem om uw wijze van beoordelen aan te passen. Bij de kwalificaties mild en streng is de afwijking van uw beoordelingen van het gemiddelde namelijk onaanvaardbaar groot.

Wij kunnen u ook nog mededelen dat 0% van uw beoordelingen van dit examen opnieuw zijn gedaan door een derde beoordelaar. Een hoog percentage wordt veroorzaakt door een te groot verschil tussen uw beoordeling en die van de andere beoordelaar. Van de door u beoordeelde kandidaten is 14% geslaagd. Kijken we naar alle kandidaten dan is het slaagpercentage 56%. De percentages kunnen afwijken ten gevolge van de kandidaten die u waren toegewezen.

We vertrouwen erop u hiermee voldoende te hebben geïnformeerd.

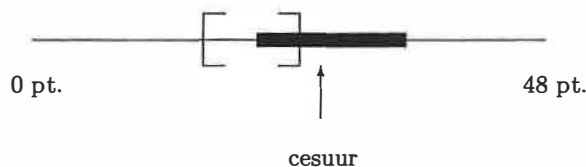
De brief aan beo 445

Doorkiesnummer 026-3521515	Ons kenmerk 05-480-252	Uw kenmerk 445	Datum 15 december 2005
-------------------------------	---------------------------	-------------------	---------------------------

E-mail
aukje.bergsma@citogroep.nl

Geachte beoordelaar,

U bent betrokken geweest bij de beoordeling van het juli-examen Schrijven van het Staatsexamen NT2. Het ging om het ochtendexamen. Zoals u misschien al weet, zullen wij na elk examen aan de beoordelaars feedback geven op hun beoordelingen. U heeft hierover van ons een brief ontvangen (kenmerk 03-480-276). In het onderstaande schema kunt u zien hoe u het gedaan heeft ten opzichte van de andere beoordelaars.



De [] geven aan waar de meerderheid van uw beoordelingen zich bevindt op de beoordelingsschaal. De zwarte balk geeft aan waar de meeste beoordelaars zich bevinden.

Algemeen geldt: Bevindt u zich links van de zwarte balk, dan beoordeelt u strenger dan de meeste beoordelaars. Bevindt u zich rechts van de zwarte balk dan beoordeelt u milder dan de meeste beoordelaars. Bevindt u zich in de zwarte balk, dan komen uw beoordelingen overeen met die van de meeste beoordelaars. In het belang van de kandidaten streven wij naar zoveel mogelijk overeenstemming tussen beoordelaars. Dan hebben kandidaten de meeste kans om op een vergelijkbare manier beoordeeld te worden.

Uw beoordelingen kunnen we kwalificeren als:

streng

Als u een andere kwalificatie dan *gemiddeld* krijgt, dan verzoeken we u met klem om uw wijze van beoordelen aan te passen. Bij de kwalificaties mild en streng is de afwijking van uw beoordelingen van het gemiddelde namelijk onaanvaardbaar groot.

Wij kunnen u ook nog mededelen dat 0% van uw beoordelingen van dit examen opnieuw zijn gedaan door een derde beoordelaar. Een hoog percentage wordt veroorzaakt door een te groot verschil tussen uw beoordeling en die van de andere beoordelaar. Van de door u beoordeelde kandidaten is 14% geslaagd. Kijken we naar alle kandidaten dan is het slaagpercentage 56%. De percentages kunnen afwijken ten gevolge van de kandidaten die u waren toegewezen.

We vertrouwen erop u hiermee voldoende te hebben geïnformeerd.

