De praktijk van de eerste en tweede correctie

Samenvatting van onderzoek naar het functioneren van het CSE

Hans Kuhlemeier en Ed Kremers
Cito, Arnhem
Versie juli 2012

Inhoud

1. Inleiding

Aan het einde van het voortgezet onderwijs leggen de leerlingen een examen af. Een examen is een door een bevoegde instantie ingestelde toetsing van leerresultaten waaraan de kandidaat bepaalde rechten of bevoegdheden kan ontlenen. Het diploma voortgezet onderwijs heeft nog steeds een grote persoonlijke, maatschappelijke en economische waarde. Dat wordt door vervolgopleidingen, werkgevers, docenten, ouders en kandidaten ook als zodanig erkend. Het examen beschermt leerlingen tegen ondeugdelijk onderwijs en geeft de samenleving zekerheid over wat er is geleerd. Het is dan ook van groot belang dat het functioneren van het examensysteem regelmatig onderzocht en geëvalueerd wordt.

De Nederlandse examens bewegen zich in het spanningsveld tussen de verantwoordelijkheid van de overheid en de vrijheid van de school. Van de ene kant stelt de overheid vanuit het oogpunt van kwaliteitsbewaking eisen aan de inhoud en organisatie van de examens. Van de andere kant krijgen scholen binnen deze eisen ruimte om het examen naar eigen onderwijsinhoudelijke en levensbeschouwelijke inzichten in te richten. De huidige examens in het voortgezet onderwijs bestaan uit een centraal examen (CE) en een schoolexamen (SE). De verantwoordelijkheid voor de afname en beoordeling van het CE en het SE heeft de wetgever in handen van de school gelegd. Zo is de correctie van het examenwerk in handen van de ‘eigen’ docent die de kandidaten heeft opgeleid. Hiermee geeft de overheid aan veel vertrouwen te hebben in de professionaliteit en integriteit van de school en de examinatoren. In dit opzicht nemen de Nederlandse examens in de wereld een unieke plaats in. Hoe uitzonderlijk het Nederlandse examensysteem is, wordt duidelijk als men het aan buitenlandse toetsdeskundigen probeert uit te leggen. De eerste reacties zijn altijd die van onbegrip en ongeloof. Kenmerkende reacties zijn ‘Nou, dat zou bij ons niet werken’, ‘Weet je wel zeker dat het werkt?’ en ‘En hoe weet je dat het werkt?’. Bij gebrek aan harde onderzoeksgegevens is een antwoord op deze vragen moeilijk te geven. In deze publicatie doen we verslag van drie studies die tot doel hebben hierover meer te weten te komen.

Examens in het voortgezet onderwijs zijn een grootschalig gebeuren (Alberts & Erens, 2012). Voor het voortgezet onderwijs maakt het Cito jaarlijks meer dan vijfhonderd examens. In 2011 namen er in het vmbo ongeveer 103.000 kandidaten deel, in het havo waren het er circa 56.700 en in het vwo bijna 40.000. In het Nederlandse voortgezet onderwijs is de correctie van het examenwerk in handen van de eigen docent. Kandidaten hebben recht op een professionele, objectieve en rechtvaardige beoordeling. Idealiter zou het niet mogen uitmaken wie het examen nakijkt. In de praktijk blijkt de ene docent echter soepeler te beoordelen dan de andere docent. Om ertoe bij te dragen dat iedere kandidaat het cijfer krijgt dat hij verdient, heeft de overheid de tweede correctie in het leven geroepen. Het examenwerk wordt daarom nog een keer nagekeken door een corrector van een andere school, de tweede corrector. Verschillen tussen eerste en tweede correctoren horen onvermijdelijk bij het correctiewerk en hebben een belangrijke functie: ze leiden niet alleen tot een meer evenwichtige beoordeling, maar ook tot intervisie (Algra, 2004). Het systeem van eerste en een tweede correctie biedt geen garantie dat twee kandidaten met hetzelfde werk ook hetzelfde cijfer krijgen. Het zou er echter wel voor moeten zorgen dat de eigen docent bij de correctie niet zomaar zijn gang kan gaan (Algra, 2004). Over de vraag naar de grootte van de verschillen tussen docenten in soepelheid zijn geen recente gegevens beschikbaar. Evenmin is bekend welke invloed de tweede correctie heeft op de uiteindelijke scores van de kandidaten. In deze publicatie geven we een antwoord op deze twee vragen. Het is gebaseerd op drie studies waarbij correctoren een steekproef van examenwerken uit de ‘echte’ examens opnieuw hebben nagekeken (Kuhlemeier, Van Rijn & Kremers, 2012; Kuhlemeier, Gitsels, Boom, Van de Kerkhof & Sinkeldam, 2012; Gitsels & Kuhlemeier, 2012).

Voor de uitvoering van de eerste en tweede correctie heeft de overheid regels opgesteld. Tweede correctoren moeten het werk van de kandidaten integraal nakijken. Dit wil zeggen dat de tweede

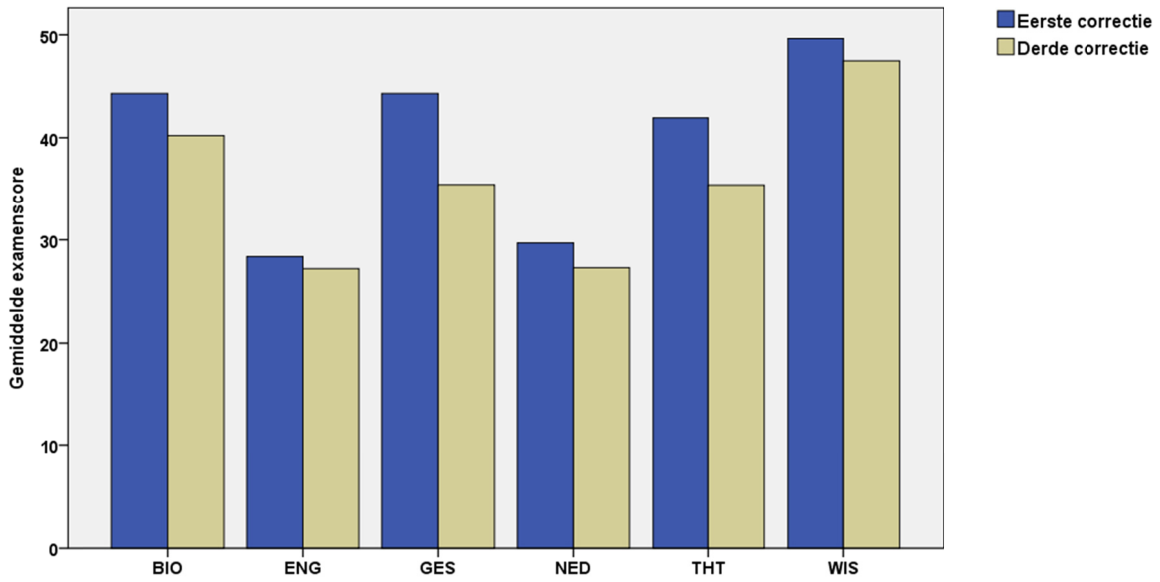
corrector alle werken nakijkt en per kandidaat het volledige examenwerk met alle vragen. Daarnaast zijn er regels voor de wijze waarop de uiteindelijke score in het overleg tussen eerste en tweede corrector tot stand moet komen. Als de tweede corrector vindt dat er sprake is van grote onzorgvuldigheid, aperte fouten of verkeerde interpretatie van de correctievoorschriften dient hij of zij er eerst in overleg met de eerste corrector uit te komen. Als dat niet lukt, kan het scoreverschil worden gemiddeld. De Inspectie van het onderwijs ziet middelen echter als een zwaktebod dat niet past bij professioneel handelende vakdeskundigen (Inspectie van het Onderwijs, 2006). Als de eerste en/of tweede corrector niet willen middelen, kan de tweede corrector zich tegenwoordig melden bij zijn eigen bevoegd gezag die dan contact kan opnemen met het bevoegd gezag van de eerste corrector. Als beiden het niet eens kunnen worden, melden zij dit bij de Inspectie en deze kan vanuit haar toezichthoudende taak bij de examens optreden. Dit kan betekenen dat de Inspectie besluit tot de inzet van een derde onafhankelijke corrector. Uiteraard kan deze procedure ook worden toegepast bij klachten over het werk van de tweede corrector. In deze publicatie doen we ook verslag van een inventariserend onderzoek naar de volledigheid van de tweede correctie en de wijze waarop het overleg tussen eerste en tweede correctoren plaatsvindt (Kuhlemeier & Kremers, 2012). Daarbij besteden we ook aandacht aan de randvoorwaarden waaronder docenten de correctie uitvoeren. De inventarisatie is een vervolg op een soortgelijk onderzoek naar de praktijk van de Centraal Schriftelijke en Praktische Examens (CSPE) in het vmbo (Kuhlemeier & Dietvorst, 2009).

2. Het Onderzoek Tweede Correctie (O2C)

In het Onderzoek Tweede Correctie (O2C) hebben onafhankelijke 'derde' correctoren ruim zeshonderd examenwerken van zes examenvakken opnieuw nagekeken (Kuhlemeier, Van Rijn & Kremers, 2012). De examens waren Nederlands vwo, Engels vmbo, wiskunde vmbo, biologie vwo, geschiedenis havo en tehatex havo. Naast de scores van de eigen docent beschikten we ook over de met de tweede corrector overeengekomen scores. Per examen zijn vijf zogeheten derde correctoren in het onderzoek betrokken. Elk examenwerk is door een steekproef van telkens twee van deze vijf correctoren twee keer nagekeken. Van elk examenwerk is zowel een geannoteerde als een blanco versie nagekeken (waarbij elke versie telkens aan een andere corrector is voorgelegd). De blanco versie was identiek aan de geannoteerde versie, met dien verstande dat de punten en aantekeningen van de eerste corrector digitaal verwijderd waren.[1] De derde correctoren hadden de opdracht de examenwerken zorgvuldig en objectief na te kijken overeenkomstig het correctievoorschrift. Allen waren ervaren docenten met ervaring als eerste en tweede corrector. Zij kregen ruim de tijd en een redelijke vergoeding. In het O2C-onderzoek zijn de scores van de 'eigen' docent vergeleken met die van de 'onafhankelijke' derde correctoren. Daarnaast is gekeken naar het verschil tussen de scores van de eerste corrector en de scores zoals vastgesteld in het overleg tussen eerste en tweede corrector.

2.1 Verschillen tussen eerste en derde correctoren

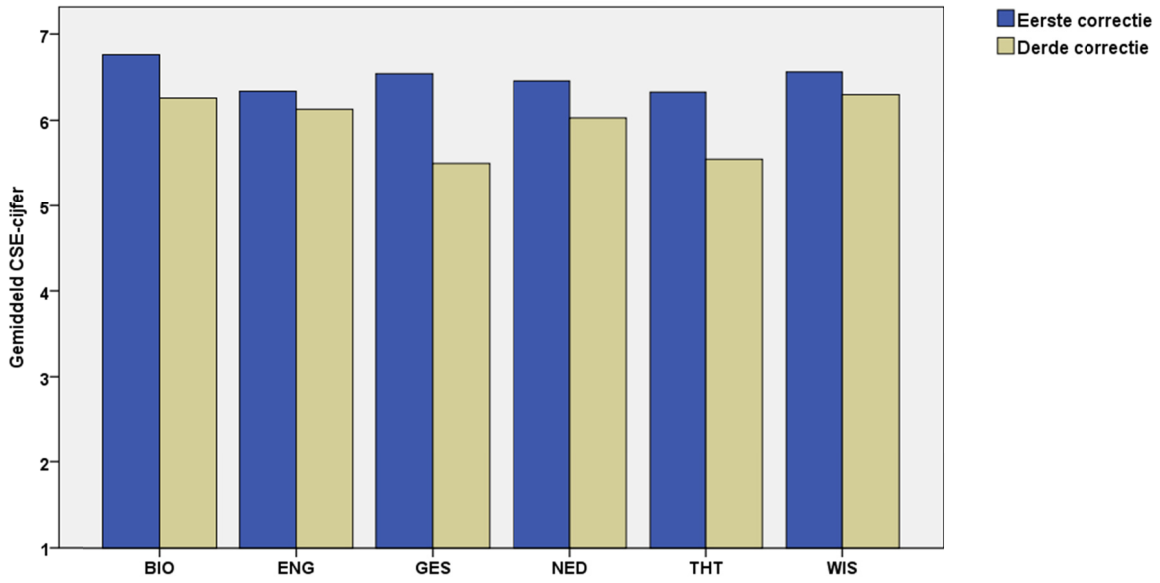
In het O2C-onderzoek zijn de scores van de eerste correctoren vergeleken met die van de derde correctoren. Figuur 1 toont de gemiddelde examenscores per examen (voor het eerste tijdvak). Het verschil tussen de examenscores varieert afhankelijk van het examen van één tot ruim negen scorepunten. Telkens zijn de scores van de eerste correctie gemiddeld hoger dan die van de derde correctoren.



Figuur 1 Gemiddelde examenscore voor de eerste en derde correctie per examen

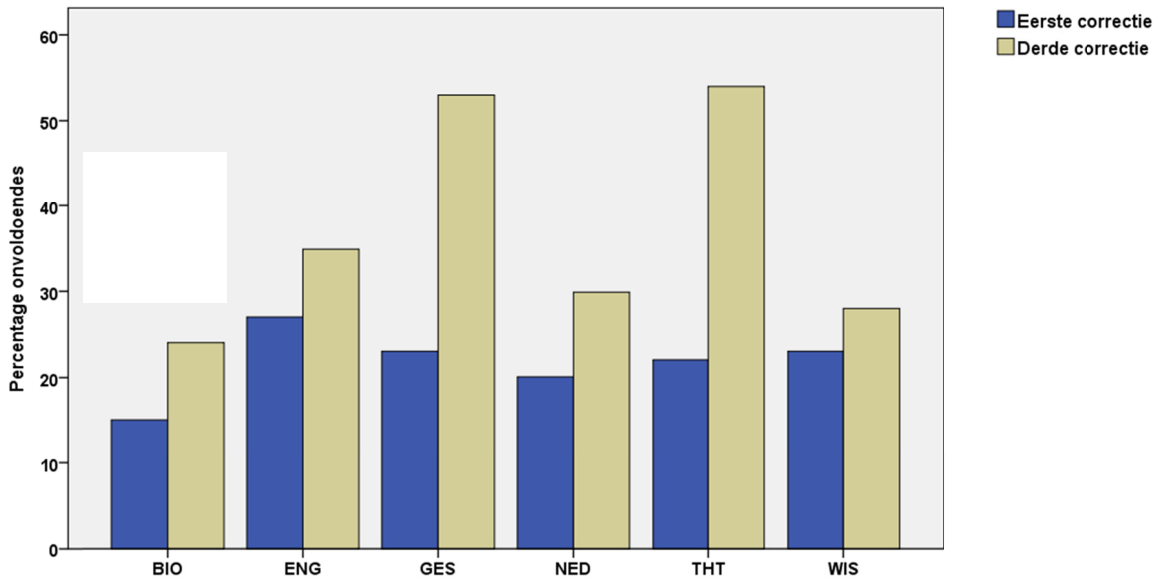
[1] In het O2C-onderzoek is ook vastgesteld dat derde correctoren zich laten beïnvloeden door de aantekeningen die de eerste corrector op het examenwerk aanbrengt. Omdat deze conclusie voor deze publicatie van onderschikt belang is, gaan we hier verder niet op in.

Het verschil in soepelheid tussen de eerste en derde correctoren zien we ook terug in de examencijfers (zie Figuur 2). De eerste correctie leidt bij alle zes examens tot hogere cijfers dan de derde correctie. Bij Engels gaat het om een vijfde punt, bij wiskunde om een kwart punt, bij Nederlands en biologie om een half punt, bij tehatex om vier vijfde punt en bij geschiedenis om een vol cijferpunt.



Figuur 2 Gemiddeld cijfer voor de eerste en derde correctie per examen

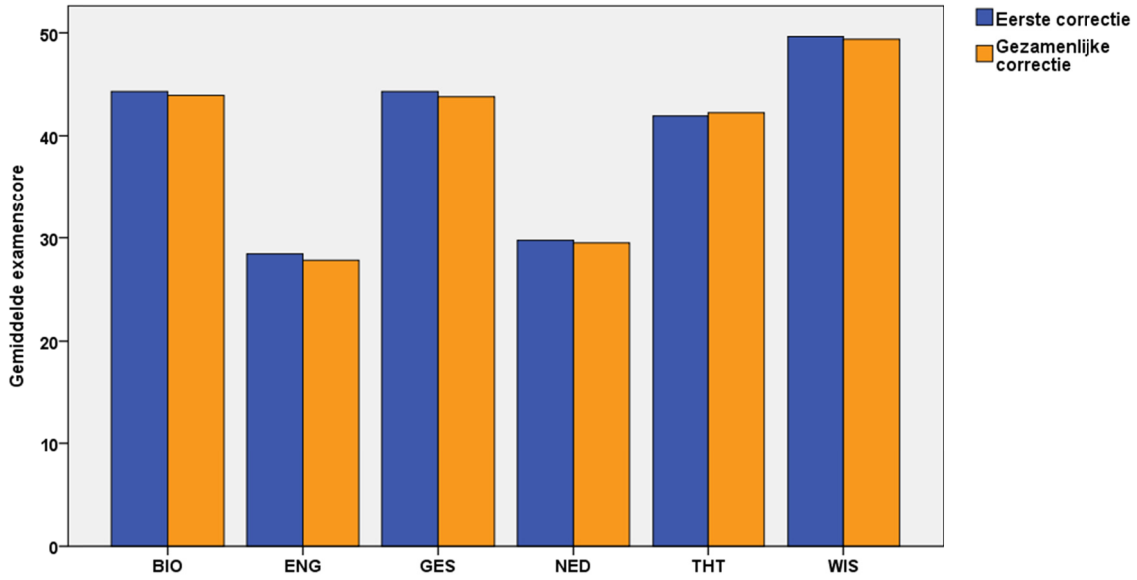
Figuur 3 geeft inzicht in de percentage onvoldoendes waartoe de eerste en derde correctie geleid zouden hebben (op basis van eerste tijdvak). Stel dat we de eerste correctoren zouden vervangen door de derde correctoren, dan stijgt het percentage onvoldoendes bij wiskunde van 23% naar 29%, bij Engels van 27% naar 35%, bij biologie van 15% naar 24%, bij Nederlands van 20% naar 31%, bij geschiedenis van 23% naar 53% en bij tehatex van 22% naar 54%.



Figuur 3 Percentage onvoldoendes voor de eerste en derde correctie per examen

2.2 De invloed van de tweede correctie

In het O2C-onderzoek zijn de scores van de eerste correctoren vergeleken met de gezamenlijk tussen eerste en tweede corrector overeengekomen scores (zie Figuur 4). De gezamenlijke scores blijken niet noemenswaard af te wijken van die van de eerste correctoren. Afhankelijk van het examen varieert het gemiddelde verschil van een tiende tot een derde scorepunt. Ook voor de rangordening van de kandidaten maakt het zeer weinig uit of dat gebeurt op basis van de eerste dan wel de gezamenlijke correctie. Dat de verschillen tussen eerste en gezamenlijke correctie gering zijn, komt ook tot uiting in het percentage onvoldoendes. Bij Engels en wiskunde stijgt het percentage onvoldoendes ten gevolge van de tweede correctie met drie percentagepunten en bij de overige examens is het verschil nul of hooguit een percentagepunt. Al met al kunnen we niet anders dan concluderen dat de eerste en gezamenlijke correctie in statistisch opzicht niet of nauwelijks onderscheidbaar zijn. Kennelijk heeft de tweede correctie weinig directe invloed op de eerste correctie.



Figuur 4 Gemiddelde examenscore voor de eerste en gezamenlijke correctie per examen

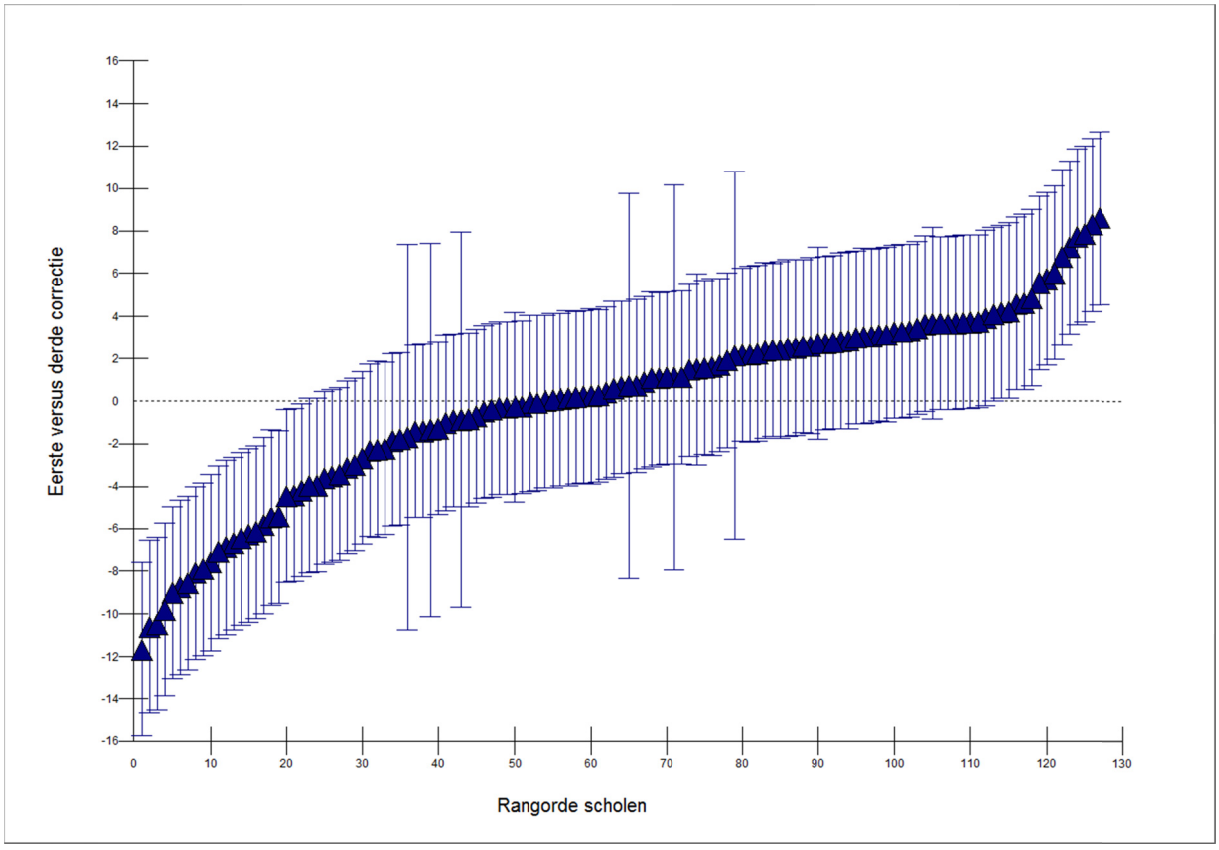
2.3 Verschillen tussen scholen

Hiervoor constateerden we dat de eerste correctoren gemiddeld aanzienlijk lagere scores toekennen dan de derde correctoren. Een overwegend te soepele beoordeling is niet zo erg als alle docenten even soepel zouden beoordelen en kandidaten niet bevoor- of benadeeld worden. Dit roept de vraag op in hoeverre leerlingen op de ene school meer profiteren van de geconstateerde soepelheid van de eigen docent dan op de andere school. De dataset van het O2C-onderzoek bevat de examenscores van de eerste, gezamenlijke en derde correctie van in totaal 803 examenkandidaten voor zes examens van in totaal 127 scholen. De gemiddelden per school zijn gebaseerd op de oordelen van vijf derde correctoren. Een probleem is dat examens verschillende maximumscores hebben en de scores daardoor moeilijk vergelijkbaar zijn. Daarom zijn de scores per vak omgezet naar het gemiddelde percentage goed in de volledige dataset. Het gemiddelde percentage goed beantwoorde vragen bedraagt voor de eerste correctie 59.6%, voor de gezamenlijke correctie 59.3% en voor de derde correctie 53.7%. Eenvoudig rekenwerk laat zien dat het gemiddelde verschil tussen de eerste en gezamenlijk correctie zeer klein is en slechts .3% van de maximumscore bedraagt. Het overeenkomstige verschil tussen de eerste en derde correctie is daarentegen aanzienlijk en bedraagt 6.4% van de maximumscore. Nagegaan is in hoeverre de verschillen tussen de drie vormen van

correctie op de ene school groter zijn dan op de andere school. De analyse is uitgevoerd met behulp van meerniveau analyse volgens het zogeheten multivariate model voor afwijkingsscores (Van den Bergh & Kuhlemeier, 1997). De belangrijkste resultaten zijn gevisualiseerd in Figuur 5 en 6.

Worden kandidaten op de ene school soepeler beoordeeld dan op de andere school?

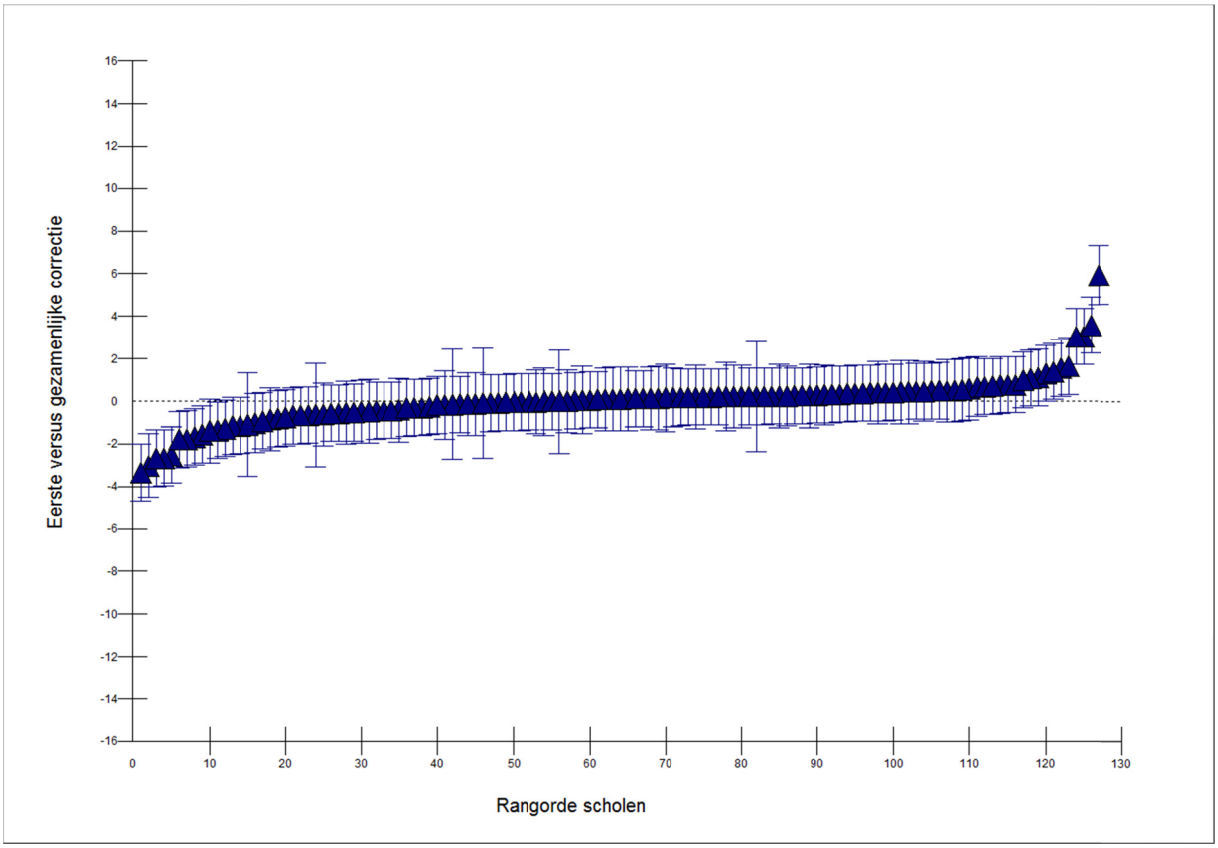
Eerder constateerden we dat de 'eigen' docent gemiddeld 6.4% hogere examenscores toekent dan de 'onafhankelijke' derde correctoren. Figuur 5 laat zien in hoeverre dit verschil op de ene school groter of kleiner is dan op de andere school. Daarbij zijn de scholen geordend van soepel naar minder soepel (met de derde correctie als vergelijkingscriterium). De horizontale stippellijn representeert het gemiddelde verschil tussen de eerste en derde correctie dat gelijk is aan nul. De driehoekjes geven het gemiddelde van de school weer in vergelijking met dat van de totale groep van 127 scholen. De verticale lijnen tonen het 95%-betrouwbaarheidsinterval rond de schoolgemiddelden. Een school verschilt significant van het gemiddelde (op 5%-niveau) als het betrouwbaarheidsinterval geen overlap heeft met nul (Goldstein & Healy, 1995).
De betrouwbaarheidsintervallen rond de schoolgemiddelden raken de nullijn in veel gevallen niet. Links onder in Figuur 5 bevindt zich een grote groep scholen waar het gemiddeld verschil tussen de eerste en derde correctie significant groter is dan 6.4% van de maximumscore (en waar de eerste corrector dus significant soepeler beoordeelde dan het gemiddelde). Rechts boven in de figuur zien we een grote groep scholen waar het gemiddeld verschil tussen de eerste en derde correctie significant kleiner is dan 6.4% (en waar de eigen docent dus significant minder soepel beoordeelde dan gemiddeld). Al met al moeten we concluderen dat kandidaten op de ene school sterker profiteren van de soepelheid van de eigen docent dan op de andere school. Deze resultaten staan op gespannen voet met het uitgangspunt dat kandidaten met dezelfde vaardigheid onafhankelijk van de school dezelfde slaagkans zouden moeten hebben.



Figuur 5 Verschillen tussen scholen in het verschil tussen de eerste en derde correctie

Heeft de tweede correctie op de ene school meer invloed dan op de andere school?

Hiervoor constateerden we dat het gemiddelde van de gezamenlijke correctie .3% lager is dan dat van de eerste correctie. Figuur 6 laat zien in hoeverre dit gemiddeld verschil op de ene school groter is dan op de andere school. Daarbij zijn de scholen geordend van soepel naar minder soepel (met de gezamenlijke correctie als vergelijkingscriterium). De horizontale stippellijn representeert nu het gemiddelde verschil tussen de eerste en gezamenlijke correctie dat gelijk is aan nul. De driehoekjes geven het gemiddelde van de school weer in vergelijking met dat van de totale groep van 127 scholen. De betrouwbaarheidsintervallen van de schoolgemiddelden in Figuur 6 raken de nullijn vrijwel altijd. Het verschil tussen het gemiddelde van de eerste en gezamenlijke correctie is dus voor nagenoeg alle scholen gelijk. Toch is er een kleine groep scholen die afwijkt van het gemiddelde. Helemaal links bevinden zich enkele scholen waar de eerste corrector in vergelijking met de tweede corrector significant soepeler was (dan het gemiddelde verschil van .3%) en helemaal rechts zien we enkele scholen waar de eerste corrector strenger was.



Figuur 6 Verschillen tussen scholen in het verschil tussen de eerste en gezamenlijke correctie

3. Het Panelonderzoek Vierde Correctie (P4C)

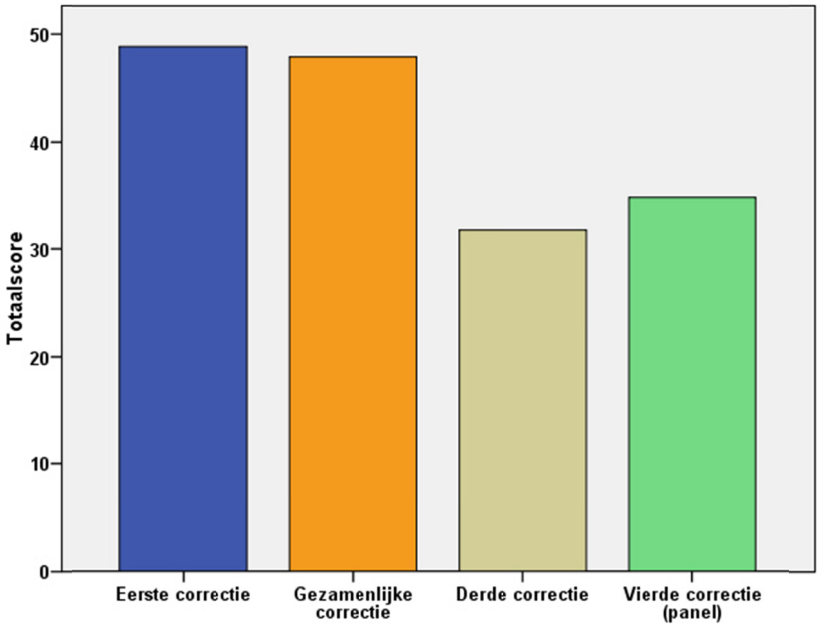
3.1 Check op de resultaten van het O2C-onderzoek

De resultaten van het O2C-onderzoek zijn besproken in een breed samengestelde resonansgroep. Die bestond uit vertegenwoordigers van het ministerie van OCW, werkgevers, werknemers, vakverenigingen, Inspectie, College voor Examens (CvE) en Cito. Binnen deze begeleidingscommissie werd over mogelijke verklaringen van de geconstateerde verschillen tussen de eerste en derde correctie verschillend gedacht. Een belangrijke vraag was in hoeverre de eerste correctoren te soepel hadden beoordeeld dan wel dat de derde correctoren te streng waren geweest. De resonansgroep concludeerde dat de opzet van het onderzoek geen eenduidig antwoord op deze vraag toeliet. De argumentatie was dat het via overleg tot stand gekomen oordeel van de eerste en tweede corrector evenveel gewicht in de schaal legt als het gemiddelde oordeel van de derde correctoren. Alvorens de resultaten te publiceren, wilde de resonansgroep meer zekerheid dat de conclusies juist waren. Er was met andere woorden behoefte aan een hard criterium waartegen de scores van de eerste en derde correctoren konden worden afgezet. Daarom werd besloten een vervolgonderzoek uit te voeren (Kuhlemeier, Gitsels, Boom, Van de Kerkhof & Sinkeldam, 2012). In het zogeheten Panelonderzoek Vierde Correctie (P4C) hebben panels van getrainde ‘vierde’ correctoren een selectie van vragen en examenwerken uit de examens geschiedenis, tehatex en Nederlands nogmaals nagekeken en vergeleken met de eerste, gezamenlijke en derde correcties uit het hoofdonderzoek. Anders dan in het O2C-onderzoek ontvingen de panelleden een training in het gebruik van het correctievoorschrift. Bovendien kregen zij volop de ruimte om met elkaar te discussiëren over de vraag, het beoordelingsmodel en de toegekende scores.

Het panelonderzoek had tot doel een antwoord te geven op de vraag: ‘In hoeverre hebben (sommige) eerste correctoren inderdaad overwegend (te) soepel nagekeken (en heeft de tweede correctie ten onrechte geen directe invloed op de uiteindelijke score)?’ Anders gezegd gaat het erom in hoeverre de eerste corrector het correctievoorschrift bij deze kandidaten correct heeft toegepast dan wel dat eventuele beoordelingsfouten bij de derde correctoren liggen. De panels geschiedenis, tehatex en Nederlands hebben ieder vijftien examenwerken opnieuw beoordeeld. Deze vijftien werken vormden geen willekeurige steekproef uit alle examenwerken van het hoofdonderzoek. Gekozen zijn de vijftien werken waarbij het verschil tussen de eerste en derde correctie het grootst was. De achterliggende gedachte was dat juist bij deze werken de kans op het vinden van te soepele of te strenge beoordelingen het grootst is. Naderhand zijn de scores van de eerste en derde correctoren vergeleken met die van het panel van vierde correctoren. Zo probeerden we erachter te komen in hoeverre de eerste correctoren te soepel hadden beoordeeld dan wel dat de derde correctoren te streng waren geweest.

Check op de resultaten geschiedenis

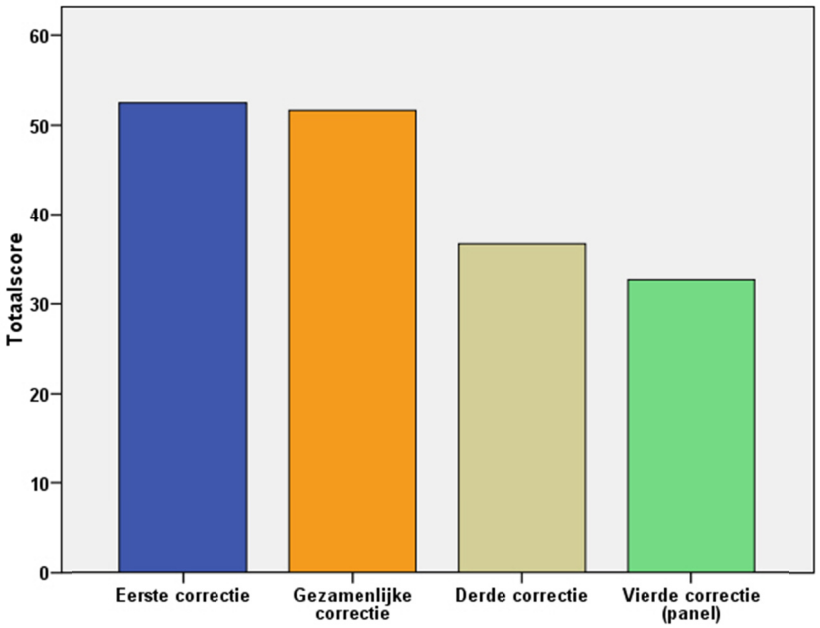
Bij geschiedenis blijkt de gemiddelde score van het panel veel dichter bij de derde dan de eerste correctie te liggen (zie Figuur 7). Kennelijk zijn de eerste (en tweede) correctoren geschiedenis te soepel geweest en waren de derde correctoren slechts in beperkte mate te streng. Uitgedrukt in cijfers op het CSE geschiedenis komt het verschil tussen de eerste, derde en vierde correctie overeen met de cijfers 7.1, 5.1 en 5.4.



Figuur 7 Gemiddelde totaalscores van de eerste, gezamenlijke, derde en vierde correctie geschiedenis

Check op de resultaten tehatex

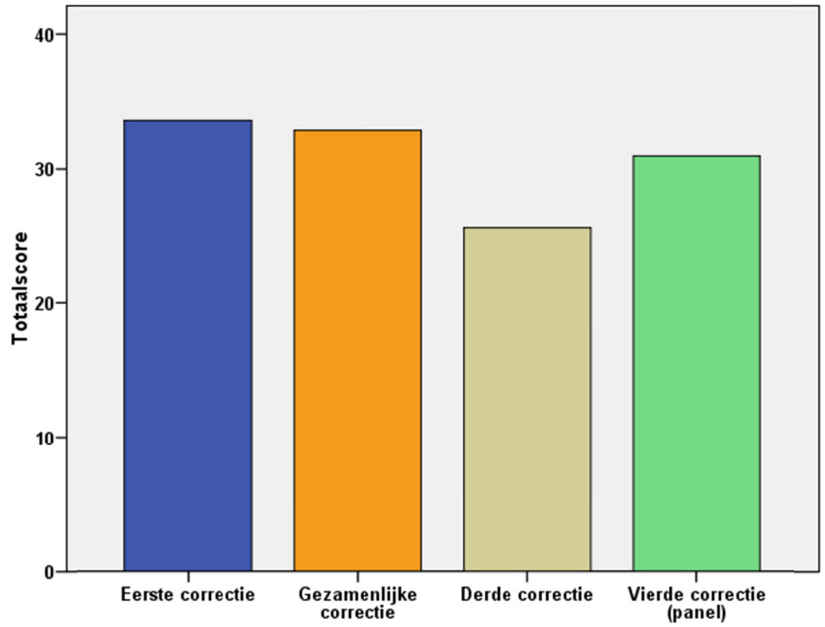
Bij tehatex blijkt de gemiddelde score van het panel nog onder het gemiddelde van de derde correctoren te liggen (zie Figuur 8). Kennelijk hebben de eerste (en tweede) correctoren te soepel beoordeeld en waren de derde correctoren niet onredelijk streng. Uitgedrukt in cijfers op het CSE tehatex komt het verschil tussen de eerste, derde en vierde correctie overeen met de cijfers 7.6, 5.7 en 5.3. In een ander vervolgonderzoek hebben vijf correctoren ongeveer de helft van de examenwerken uit het O2C-onderzoek nogmaals beoordeeld (Gitsels & Kuhlemeier, in voorbereiding). De vijf correctoren bleken nog strenger te hebben beoordeeld dan de derde correctoren uit het O2C-onderzoek.



Figuur 8 Gemiddeld totaalscores van de eerste, gezamenlijke, derde en vierde correctie tehatex

Check op de resultaten Nederlands

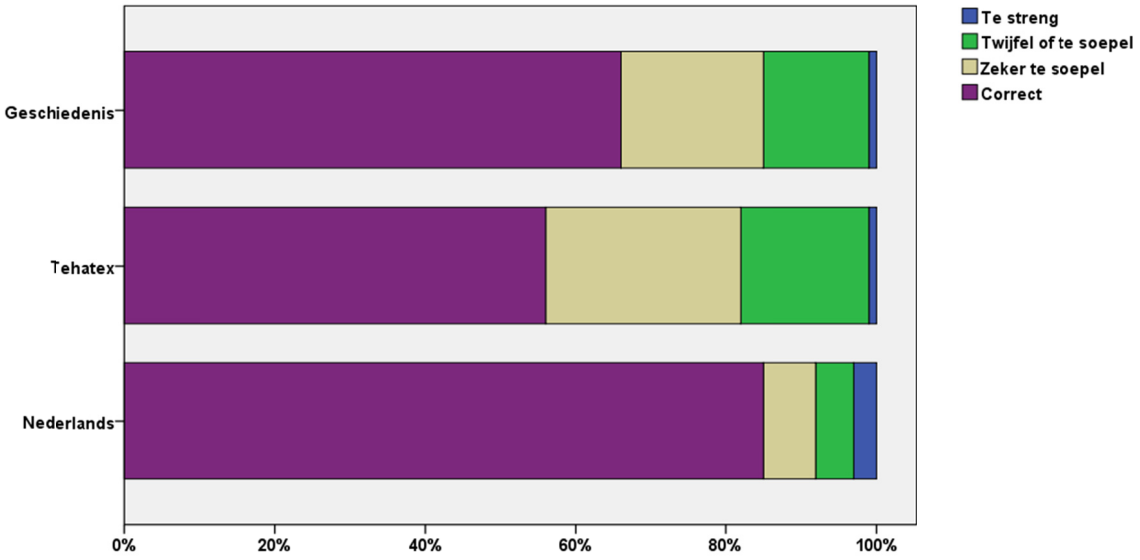
De bevindingen van het panel Nederlands wijken af van die bij geschiedenis en tehatex (zie Figuur 9). Bij Nederlands ligt de gemiddelde score van het panel dichter bij de eerste dan de derde correctie. De eerste correctoren Nederlands zijn dus niet onredelijk soepel geweest is en de derde correctoren waren te streng. Uitgedrukt in cijfers op het CSE Nederlands komt het verschil tussen de eerste, derde en vierde correctie overeen met de cijfers 7.1, 5.7 en 6.7. In de paneldiscussies speelde het verlag van de examenbespreking uit Levende Talen een belangrijke rol. De panelleden vonden dat verslag een nuttig hulpmiddel bij de beoordeling. De vergelijking van de verschillen tussen de scores van de eerste beoordeling vóór discussie en de tweede beoordeling na discussie laat zien dat het gebruik van de examenbesprekingen over het algemeen tot hogere scores leidt. Het veelvuldig gebruik van de examenbesprekingen verklaart wellicht mede waarom het verschil tussen de eerste en derde correctie bij Nederlands kleiner is dan bij geschiedenis en tehatex.



Figuur 9 Gemiddeld totaalscores van de eerste, gezamenlijke, derde en vierde correctie Nederlands

Check op de toepassing van het correctievoorschrift

De panelleden hebben ook beoordeeld in hoeverre de eerste corrector het correctievoorschrift bij deze vijftien examenwerken juist heeft toegepast. Voor elke beoordeelde vraag maakten zij een keuze uit de volgende vier antwoordmogelijkheden: a) de eerste corrector heeft het correctievoorschrift correct toegepast, b) er is twijfel of de eerste corrector te soepel beoordeelde, c) de eerste corrector heeft met zekerheid te soepel beoordeeld en d) de eerste corrector heeft te streng beoordeeld. De resultaten zijn weergegeven in Figuur 10.



Figuur 10 De mate waarin eerste correctoren het correctievoorschrift correct toepassen per examen

Naar het oordeel van het panel geschiedenis heeft de eerste corrector het correctievoorschrift bij 66% van de nagekeken antwoorden juist toegepast. Bij 14% van de antwoorden staan de vierde correctoren in dubio of de eerste corrector de kandidaat te soepel beoordeelde en de kandidaat benadeeld is. Bij 19% van de antwoorden is het panel van mening dat de eerste corrector de kandidaat op een voor hen onverklaarbare wijze bevoordeeld heeft. De eerste correctoren beoordeelden hier bijvoorbeeld volgens het principe 'alles wat niet echt fout is, is goed' of rekenden evident foute antwoord goed. Daar staat tegenover dat de kandidaat bij 1% van de beoordelingen benadeeld is.
Naar het oordeel van het panel tehatex heeft de eerste corrector het correctievoorschrift bij 56% van de nagekeken antwoorden correct toegepast, bij 17% bestond twijfel of de kandidaat te soepel beoordeeld was, 26% was met zekerheid te soepel beoordeeld en 1% was te streng beoordeeld. Voor Nederlands bedragen de overeenkomstige percentages 85% (correct), 5% (mogelijk te soepel), 7% (zeker te soepel) en 3% (te streng).

De panels geschiedenis, tehatex en Nederlands concluderen verder dat er eerste correctoren zijn die een aanzienlijk deel van de antwoorden onverklaarbaar toegeeflijk beoordelen. Dit wil zeggen dat de soepelheid niet kan worden toegeschreven aan een gebrek aan vakkennis, onvolkomenheden in het examen, de wijze waarop de kandidaat het antwoord geformuleerd heeft en dergelijke. Bij Nederlands doet de onverklaarbaar soepele beoordeling zich voornamelijk voor bij het toekennen van aftrekpunten vanwege incorrecte formuleringen en onjuist taalgebruik in de samenvattingsopdracht. Volgens de panels moet hier welhaast sprake zijn van opportunistisch-strategisch beoordelingsgedrag en een gebrek aan professionaliteit en/of integriteit.

Al met al bevestigt het P4C-onderzoek dat de eerste correctoren geschiedenis en tehatex de 'eigen' leerlingen te soepel beoordelen en dat de derde correctoren niet onredelijk streng zijn geweest. Bij Nederlands waren de eerste correctoren volgens het panel doorgaans niet te soepel. Wel zijn er ook bij Nederlands eerste correctoren die opportunistisch-strategisch beoordelingsgedrag vertonen. De beperkte omvang van het beoordelaarsonderzoek laat het echter niet toe om de grootte van deze groep precies te bepalen.

3.2 Examenkenmerken en verschillen in soepelheid

Aan verschillen tussen correctoren kunnen verschillende oorzaken ten grondslag liggen. Een daarvan is gelegen in het examen zelf. Docenten moeten hun leerlingen beoordelen volgens het correctievoorschrift. Dat bestaat onder meer uit algemene en vakspecifieke regels voor de beoordeling en een beoordelingsmodel. In het P4C-onderzoek is ook nagegaan in hoeverre de verschillen tussen correctoren samenhangen met kenmerken van de vraag en het beoordelingsmodel. Daartoe hebben de drie panels een selectie van vijf vragen uit hun examen nogmaals nagekeken. Geselecteerd zijn de vijf vragen waarbij het gemiddelde verschil tussen de scores van de eerste minus de derde correctie het sterkst positief is. De vierde correctoren hebben van deze vijf vragen niet alle beschikbare antwoorden nogmaals nagekeken. Per vraag is volstaan met een selectie van de antwoorden van twaalf kandidaten. Geselecteerd zijn die antwoorden waarbij het verschil tussen de eerste en derde correcties gemiddeld het grootste is. De veronderstelling is dat de kans op het vinden van aanwijzingen voor verbetering van het examen bij deze antwoorden het grootst is. Hieronder bespreken we eerste twee examenvragen die zeer lastig te beoordelen bleken. Daarna gaan we in op kenmerken van het beoordelingsmodel die aanleiding gaven tot grote verschillen tussen correctoren. Voor een uitgebreidere bespreking dan in het bestek van deze publicatie mogelijk is, wordt verwezen naar Kuhlemeier, Gitsels, Boom, Van de Kerkhof en Sinkeldam (2012).

Eerste voorbeeld van een lastige examenvraag

Bij het beoordelen van de antwoorden op open vragen zijn verschillen tussen correctoren niet volledig te vermijden. Hoe lastig de beoordeling van open vragen kan zijn, illustreren we aan de hand van een voorbeeld uit het examen geschiedenis havo 2009. De gepresenteerde antwoorden van kandidaten zijn 'echte' antwoorden waarbij de oorspronkelijke formulering en spelling intact gelaten zijn.

Vraag 20 uit het examen geschiedenis havo 2009 is een zogeheten noemvraag. De kandidaat wordt gevraagd drie voordelen te noemen van de Spaanse verovering van Antwerpen voor de Hollandse nijverheid. De inleiding op de vraag vermeldt dat de Hollandse nijverheid en handel profiteerden van de verovering van Antwerpen door het Spaanse leger. De precieze formulering is als volgt:

De Hollandse nijverheid en handel profiteerden van de verovering van Antwerpen door het Spaanse leger.

3p **20** Noem drie voordelen voor de Hollandse nijverheid van de Spaanse verovering van Antwerpen.

Het beoordelingsmodel bij vraag 20 is van het type 'Een voorbeeld van een juist antwoord is'. De corrector kan 0, 1, 2 of 3 punten toekennen. Het beoordelingsmodel geeft vier voorbeelden van juiste voordelen waarbij de kandidaat per juist voordeel één punt krijgt met een maximum van 3 punten. Het beoordelingsmodel is hieronder onverkort weergegeven:

20 maximumscore 3

Voorbeeld van een juist antwoord is (drie van de volgende):
Door de verovering van Antwerpen door de Spanjaarden:

- kwamen geschoolde arbeidskrachten uit Antwerpen/de Zuidelijke Nederlanden naar Holland.
- verdween de Antwerpse nijverheid als concurrent.
- konden de opstandige gewesten de Schelde afsluiten/de toegang tot Antwerpen blokkeren.
- vestigden zuidelijke immigranten zich in Holland waar zij met hun kennis/bedrijfskapitaal de nijverheid versterkten.

per juist voordeel 1

Overeenkomstig de vraagstelling moet de kandidaat drie voordelen opnoemen van de Spaanse verovering van Antwerpen voor de Hollandse nijverheid. Antwoorden die niet specifiek over nijverheid gaan, moeten volgens de maker van het examen fout gerekend worden. Kandidaten blijken bij deze vraag regelmatig voordelen voor de handel te noemen in plaats van voor de nijverheid (zie onderstaande voorbeelden van 'echte' antwoorden van kandidaten die deelnamen aan het examen geschiedenis havo 2009).

De Hollandse nijverheid profiteerde van de Spaanse verovering van Antwerpen omdat ze nu minder concurrentie hadden, de handelspositie van Antwerpen nu naar Holland verschoof en dus beter werd en omdat veel mensen naar Holland trokken waardoor het dichter bevolkt werd en door die redenen dus het middelpunt van de handel werd.

Drie voordelen voor de Hollandse nijverheid van de Spaanse verovering van Antwerpen zijn:
De val van Antwerpen. Hierdoor kwam de handel naar Amsterdam
Rijke kooplieden-regenten vertrokken uit Antwerpen naar Holland en namen daarbij hun kapitaal en kennis mee
De Schelde werd afgesloten waardoor er geen schepen meer naar Antwerpen konden gaan en de handel nog meer in Holland gevestigd werd

De kooplieden en handelaren uit Antwerpen kwamen naar Holland, daardoor nam de kennis toe in Holland.
Door de afsluiting van de Schelde vestigde zich in Holland een stapelmarkt
Doordat de Schelde was afgesloten en hier meer handel kwam, namen de arbeidsplaatsen toe. Dat was een voordeel.

De Antwerpse haven was niet veilig dus gingen ze naar Hollandse havens voor nijverheid producten.
Door de verovering van Antwerpen vluchtte veel mensen naar Holland. Deze mensen gingen verder hun werk doen in Holland dus er was een overvloed aan werknemers.
Er waren ook mensen die vluchtte, omdat ze heel rijk waren en veel kennis hadden deze hielpen mee aan de financiering en vernieuwing van de nijverheid.

Strengere correctoren kennen aan voordelen voor de handel terecht geen punten toe, terwijl hun soepelere collega's dat wel doen. Laatstgenoemden beargumenteren de toegekende punten door erop te wijzen dat nijverheid en handel sterk met elkaar verbonden zijn en dat voordelen voor de nijverheid ook ten goede komen aan de handel. Overigens had dit beoordelingsprobleem wellicht verminderd kunnen worden door het woordje handel uit de inleiding op de vraag te schrappen en aan het beoordelingsmodel een opmerking toe te voegen dat voordelen voor de handel fout gerekend moeten worden.

Tweede voorbeeld van een lastige examenvraag

Het tweede voorbeeld betreft vraag 1 uit het examen geschiedenis havo uit 2009. De inleiding op deze vraag vermeldt dat de Franse regering in de Coalitieoorlogen de dienstplicht invoerde. De vraag is een zogeheten uitlegvraag en luidt "Leg uit dat zij hiermee de betrokkenheid van de Franse burgers bij de staat kon vergroten".

De maximumscore bij deze vraag is 2 punten. Het beoordelingsmodel beschrijft een voorbeeld van een juist antwoord dat uit twee varianten bestaat. Volgens de eerste variant is de juiste uitleg van het gegeven dat de Franse regering door de dienstplicht in te voeren de betrokkenheid van de burgers bij de staat kon vergroten dat een groot aantal (jonge) mannen meer in aanraking kwam met de idealen van de Franse revolutie of het Franse nationalisme. Bij de tweede variant is de juiste uitleg dat een groot aantal (jonge) mannen meer onder invloed van politieke commissarissen kwam. Deze tweede

variant is op verzoek van het CvE aan het beoordelingsmodel toegevoegd. Het beoordelingsmodel is hieronder integraal opgenomen.

1	**maximumscore 2** Voorbeeld van een juist antwoord is: Door de invoering van de dienstplicht kwam een groot aantal (jonge) mannen meer in aanraking met de idealen van de Revolutie/het Franse nationalisme / onder invloed van politieke commissarissen.

Hieronder staan vijf voorbeelden van 'echte' antwoorden van kandidaten die in de praktijk aanleiding gaven tot grote scoreverschillen tussen eerste correctoren.

De dienstplicht houdt in dat alle mannen tussen de 18 en 25 jaar (of ouder) in het leger moesten dienen. Doordat deze mannen het leger in moesten kregen ze meer mee van de oorlogen in de tijd & leerden ze vechten voor hun vaderland. De betrokkenheid van de burgers werd vergroot doordat mannen hun vrouwen en kinderen en ouders achter lieten. Deze leefden met de oorlog mee omdat een geliefde van hen het leger in ging en oorlog ging voeren.

De dienstplicht was ingesteld voor mannen van 18-25 jaar. De betrokkenheid van de franse burgers wordt hierdoor vergroot doordat de burgers nu zelf ten strijd moeten gaan

Door de dienstplicht was iedereen vanaf een bepaalde leeftijd verplicht mee te vechten in de oorlog. Hierbij werden dus de burgers betrokken bij de beslissingen die de staat nam (bij deze dus een oorlog). Bijna alle mannen van een bepaalde leeftijd (meestal vanaf 18) werden opgeroepen dat ze moesten vechten en wel elke vrouw had een vader, broer, man of zoon die ging vechten.

De burger werd betrokkener bij de staat omdat er de kans was dat die in het leger moest en dus afhankelijk was van de beslissingen van de staat op bv. militair gebied. Ook kende iedereen wel iemand die in het leger zat of zou gaan. Ze trokken het lot van die mensen aan en gingen dus ook de beslissingen van de staat volgen, hierdoor werden de burgers meer betrokken tot de staat.

Vrijwel alle mannen in Frankrijk die oud genoeg waren gingen in dienstplicht. Hierdoor was een enorm deel van de bevolking direct bij oorlogen betrokken.

Waarom zijn deze vijf antwoorden zo verschillende beoordeeld? Een eerste reden is dat de antwoorden op geen enkele wijze zijn terug te vinden in het beoordelingsmodel. Zo verwijst geen van de kandidaten in zijn of haar uitleg naar de idealen van de Franse revolutie, het Franse nationalisme of de invloed van politieke commissarissen die in het beoordelingsmodel genoemd worden. Voor zover kandidaten andere historische verklaringen geven, zijn deze meestal niet (door docenten) terug te vinden in gezaghebbende wetenschappelijke publicaties.

Een tweede reden heeft te maken met de aard van het beoordelingsmodel. Het bestaat uit een beoordelingsschaaltje met de toegestane scores 0, 1 en 2. Het beoordelingsmodel geeft alleen een voorbeeld van een volledig juist antwoord. Omdat alleen de maximumscore aan de hand van een voorbeeld omschreven is, biedt het beoordelingsmodel de corrector weinig houvast bij het toekennen van 1 of 0 punten. Het herkennen en waarderen van half en geheel foute antwoorden wordt met andere woorden aan de vakinhoudelijke deskundigheid van de corrector overgelaten.

Een derde reden is dat kandidaten vaak zuiver psychologische verklaringen geven die op zich niet onlogisch zijn, maar die niet door gezaghebbende wetenschappelijke publicaties gestaafd worden (en

dus fout gerekend zouden moeten worden). Zo verwijzen veel kandidaten ernaar dat de Franse overheid de burgers bij de staat wist te betrekken doordat iedere soldaat wel een familielid heeft dat met hem meeleeft. Verschillen in soepelheid ontstaan waar strenge correctoren een verwijzing naar termen uit het beoordelingsmodel of de vakliteratuur eisen en soepele correctoren genoegen nemen met een niet-historisch antwoord.

Examenkenmerken en verschillen in soepelheid

De panels hebben ieder zestig antwoorden op vijf vragen uit het desbetreffende examen opnieuw nagekeken en vervolgens besproken. De bespreking was gericht op het vinden van verklaringen voor de geconstateerde verschillen tussen correctoren in soepelheid. Verschillen in soepelheid blijken vooral voor te komen in de volgende situaties:

- Het beoordelingsmodel is onvolledig en kandidaten geven veelvuldig antwoorden die niet in het beoordelingsmodel voorkomen. De corrector is dan aangewezen op algemene regel 3.3 (d.w.z.: indien een antwoord op een open vraag niet in het beoordelingsmodel voorkomt en dit antwoord op grond van aantoonbare, vakinhoudelijke argumenten als juist of gedeeltelijk juist aangemerkt kan worden, moeten scorepunten worden toegekend naar analogie of in de geest van het beoordelingsmodel) en de vakspecifieke regel dat vakinhoudelijke argumenten afkomstig moeten zijn uit gezaghebbende, wetenschappelijke publicaties. Het onderzoek maakt aannemelijk dat de scores van de correctoren verder uiteenlopen naarmate zij vaker een beroep moeten doen op algemene regel 3.3 en de vakspecifieke regel.
- Het beoordelingsmodel geeft alleen een voorbeeld van een volledig juist antwoord en biedt geen steun bij het herkennen van en het toekennen van punten aan minder goede of foute antwoorden. Daardoor moet de corrector een beroep doen op algemene regel 3.2 (d.w.z. de regel voor het zelfstandig toekennen van scorepunten aan gedeeltelijk juiste antwoorden). Uit de panelbeoordelingen komt naar voren dat algemene regel 3.2 vaak aanleiding geeft tot verschillen tussen correctoren in soepelheid.
- Een inconsistentie tussen vraag en beoordelingsmodel geeft aanleiding tot verschillen tussen correctoren in soepelheid. Bijvoorbeeld:
    a. de examenvraag gaat over burgers terwijl in het beoordelingsmodel mannen i.c. dienstplichten centraal staan;
    b. Er wordt een kort en bondig antwoord gevraagd, maar het beoordelingsmodel bestaat uit lange en complexe zinnen.
- De kandidaat geeft meer antwoorden dan volgens de vraagstelling is toegestaan. De corrector moet dan algemene regel 3.5 toepassen (d.w.z. de regel waarbij antwoorden boven het gevraagde aantal niet in de beoordeling betrokken mogen worden). De ene corrector past deze algemene regel zoals het hoort wel toe en de ander doet dat ten onrechte niet en geeft de kandidaat ook punten voor antwoorden boven het toegestane aantal.
- Het aantal te beoordelen inhoudselementen is groter dan de maximumscore. Om één punt te verdienen moet de kandidaat bijvoorbeeld zowel een juiste uitleg geven als een juiste bron noemen. De ene corrector kent zoals het hoort 0 punten toe, maar de andere corrector kent dat éne punt ook toe als de kandidaat alleen een juiste uitleg geeft of alleen een juiste bron noemt.
- De kandidaat geeft een foute toelichting bij een goed antwoord of een goede toelichting bij een fout antwoord zonder dat in de vraagstelling om verduidelijking of uitleg gevraagd wordt. Verschillen in soepelheid ontstaan als de ene corrector de toelichting wel in de beoordeling betrekt en de ander dat niet doet.
- Het voorschrift van een als 0-1 te scoren vraag beschrijft meerdere kenmerken van een goed antwoord (element) zonder dat volledig duidelijk is hoeveel en welke kenmerken in het antwoord van de kandidaat aanwezig moeten zijn om dat ene punt te mogen toekennen. De ene corrector is al blij met één van de kenmerken, terwijl de ander eist dat alle kenmerken in het antwoord aanwezig zijn.
- Het beoordelingsmodel is geformuleerd in academische vaktaal/jargon, terwijl kandidaten hun antwoorden in alledaags Nederlands formuleren. Een voorbeeld is de omschrijving van een juist

antwoord ‘Het is een enorme ‘blow-up’ van een (klein) gebruiksvoorwerp’ uit het correctievoorschrift tehatex. De ene corrector eist een expliciete verwijzing naar een blow-up of een enorme uitvergroting van een gebruiksvoorwerp, terwijl de ander genoegen neemt met ieder antwoord waarin iets van grootte naar voren komt.

- Een beoordelingsmodel laat alleen de scores 0 en 2 toe, waarbij gedeeltelijk juiste antwoorden de score 0 moeten krijgen. Verschillen tussen correctoren ontstaan als de ene corrector voor een gedeeltelijk juist antwoord 0 punten toekent en de ander dat antwoord in strijd met het beoordelingsmodel toch met één scorepunt honoreert.
- Bij lastig te beoordelen antwoorden baseert de ene corrector de puntentoekenning op het verslag van examenbesprekingen terwijl de andere corrector dat niet doet en zelfstandig de algemene scoringsregels probeert toe te passen. De ervaringen van het panel Nederlands laten zien dat het gebruik van de examenbesprekingen doorgaans tot hogere scores leidt.

Bij de interpretatie van deze resultaten moet men bedenken dat de onderzochte examens, kandidaten en examenvragen geen representatieve steekproef vormen uit de totale verzameling van examens, kandidaten en examenvragen. De drie examens uit het P4C-onderzoek zijn vooral gekozen vanwege het grote aantal open vragen. Hadden we bijvoorbeeld voor examens zonder open vragen gekozen, dan waren de resultaten ongetwijfeld gunstiger geweest (maar had het onderzoek ook minder verbeteringssuggesties opgeleverd). Daarnaast hebben de panels alleen vragen en examenwerken beoordeeld waarbij het verschil tussen de eerste minus de derde correctie sterk positief was. De kans op het aantreffen van problematische vragen is daardoor veel groter dan wanneer we de panels een representatieve steekproef van vragen hadden voorgelegd. De resultaten van het P4C-onderzoek zijn derhalve niet geldig voor het gemiddelde examen, de gemiddelde examenvraag of de gemiddelde corrector.

4. Het vragenlijstonderzoek naar de praktijk van de eerste en tweede correctie

In het schooljaar 2011-2012 heeft Cito een vragenlijstonderzoek uitgevoerd naar de praktijk van de eerste en tweede correctie. Van deze inventarisatie is verslag gedaan in Kuhlemeier en Kremers (2012). Het onderzoek had tot doel een antwoord te geven op de volgende vragen:

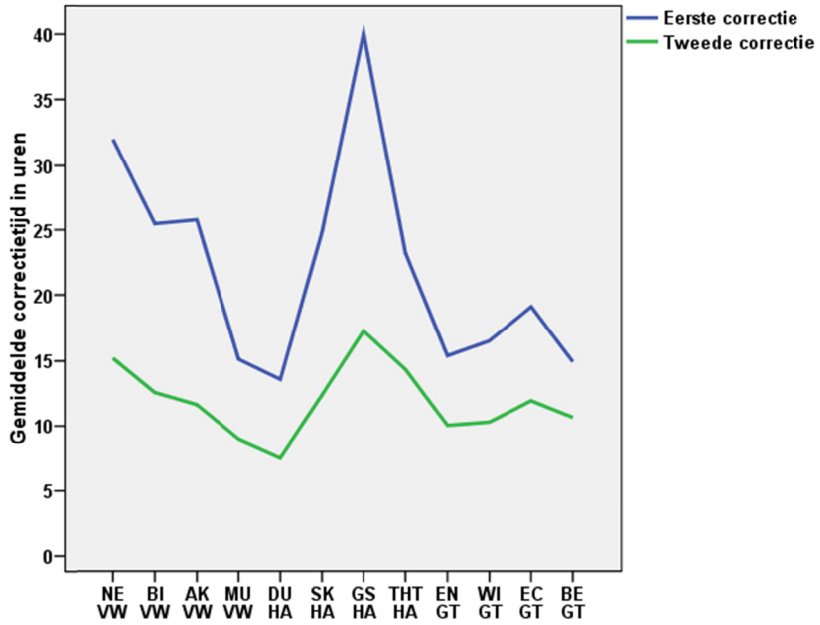
- Hoeveel tijd kost de eerste en tweede correctie?
- Onder welke omstandigheden worden de eerste en tweede correctie uitgevoerd?
- In hoeverre voeren docenten de tweede correctie integraal uit?
- Hoe denken eerste en tweede correctoren over de zorgvuldigheid en objectiviteit van elkaars beoordelingen?
- Hoe vindt het overleg tussen eerste en tweede corrector plaats en hoe komen de uiteindelijke scores tot stand?

In het onderzoek zijn twaalf examens betrokken, te weten Nederlands vwo, Biologie vwo, Aardrijkskunde vwo, Muziek vwo (regulier en cbt), Duits havo, Scheikunde havo, Geschiedenis havo, Tehatex havo, Engels gt , Wiskunde gt, Techniek gt en Beeldende vakken gt. Van de 6000 verzonden vragenlijsten kwamen er 3695 (62%) ingevuld retour. Alle respondenten hadden als eerste en/of tweede corrector aan het centraal schriftelijk examen deelgenomen. De relatief hoge respons betekent niet noodzakelijkerwijs dat de gegevens representatief zijn voor de examenpraktijk in Nederland. Zo zouden docenten die de correctie niet volgens de wettelijke voorschriften uitvoeren in de responsgroep ondervertegenwoordigd kunnen zijn.

4.1 Tijdbesteding aan de eerste en tweede correctie

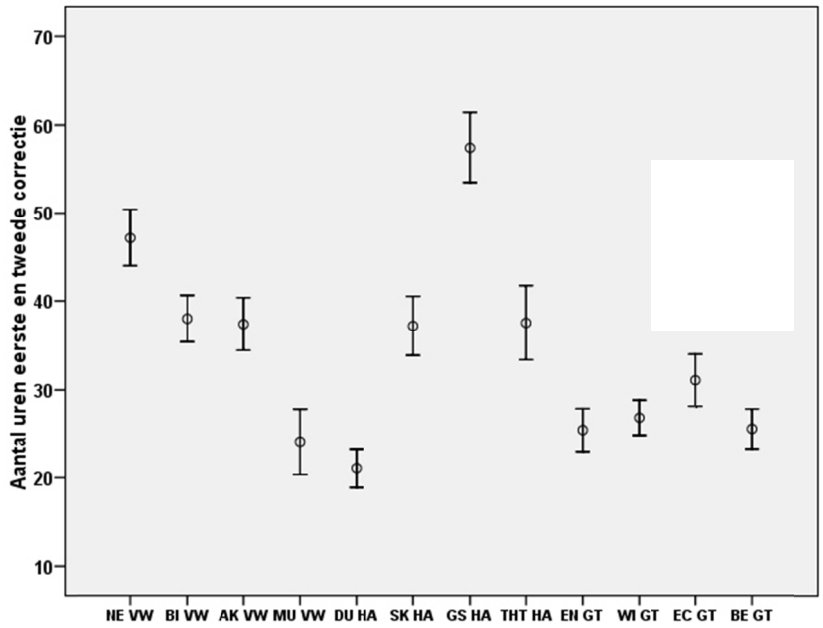
Het Platform VVVO (2008a,b) heeft erop gewezen dat de correctie voor veel docenten een hoge taakbelasting met zich meebrengt. In het vragenlijstonderzoek is getracht meer te weten te komen over de tijdbesteding en de randvoorwaarden waaronder de eerste en tweede correctie worden uitgevoerd.

Als eerste corrector kijkt de gemiddelde docent het werk van 38 kandidaten na en als tweede corrector beoordeelt hij of zij veertig kandidaten. Hoewel het aantal te beoordelen kandidaten ongeveer gelijk is, kost de eerste correctie docenten veel meer tijd dan de tweede correctie, respectievelijk 22 en 12 uur. Dit verschil zien we ook terug in het gemiddeld aantal dagen: vier dagen voor de eerste correctie en drie dagen voor de tweede correctie. Figuur 11 laat zien dat docenten veel meer tijd aan de eerste correctie besteden dan aan de tweede correctie.



Figuur 11 Tijdbesteding aan de eerste en tweede correctie per examen

Hoeveel tijd besteden docenten aan de eerste en tweede correctie samen? De gemiddelde docent corrigeert het werk van 79 examenkandidaten en besteedt daar 35 klokuren aan in zeven dagen. De correctielast is voor het ene examen veel hoger dan voor het andere examen. Dat valt op te maken uit Figuur 12. Daarin is voor elk examen het gemiddeld aantal uren voor de eerste en tweede correctie samen weergegeven. De verticale lijnen tonen het 95%-betrouwbaarheidsinterval rond het gemiddelde. Examens hebben een significant verschillende correctielast als de betrouwbaarheidsintervallen elkaar niet overlappen. Te zien is onder meer dat geschiedenisdocenten bijna drie keer zoveel tijd in de correctie steken als docenten Duits.



Figuur 12 Gemiddelde correctietijd voor de eerste en tweede correctie in uren per examen

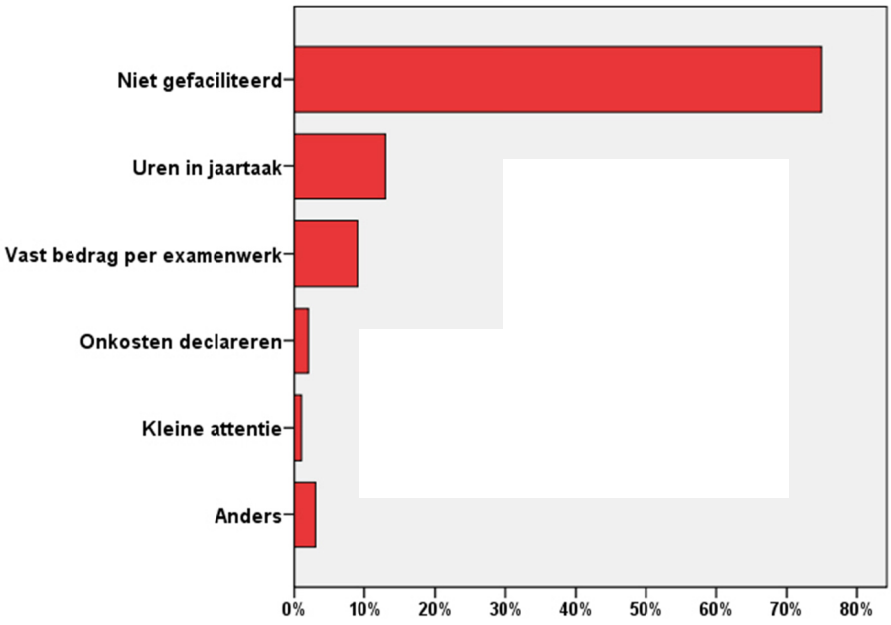
De correctietijd blijkt zeer ongelijk verdeeld over de docenten. De tien procent met de lichtste correctielast kijkt het werk van hooguit 33 kandidaten na en besteedt daar tot tien klokuren aan in hooguit vier dagen. De tien procent met de zwaarste correctielast corrigeert het werk van minimaal 140 kandidaten en besteedt daar ten minste zeventig uur aan in twaalf dagen of meer.

Docenten die lesgeven in de gemengde en theoretische leerweg besteden als groep minder tijd aan de correctie dan vwo- en havo-docenten (27 versus 40 en 38 uur). Gammadocenten steken meer tijd in de correctie dan bètadocenten (respectievelijk 42 versus 34 uur) die daar op hun beurt meer tijd aan besteden dan talendocenten (31 uur) en kunstdocenten (29 uur).

Bij de interpretatie wordt opgemerkt dat alleen gevraagd is naar de tijdbesteding aan het nakijken als zodanig. Docenten besteden daarnaast nog tijd aan andere zaken die met de correctie samenhangen, zoals intercollegiaal overleg over de toepassing van het correctievoorschrift, raadplegen van de examenverslagen van de vakverenigingen en de administratieve afhandeling. De hier gerapporteerde correctietijd vormt dan ook een onderschatting van de totale tijdbesteding aan de eerste en tweede correctie.

4.2 Randvoorwaarden van de eerste en tweede correctie

Wat doen scholen om docenten in staat te stellen de eerste en tweede correctie naar behoren uit te voeren? Figuur 13 laat zien hoe scholen de eerste correctie faciliteren. Van 13% van de docenten zijn de uren voor de eerste correctie opgenomen in de jaartaak (bijvoorbeeld in de opslagfactor), 8% krijgt een vast bedrag per examenwerk, 2% kan onkosten declareren (bijvoorbeeld telefoonkosten), 1% ontvangt een kleine attentie (bijv. een boekenbon) en bij 3% wordt de eerste correctie op een andere, niet nader omschreven manier gefaciliteerd.



Figuur 13 De wijzen waarop scholen de eerste correctie faciliteren

Scholen faciliteren de tweede correctie anders dan de eerste correctie. De verschillen zitten hem vooral in het jaartaakbeleid en het stukloon. De uren voor de tweede correctie zijn bij 5% van de docenten opgenomen in de jaartaak tegen 13% voor de eerste correctie. Voor de tweede correctie krijgt 39% van de docenten een vast bedrag per examenwerk tegen 9% voor de eerste correctie.

Hoe tevreden zijn docenten over het schoolbeleid? Driekwart is van mening dat de eerste correctie op school niet gefaciliteerd wordt en bij de tweede correctie gaat het om bijna de helft. Met het schoolbeleid ten aanzien van de eerste en tweede correctie is respectievelijk 39% en 49% ontevreden of zeer ontevreden.
Veel docenten ervaren de eerste en tweede correctie als een zware belasting. Zo ervaart 39% de eerste correctie als belastend of zeer belastend en voor de tweede correctie bedraagt dit percentage 56%.

Overigens is in de vragenlijst alleen gevraagd naar expliciete faciliteringsmaatregelen. Veel scholen kennen de stilzwijgende afspraak dat docenten de correctie uitvoeren in de tijd die vrijkomt doordat de ingeroosterde lessen in examenklassen bij aanvang van het CSE niet meer gegeven worden. In het onderzoek is niet nagegaan in hoeverre de vrijgekomen tijd toereikend is en daadwerkelijk aan de correctie besteed kan worden.

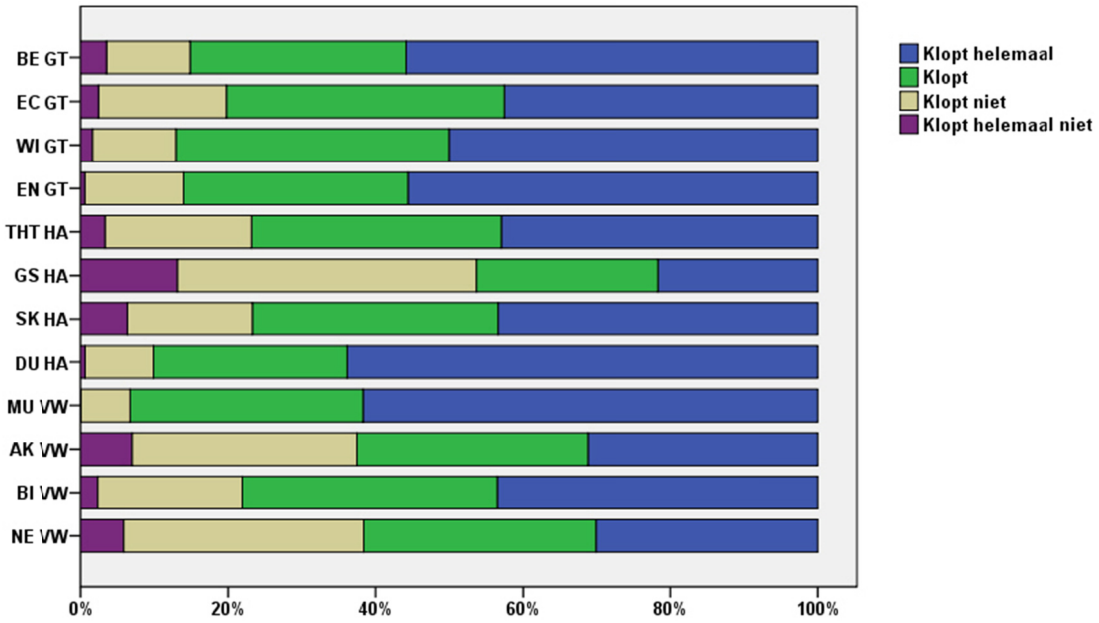
4.3 Volledigheid van de tweede correctie

Volgens het Examenbesluit moeten tweede correctoren het werk van de kandidaten integraal nakijken. Dit wil zeggen dat de tweede corrector alle werken nakijkt en per kandidaat het volledige examenwerk met alle vragen. Volgens opgave van de eerste correctoren komt het nauwelijks voor dat de tweede correctie achterwege blijft. Slechts van één procent van de kandidaten is het werk niet door een tweede corrector nagekeken. Het verhaal van de eerste corrector die een aantal pagina's dichtplakte met lijm en het examenwerk enkele dagen later in dezelfde staat weer terugkreeg, lijkt dus vooral een mooi verhaal.

Gevraagd naar de volledigheid waarmee de tweede correctie is uitgevoerd, zegt 9% van de eerste correctoren hier geen zicht op te hebben, 31% vindt dat de tweede correctie zeer volledig gedaan is, 39% volledig, 18% min of meer volledig, 2% onvolledig en 1% zeer onvolledig.

Aan de hand van vijf stellingen is geïnventariseerd welke strategieën docenten bij de tweede correctie hanteren. De eerste stelling beschrijft de situatie waarin de tweede corrector het examenwerk integraal nakijkt. De letterlijke formulering is 'Ik heb het examenwerk van alle kandidaten helemaal nagekeken (als ware het een eerste correctie)'. De docenten is gevraagd in hoeverre deze uitspraak op hen van toepassing is. Zij konden daarbij kiezen uit de antwoordmogelijkheden 'Klopt helemaal', 'Klopt', 'Klopt niet' en 'Klopt helemaal niet'. Van de tweede correctoren heeft 44% het examenwerk van alle kandidaten helemaal nagekeken, als ware het een eerste correctie. Van hen mag worden aangenomen dat zij de tweede correctie overeenkomstig het overheidsbeleid uitvoeren. Bijna één derde (32%) van de tweede correctoren koos voor de antwoordmogelijkheid 'Klopt', 20% voor 'Klopt niet' en 4% voor 'Klopt helemaal niet'. Zij voeren de tweede correctie niet helemaal of helemaal niet volgens de wettelijke voorschriften uit.

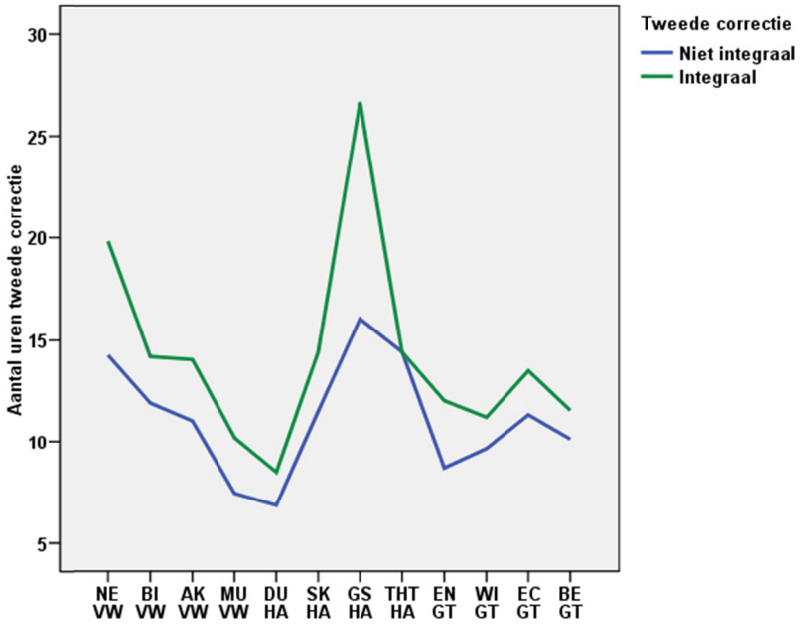
Figuur 14 toont de verdeling van de antwoorden per examen. Bijna de helft van de geschiedenisdocenten herkent zichzelf helemaal niet in de stelling. Waarschijnlijk niet toevallig is geschiedenis ook het examen met de zwaarste correctielast (zie ook Figuur 12). Tweede correctoren muziek, Duits, Engels, wiskunde, economie en beeldende vakken kijken naar eigen zeggen vaak integraal na. Niet onverwacht zijn dit examenvakken met een relatief lichte correctielast (zie ook Figuur 12).



Figuur 14 De mate waarin tweede correctoren integraal nakijken per examen

De overige vier stellingen beschrijven niet-integrale correctiewijzen. Eén procent van de tweede correctoren heeft eerst het examenwerk van enkele kandidaten steekproefsgewijs nagekeken en is daarna gestopt. Deze docenten voeren de tweede correctie niet volgens de wettelijke voorschriften uit. Hetzelfde geldt voor de 1% die eerst enkele vragen nakeek waarover discussie te verwachten valt en daarna stopte. Acht procent van de tweede correctoren corrigeerde in eerste instantie een steekproef van kandidaten en beoordeelde daarna nog meer kandidaten. Negen procent keek eerst enkele vragen na waarover discussie te verwachten valt en beoordeelde daarna nog andere vragen. De beide laatstgenoemde strategieën sluiten integrale correctie niet uit. Het is mogelijk dat de tweede corrector na de steekproef van vragen of examenwerken volledig te hebben beoordeeld de overige vragen en examenwerken alsnog volledig nakijkt. Het onderzoek maakt echter aannemelijk dat het percentage docenten dat eerst steekproefsgewijs nakijkt en uiteindelijk toch nog alle vragen en kandidaten corrigeert niet hoger is dan één procent.

Hiervoor constateerden we dat 44% van de docenten het examenwerk naar eigen zeggen helemaal heeft nagekeken, als ware het een eerste correctie. Nagegaan is hoeveel docenten zich volledig herkennen in de eerste stelling over integrale correctie en zich daarnaast helemaal niet herkennen in de vier stellingen die niet-integrale correctiewijzen beschrijven. Van alle tweede correctoren voldoet bijna één derde (31%) aan dit samengestelde criterium. Van hen mag met een grote mate van zekerheid worden aangenomen dat zij de tweede correctie volledig overeenkomstig de wettelijke richtlijnen uitvoeren. Figuur 15 laat zien dat tweede docenten die integraal nakijken meer tijd in de tweede correctie investeren dan docenten die niet alle kandidaten en/of vragen nakijken (met uitzondering van tehatex).



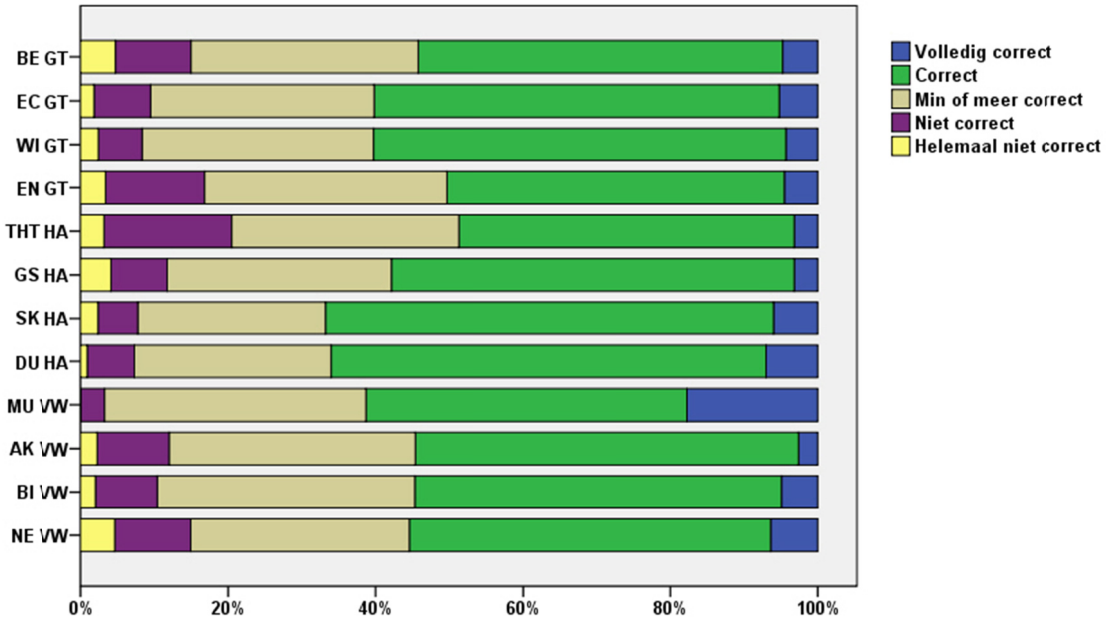
Figuur 15 Tijdbesteding aan de tweede correctie voor integrale en niet-integrale tweede correctie

De vijf stellingen geven geen uitputtende beschrijving van de strategieën die docenten bij de tweede correctie kunnen toepassen. De docenten is daarom de ruimte geboden om hun correctiestrategie toe te lichten. Hieronder vatten we de belangrijkste toelichtingen kort samen:

- Docenten die naar eigen zeggen steekproefsgewijs nakijken, kijken de open vragen vaak wel volledig na, maar de meerkeuzevragen niet. Een deel van hen kijkt de meerkeuzevragen helemaal niet na en een ander deel beperkt zich tot een steekproef uit de meerkeuzevragen.
- Een deel van de docenten die steekproefsgewijs nakijken, corrigeert alleen of vooral het werk van kandidaten op de grens tussen voldoende en onvoldoende voor het CE en/of het SE.
- Een van de redenen waarom de tweede correctie niet integraal wordt uitgevoerd is dat de eerste correctie goed is uitgevoerd. De tweede corrector vindt integrale correctie dan niet nodig. Ook komt het voor dat de eerste correctie zo slecht is uitgevoerd dat de tweede corrector het examenwerk na dat gedeeltelijk te hebben nagekeken terugstuurt naar de eerste corrector.

4.4 Zorgvuldigheid en objectiviteit van de eerste en tweede correctie

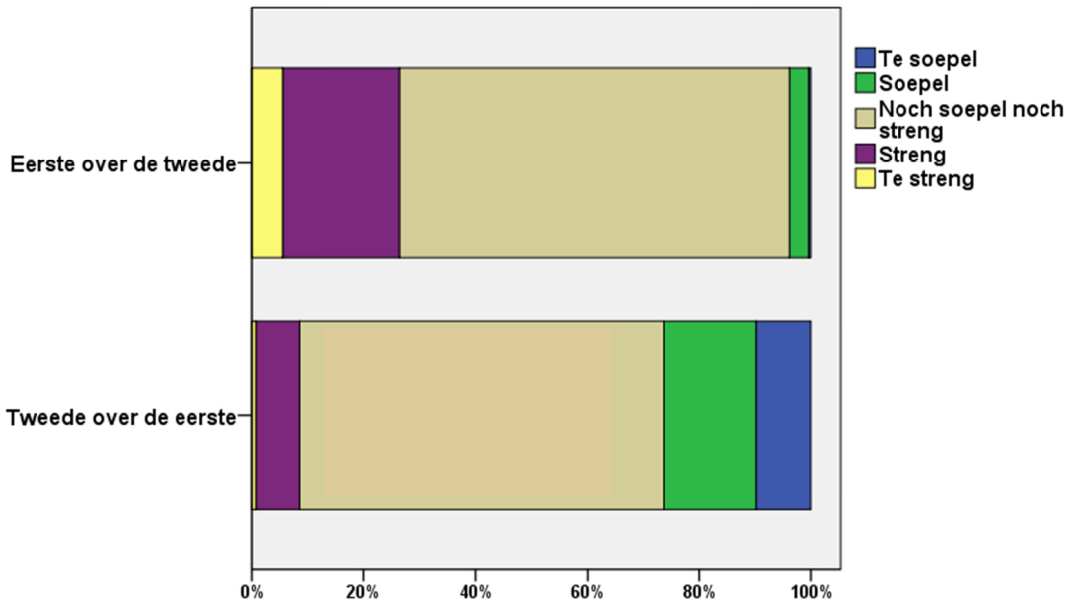
Examinatoren worden geacht het werk van de kandidaten niet alleen volledig maar ook zorgvuldig na te kijken.
Volgens de tweede correctoren past 5% van de eerste correctoren het correctievoorschrift volledig correct toe, 53% doet dat correct, 31% min of meer correct, 9% niet correct en 3% helemaal niet correct. Figuur 16 toont de verdeling van de antwoorden per examen. Docenten tehatex passen het correctievoorschrift volgens de collega's het minst correct toe.



Figuur 16 De correcte toepassing van het correctievoorschrift

Hoe denken eerste en tweede correctoren over de zorgvuldigheid van elkaars beoordelingen? Tweede correctoren zijn kritischer over de zorgvuldigheid van de eerste correctie dan eerste correctoren over de zorgvuldigheid van tweede correctie. Zo vindt 24% van de eerste correctoren dat de tweede corrector zeer zorgvuldig nakeek, terwijl slechts 10% van de tweede correctoren van mening is dat de eerste correctie zeer zorgvuldig is uitgevoerd.

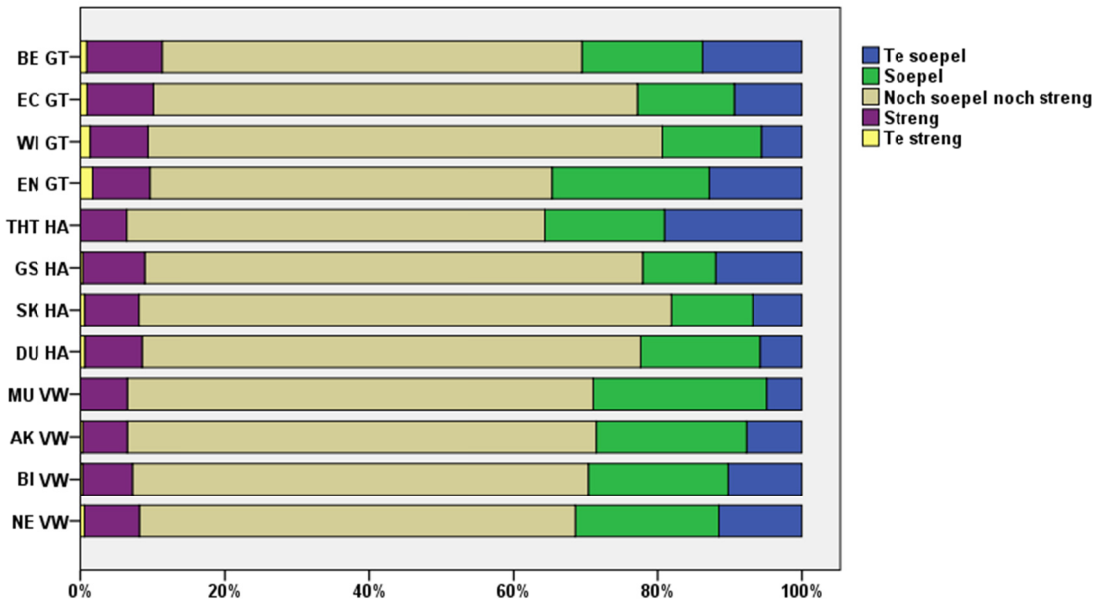
Behalve volledig en zorgvuldig moet de correctie ook objectief zijn. Correctoren mogen kandidaten niet bevoordelen of benadelen door een te soepele of te strenge beoordeling. Met betrekking tot de objectiviteit van de beoordeling is eerste en tweede correctoren gevraagd naar de soepelheid van elkaars beoordelingen. Daarbij konden zij kiezen uit de antwoordmogelijkheden 'Te soepel (d.w.z. kandidaten zijn bevoordeeld)', 'Soepel', 'Noch te soepel noch te streng', 'Streng' en 'Te streng (d.w.z. kandidaten zijn benadeeld)'. Figuur 17 laat zien hoe de antwoorden van de eerste en tweede correctoren verdeeld zijn over de vier antwoordmogelijkheden. De bovenste balk betreft de mening van de eerste corrector over de soepelheid van de tweede corrector en de onderste balk de mening van de tweede corrector over de soepelheid van de eerste corrector.



Figuur 17 Mening van eerste en tweede correctoren over de soepelheid van elkaars beoordelen

De oordelen van de eerste en tweede correctoren over de soepelheid van elkaars beoordelingen lopen ver uiteen. Tien procent van de tweede correctoren vindt dat de eerste corrector te soepel beoordeeld heeft in de zin dat kandidaten bevoordeeld zijn. Daarentegen vindt vrijwel geen enkele eerste corrector (0%) de tweede corrector te soepel. Het omgekeerde patroon zien we bij de percentages strenge en te strenge beoordelingen. Slechts negen procent van de tweede correctoren vindt de eerste corrector streng of te streng. Daarentegen is 26% van de eerste correctoren van mening dat de tweede corrector streng of te streng was. Een mogelijke verklaring veronderstelt dat docenten kritischer staan tegenover de objectiviteit van het werk van een onbekende collega dan tegenover hun eigen werk. Een andere verklaring stelt dat docenten als eerste corrector minder objectief beoordelen dan als tweede corrector. Als eerste correctoren beoordelen docenten immers hun 'eigen' kandidaten terwijl zij met de kandidaten die zij als tweede corrector beoordelen geen persoonlijke band hebben.

Figuur 18 laat zien hoe tweede correctoren van de twaalf examens denken over de soepelheid van de eerste corrector. Volgens de tweede correctoren is de neiging tot het geven van te soepele oordelen bij docenten tehatex het sterkst ontwikkeld. Waarschijnlijk niet toevallig is tehatex in het O2C-onderzoek ook het examen waarbij de eerste correctoren in vergelijking met de derde correctoren het meest soepel beoordeelden.



Figuur 18 Soepelheid van de eerste corrector per examen

4.5 Het overleg tussen de eerste en tweede corrector

De uiteindelijke scores moeten in overleg tussen eerste en tweede corrector worden vastgesteld. Aan deze wettelijke eis wordt vrijwel altijd voldaan. Slechts één procent van de eerste correctoren rapporteert dat er geen overleg met de tweede corrector heeft plaatsgevonden.

De vorm van het overleg

De wet- en regelgeving stelt geen eisen aan de vorm van het gezamenlijke overleg. Dat wordt aan de scholen zelf overgelaten. Het gezamenlijk overleg wordt in 96% van de gevallen via de telefoon gevoerd, 5% via e-mail of vergelijkbaar en 1% via een persoonlijke ontmoeting op afspraak. Van communicatiemiddelen zoals SKYPE en *video conferencing* wordt nog nauwelijks gebruik gemaakt, net als van schriftelijk overleg en andere, niet nader omschreven vormen van overleg.

De duur van het overleg

Het gezamenlijk overleg vergt gemiddeld ongeveer vijftig minuten. De gespreksduur loopt sterk uiteen. De tien procent docenten met de kortste gespreksduur overlegde tot een kwartier en de tien procent met de langste gespreksduur minimaal twee uur.

De sfeer van het overleg

De sfeer van het overleg is doorgaans goed. Vijf procent van de docenten vond de sfeer onplezierig en slechts één procent zeer onplezierig.

Meningsverschillen

De overheid heeft voorschriften opgesteld voor de wijze waarop de uiteindelijke scores tot stand moeten komen. Als de tweede corrector vindt dat er sprake is van grote onzorgvuldigheid, aperte fouten of verkeerde interpretatie van de correctievoorschriften dient hij of zij er eerst in overleg met de eerste corrector uit te komen. In 8% van de gesprekken verschillen eerste en tweede corrector geen enkele keer van mening over de toegekende scores, in 76% is dat af en toe het geval, in 13% regelmatig, in 2% vaak en in 1% zeer vaak. Tweede correctoren rapporteren meer meningsverschillen dan eerste correctoren.

In het geval van meningsverschillen moeten de eerste en tweede corrector eerst proberen er samen uit te komen. Vrijwel alle eerste en tweede correctoren rapporteren dat zij er altijd samen uitgekomen zijn, dit wil zeggen dat er een oplossing gevonden is waarmee beiden kunnen leven.

Middelen van scores

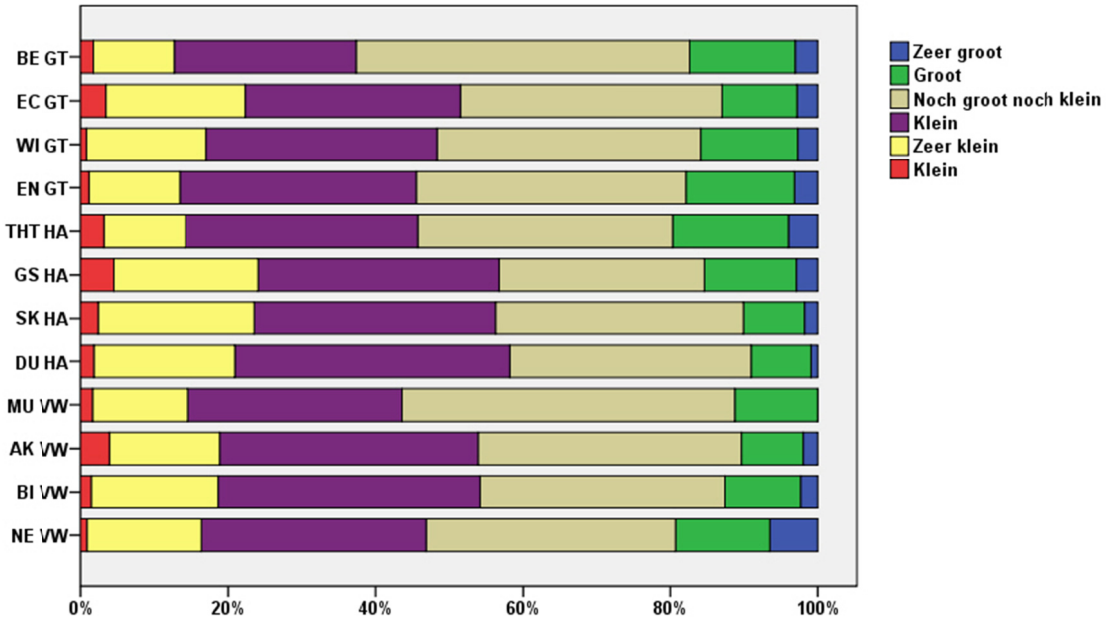
Als eerste en tweede corrector het niet eens kunnen worden, kan het verschil tussen de beide scores worden gemiddeld. Zowel eerste als tweede corrector kunnen middelen echter weigeren. Negen procent van de eerste correctoren heeft nooit een scoreverschil gemiddeld omdat zij het altijd met de tweede corrector eens waren. Van de eerste correctoren rapporteert 43% dat er wel scoreverschillen waren, maar dat er nooit het gemiddelde van beide scores genomen is. Bij 42% is er af en toe een scoreverschil gemiddeld, bij 5% regelmatig, bij 1% vaak en bij 0% zeer vaak.

Inroepen hulp van derden

Als eerste en tweede corrector er samen niet uitkomen, kan de hulp van derden worden ingeroepen. Dat kan een vakcollega zijn, maar ook een lid van de directie of in laatste instantie de Inspectie. Die kan besluiten tot de inzet van een onafhankelijke derde corrector. In 97% van het gezamenlijk overleg is er geen hulp van derden ingeroepen, bij 2% is een collega van de ‘eigen’ school geraadpleegd, bij 1% is dat een collega van de andere school, bij 1% de directie van de ‘eigen’ school, bij 1% de directie van de andere school en bij 0% kwam de Inspectie tussenbeide. De eerste correctoren rapporteren vier gevallen waarin de Inspectie is ingeschakeld en de tweede correctoren zeven gevallen. Eén docent viel twee keer in de prijzen en kreeg zowel als eerste als tweede corrector met de Inspectie te maken. Van de overige negen gevallen waarin de Inspectie bemiddelde waren in acht gevallen beide directies betrokken en in één geval slechts één van beide directies.

4.6 De invloed van de tweede correctie op de uiteindelijke examenscores

Hoe beoordelen eerste correctoren de invloed van de tweede correctie op de uiteindelijke scores? Van de eerste correctoren beoordeelt 1% deze invloed als zeer groot, 4% als groot, een kwart (26%) als noch groot noch klein, 31% als klein, 30% als zeer klein en 7% als nihil. Figuur 19 geeft inzicht in de verschillen tussen examens. Bij het examen scheikunde en geschiedenis is de invloed van de tweede corrector het kleinst en bij muziek en beeldende vakken het grootst. Kennelijk wordt de directe bijdrage van de tweede correctie aan de totstandkoming van de uiteindelijke scores als vrij marginaal ervaren. Deze uitkomst komt overeen met de geringe directe invloed van de tweede correctie zoals vastgesteld in het O2C-onderzoek (Kuhlemeier, Van Rijn & Kremers, 2012).



Figuur 19 De invloed van de tweede correctie op de uiteindelijke examenscores

5. Samenvatting

In deze publicatie is verslag gedaan van drie studies naar het functioneren van de eerste en tweede correctie van de centraal schriftelijke examens in het voortgezet onderwijs. In het hoofdonderzoek hebben in totaal dertig onafhankelijke correctoren het examenwerk van 803 examenkandidaten van 127 scholen opnieuw nagekeken. Daarbij zijn de scores van de derde correctoren vergeleken met die van de 'eigen' docent. In het tweede onderzoek hebben panels van getrainde correctoren een selectie van vragen en examenwerken nog een keer nagekeken. Het panelonderzoek was bedoeld als check op de resultaten van het hoofdonderzoek. Daarnaast is nagegaan in hoeverre verschillen tussen correctoren in soepelheid samenhangen met de aard van de vraag en het beoordelingsmodel. Het derde onderzoek was een schriftelijke enquête naar de praktijk van de eerste en tweede correctie waaraan in totaal 3695 docenten hebben meegedaan (respons: 62%). De conclusies kunnen als volgt worden samengevat:

- Eerste correctoren kennen gemiddeld zes procent hogere scores toe dan 'onafhankelijke' derde correctoren (voor het eerste tijdvak). Het verschil tussen de eerste en derde correctie is niet voor elk examen gelijk. Stel dat we de eerste correctoren zouden vervangen door de derde correctoren, dan stijgt het percentage onvoldoendes bij wiskunde van 23% naar 28%, bij Engels van 27% naar 35%, bij biologie van 15% naar 25%, bij Nederlands van 20% naar 30%, bij geschiedenis van 23% naar 53% en bij tehatex van 22% naar 54%.
- Een toegeeflijke beoordeling leidt niet tot ongelijkheid als alle kandidaten in dezelfde mate bevoordeeld worden. Er zijn echter grote verschillen tussen scholen in de soepelheid van de beoordeling. Op de ene school profiteren leerlingen daar veel meer van dan op de andere school. Zoals Sanders dat al in 1983 treffend formuleerde, kan "het voor leerlingen bijzonder veel uitmaken wanneer zij niet door een milde beoordelaar, maar door een strenge beoordelaar beoordeeld worden. Jammer genoeg (of gelukkig maar) weten zij, noch hun beoordelaars, niet of zij bevoordeeld of benadeeld worden" (pag. 171). De geconstateerde verschillen in soepelheid betekenen dat kandidaten met eenzelfde vaardigheidsniveau op de ene school gemakkelijker een voldoende behalen dan op de andere school.
- Bij geschiedenis, tehatex en Nederlands blijken er docenten te zijn die hun kandidaten onverklaarbaar toegeeflijk beoordelen zonder dat de tweede correctie dat corrigeert. Het vermoeden bestaat dat deze docenten hun rol als examinator niet kunnen en/of willen scheiden van hun rol als opleider. Hoewel begrijpelijk vanuit het standpunt van de examinator die de kandidaten heeft opgeleid, is dit een punt van zorg. Docenten die strategisch-opportunistisch beoordelen, realiseren zich wellicht onvoldoende dat zij hun leerlingen er niet mee helpen als zij te soepel beoordelen en dat zij leerlingen van andere scholen benadelen. De beperkte omvang van het beoordelaarsonderzoek laat het niet toe om de grootte van deze groep precies te bepalen. Wel kan op basis van het vragenlijstonderzoek een eerste voorlopige schatting worden gedaan. Volgens de tweede correctoren beoordeelt tien procent van de eerste correctoren te soepel in de zin dat kandidaten bevoordeeld worden. Over de motieven van de eerste correctoren tasten we hier in het duister. We weten bijvoorbeeld niet in welke mate hier sprake is geweest van welbewuste bevoordeling dan wel van een onbewuste neiging om het voor 'eigen' leerling op te nemen.
- De directe invloed van de tweede corrector op de uiteindelijke scores blijkt klein. De huidige procedure waarbij de ene docent de ander adviseert en controleert, biedt geen afdoende oplossing voor mogelijke bevoordeling of benadeling. Dit resultaat komt overeen met de bevinding van het vragenlijstonderzoek dat docenten de invloed van de tweede correctie op de uiteindelijke scores doorgaans als klein ervaren.
- Blijkens het vragenlijstonderzoek corrigeert de gemiddelde docent het werk van 79 kandidaten en besteedt daar 35 klokuren aan in zeven dagen. De correctielast is echter zeer ongelijk over docenten verdeeld. Geschiedenisdocenten besteden bijvoorbeeld bijna drie keer zoveel tijd aan de correctie als docenten Duits en muziek. Veel docenten ervaren de eerste en tweede correctie als een zware belasting.

- Wat doen scholen om docenten in staat te stellen de eerste en tweede correctie naar behoren uit te voeren? Van 13% van de docenten zijn de uren voor de eerste correctie opgenomen in de jaartaak (bijvoorbeeld in de opslagfactor), 8% krijgt een vast bedrag per examenwerk, 2% kan onkosten declareren (bijvoorbeeld telefoonkosten), 1% ontvangt een kleine attentie (bijv. een boekenbon) en bij 3% wordt de eerste correctie op een andere, niet nader omschreven manier gefaciliteerd. Scholen faciliteren de tweede correctie anders dan de eerste correctie. De verschillen zitten hem vooral in het jaartaakbeleid en het stukloon. Voor de tweede correctie zijn de uren minder vaak in de jaartaak opgenomen, maar daar staat tegenover dat de docenten vaker een vast bedrag per examenwerk krijgen dan voor de eerste correctie. Overigens is in de vragenlijst alleen gevraagd naar expliciete faciliteringsmaatregelen. Veel scholen kennen de stilzwijgende afspraak dat docenten de correctie uitvoeren in de tijd die vrijkomt doordat de ingeroosterde lessen in examenklassen bij aanvang van het CSE komen te vervallen. In het onderzoek is niet nagegaan in hoeverre deze vrijgekomen tijd toereikend is en daadwerkelijk aan de correctie besteed kan worden.
- Hoe tevreden zijn docenten over het schoolbeleid? Driekwart is van mening dat de eerste correctie op school niet gefaciliteerd wordt en bij de tweede correctie gaat het om bijna de helft. Met het schoolbeleid ten aanzien van de eerste en tweede correctie is respectievelijk 39% en 49% ontevreden of zeer ontevreden.
- De tweede correctie blijft vrijwel nooit achterwege, maar slechts een derde corrigeert integraal overeenkomstig de wettelijke voorschriften. Twee derde van de tweede correctoren kijkt niet alle werken na en/of per kandidaat niet alle vragen. Alhoewel het aantal beoordeelde kandidaten vrijwel gelijk is, steken docenten bijna twee keer zoveel tijd in de eerste correctie als in de tweede correctie.
- Docenten noemen onder meer de volgende redenen waarom de tweede correctie sneller gaat dan de eerste correctie:
  - De tweede correctie wordt minder zorgvuldig uitgevoerd;
  - Men heeft meer ervaring met de toepassing van het correctievoorschrift;
  - De eerste corrector stuurde het examenwerk pas laat op;
  - Men moet te veel examens nakijken;
  - Men beschikt over de puntentoekenning en aantekeningen van de eerste corrector;
  - Als de leerling het antwoord onduidelijk geformuleerd heeft of als de examenvraag slecht geformuleerd is, gaat men sneller akkoord met het voorstel van de eerste corrector.
- Begrijpelijkerwijs voeren correctoren van examens met een zware correctielast de tweede correctie minder volledig, minder zorgvuldig en minder objectief uit dan collega's met een lichte correctielast. Als men de tweede correctie in overeenstemming zou willen brengen met de wettelijke voorschriften, lijken maatregelen nodig.
- Het gezamenlijk overleg tussen eerste en tweede corrector vergt gemiddeld ongeveer vijftig minuten en vindt vrijwel altijd via de telefoon plaats. De sfeer van het overleg is doorgaans goed, er zijn relatief weinig meningsverschillen, er wordt vrijwel altijd een oplossing gevonden waarmee beiden kunnen leven en er wordt zelden de hulp van derden ingeroepen.
- Verschillen in soepelheid tussen correctoren blijken ook samen te hangen met kenmerken van het examen. De onderzochte beoordelingsmodellen bieden de corrector vaak onvoldoende steun en dragen daarmee bij aan verschillen tussen correctoren. Verschillen in soepelheid doen zich onder meer voor als het beoordelingsmodel de antwoorden van de leerlingen niet goed dekt en alleen de maximumscore in het beoordelingsmodel omschreven is. De panels hebben een groot aantal aanbevelingen gedaan voor aanpassing van de beoordelingsmodellen. Nader ontwikkelingsonderzoek zal moeten uitwijzen in hoeverre deze suggesties tot de beoogde verbeteringen leiden.

6. Aanbevelingen

De bevindingen van de drie studies staan op gespannen voet met het uitgangspunt dat kandidaten met dezelfde vaardigheid ongeacht de school waarop zij zitten een gelijke kans hebben om voor het examen te slagen (De Groot & Wijnen, 1983). De bevinding dat het directe effect van de tweede correctie op de uiteindelijke scores klein is, betekent niet dat de tweede correctie zinloos is en afgeschaft zou moeten worden. Behalve een direct effect heeft de tweede correctie immers ook een indirect effect. Het zorgt ervoor, aldus Algra (2004), dat de eerste corrector 'niet zo maar zijn gang kan gaan' (p. 1). De wetenschap dat er een tweede correctie plaatsvindt, kan naar soepelheid neigende docenten ervan weerhouden hun kandidaten al te zeer te bevoordelen. We pleiten er dan ook niet voor om de correctie bij de 'eigen' docent weg te halen. De ervaringen in het Verenigd Koninkrijk laten zien dat aan 'onafhankelijke' correctie door externe correctoren vele nadelen verbonden zijn. Wel kunnen op grond van ons onderzoek binnen de huidige examensystematiek verbeteringssuggesties worden gedaan. Hieronder doen we aanbevelingen voor de regelgeving en examenprocedures, het schoolbeleid, de opleiding en training van docenten en het ontwerp en de constructie van het examen.

Regelgeving en examenprocedures

Voor de uitvoering van de eerste en tweede correctie bestaan wettelijke regels en uitvoeringsprotocollen (o.a. Ministerie van OCW, 2012; VO-raad, 2012). Op basis van de drie studies kunnen de volgende verbeteringssuggesties worden gedaan:

- In de huidige procedure vindt de tweede correctie na de eerste correctie plaats. Een alternatief is de examenwerken eerst digitaal te scannen en de kopie vervolgens naar de eerste en tweede correctoren te verzenden. De eerste en tweede correctie kunnen dan 'tegelijkertijd' plaatsvinden. De tweede corrector hoeft niet meer te wachten tot de 'eigen' docent klaar is. Overigens is het de vraag hoe groot deze tijdwinst in de praktijk zal zijn. Vrijwel alle eerste correctoren zijn immers ook tweede corrector. Wel voorkomt het gelijktijdig nakijken de tijdsklem die optreedt als de eerste corrector het werk pas zeer laat naar de tweede corrector opstuurt. Een tweede voordeel is dat alle correctoren het examenwerk integraal moeten nakijken, dus ook de tweede correctoren. Ook op dit voordeel valt wat af te dingen. De tweede corrector moet het examenwerk namelijk ook integraal nakijken als de 'eigen' docent het werk zeer zorgvuldig en objectief heef nagekeken en een steekproefsgewijze controle voldoende garantie zou hebben geboden. Omdat integrale tweede correctie tegenwoordig wettelijk verplicht is, is dit tegenargument formeel gezien niet erg sterk en alleen van praktische importantie. Een derde voordeel van gelijktijdige eerste en tweede correctie is dat alle correctoren een blanco examen onder ogen krijgen. Daardoor wordt de tweede corrector niet beïnvloed door de scores en aantekeningen van eerste corrector. Over de vraag of het wenselijk is dat de tweede corrector niet meer kan beschikken over annotaties van de eerste corrector zijn de meningen echter verdeeld. De aantekeningen van de eerste corrector op het examenwerk laten zien hoe de score tot stand is gekomen. Deze informatie biedt de tweede corrector steun bij het beoordelen; bovendien kan het goed van pas komen in het overleg tussen eerste en tweede corrector waarin de uiteindelijke score wordt vastgesteld. Het lijkt dan ook begrijpelijk dat tweede correctoren de annotaties van de eerste corrector over het algemeen sterk op prijs stellen.
- Voorgesteld wordt het scannen en vervolgens gelijktijdig corrigeren van examenwerk in de praktijk uit te proberen. Het ligt voor de hand deze proef uit te voeren bij de kernvakken Nederlands, Engels en wiskunde. De logistieke uitdagingen van gelijktijdige digitale correctie van gescande examenwerken zullen niet gering zijn. Vanwege het kleinere aantal kandidaten zou met het tweede tijdvak begonnen kunnen worden. Een aanvullende reden is dat het tweede tijdvak wellicht meer uitnodigt tot soepel beoordelen dan het eerste tijdvak. Er staat voor de kandidaten immers meer op het spel.
- Het toezicht op de uitvoering van de eerste en tweede correctie is in handen van de Inspectie van het Onderwijs. Het huidige toezicht biedt correctoren weinig steun. Te overwegen valt het Inspectietoezicht te intensiveren. Een voor de hand liggend middel is steekproefsgewijze controle

van nagekeken examenwerk tijdens en kort na de examencampagne. Vanwege het arbeidsintensieve en tijdrovende karakter zal de herbeoordeling van het examenwerk voor de kandidaten te laat komen.
- Het ene examen kent een veel grotere correctielast dan het andere examen. Arbeidsintensieve vakken zoals geschiedenis en Nederlands zouden altijd aan het begin van het examenrooster geplaatst kunnen worden, zodat correctoren een langere periode hebben voor de correctie.
- Overeenkomstig algemene scoringsregel 7 mogen correctoren niet zelfstandig afwijken van het beoordelingsmodel. In het geval van onvolkomenheden of fouten in het examen moet de corrector het werk beoordelen alsof het examen juist is, waarbij de corrector de vermeende fout aan het CvE kan melden. In de praktijk blijken zelfstandig afwijkende correctoren dat maar zelden te doen. Te overwegen valt de bestaande regelgeving op dit punt te verhelderen en aan te scherpen. Algemene scoringsregel 7 zou zo veranderd kunnen worden dat correctoren die zelfstandig afwijken, verplicht worden dat aan het CvE te melden (met vermelding van de vermeende onvolkomenheid of fout in de vraag of het beoordelingsmodel). Daartoe zal de meldingsprocedure eenvoudiger en toegankelijker gemaakt moeten worden. Een voor de hand liggende mogelijkheid is het toevoegen van een module aan WOLF waarmee correctoren vermeende onvolkomenheden in het beoordelingsmodel kunnen doorgeven.
- Tijdens de paneldiscussies speelden de verslagen van de regionale en landelijke examenbesprekingen een belangrijke rol. Deze verslagen bestaan grotendeels uit een opsomming van lastig te beoordelen antwoorden die goed te rekenen zijn (en veel minder vaak uit antwoorden die fout gerekend moeten worden). Naar de mening van de panels leidt het gebruik van de examenbesprekingen over het algemeen tot een versoepeling van de beoordeling. Verschillen tussen correctoren in soepelheid ontstaan waar de ene corrector de examenbesprekingen wel gebruikt en de ander dat niet doet. Dit pleit voor een proactieve, integrale en centrale regievoering over de examenbesprekingen, bij voorkeur uit te voeren onder de gezamenlijke verantwoordelijkheid van het CvE, Cito en de vakverenigingen. Overigens kan men zich afvragen in hoeverre bespreking van het examen achteraf nog nodig is als de examenconstructie meer gebaseerd zou worden op analyse van feitelijke antwoorden van kandidaten tijdens de constructie- en testfase. Ook hier lijkt te gelden dat voorkomen beter is dan genezen.

Het schoolbeleid

Het vragenlijstonderzoek bevestigt dat de eerste en tweede correctie vaak onder tijdsdruk worden uitgevoerd. Mede daardoor wordt de tweede correctie lang niet altijd volledig en zorgvuldig uitgevoerd. De volgende vier aanbevelingen zijn gericht op het schoolbeleid:
- In het advies ‘Examinering: Draagvlak en toegankelijkheid’ vraagt de Onderwijsraad (2006) de scholen om in het taak- en vergoedingenbeleid rekening te houden met het werk dat de correctie met zich meebrengt. Een voor de hand liggende maatregel is de docenten hiervoor vrij te roosteren. Het Platform VVVO (2008b) adviseert de voor de correctie benodigde tijd “zichtbaar en geoormerkt in de taakbelasting van de betrokken docenten op te nemen” (p. 1).
- Vastgesteld is dat de correctielast van het ene examen veel groter is dan van het andere. Voor zover scholen de correctie al via expliciete maatregelen faciliteren, lijkt er sprake van een voor iedereen geldende aanpak. Te overwegen valt het huidige ‘one-size-fits-all’-beleid te vervangen door een gedifferentieerde aanpak afhankelijk van de omvang van de correctielast.
- In de huidige situatie hebben schoolleiders weinig zicht op de correctie van de examens. Te overwegen valt een module aan WOLF toe te voegen die schoolleiders inzicht geeft in de uitvoering en kwaliteit van de eerste en tweede correctie. Zo kunnen schoolleiders eventuele problemen vroegtijdig signaleren en zo nodig oplossen.
- Examencijfers vervullen tegenwoordig allerlei functies waarvoor ze oorspronkelijk niet bedoeld waren. Een voorbeeld is het gebruik van examenresultaten voor publieke verantwoording van de kwaliteit van de school (denk aan de ranglijsten van scholen). Een andere voorbeeld is het gebruik van examencijfers voor personeelsbeoordeling, ook wel prestatiedifferentiatie of ‘loon naar lesgeven’ genoemd. Deze nieuwe examenfuncties kunnen strategisch-opportunistisch

beoordelingsgedrag in de hand werken en de neiging tot lankmoedig beoordelen versterken. Docenten die welbewust te soepel nakijken, bevoordelen hun eigen school en leerlingen (en wellicht zichzelf), maar duperen leerlingen van scholen waar wel integer beoordeeld wordt. Deze ongewenste neveneffecten kunnen de nu nog hoge waarde van het diploma op termijn aantasten. Te overwegen valt schoolleiders voor te lichten over de voor- en nadelen van het sturen op examencijfers.

Opleiding en training van docenten

Docenten zijn doorgaans niet geschoold in het examineren en beoordelen van kandidaten. Zij kunnen zich bijvoorbeeld nog niet als examinator laten certificeren. Enkele mogelijke maatregelen zijn:

- Het opnemen van een module 'examineren' in de initiële opleiding voor docenten die mede gericht is op de correctie van het CSE en waarbij het accent ligt op 'Wat betekent het voor mij als docent dat ik ook examinator ben?'
- Het aanbieden van vergelijkbare modules in de postinitiële opleidingen.
- Het trainen van docenten in het gebruik van het beoordelingsmodel en de algemene scoringsregels (onder meer met gebruikmaking van materiaal uit het O2C- en P4C-project).

Het ontwerp en de constructie van het examen

De geconstateerde verschillen tussen correctoren en de overwegend te soepele beoordeling kunnen gedeeltelijk verklaard worden vanuit kenmerken van het examen. Een aandachtspunt is aanpassing van de beoordelingsmodellen bij de open vragen, zodat correctoren meer steun krijgen bij het nakijken en minder vaak zelfstandig een beroep op de algemene scoringsregels hoeven te doen. In het panelonderzoek is een groot aantal aanbevelingen gedaan die verschillen tussen correctoren en de verleiding tot te soepel beoordelen kunnen tegengaan (zie Kuhlemeier e.a., 2012). De twee belangrijkste aanbevelingen zijn:

- Verschillen tussen correctoren in soepelheid doen zich vooral voor als het beoordelingsmodel de verzameling van feitelijke antwoorden niet goed dekt en alleen de maximumscore in het beoordelingsmodel omschreven is. De belangrijkste aanbeveling is het verzamelen van antwoorden van kandidaten tijdens de ontwerp- en testfase van het examen. Momenteel gebeurt het slechts op beperkte schaal dat Cito open vragen aan een kleine groep studenten onder gecontroleerde omstandigheden voorlegt. Door de antwoorden vervolgens na te kijken krijgt de examenmaker kwalitatieve informatie ter verdere verbetering van het beoordelingsmodel. Dit draagt ertoe bij dat correctoren minder vaak een beroep hoeven te doen op de verslagen van de examenbesprekingen en de algemene en vakspecifieke scoringsregels. Verdere verbetering van het beoordelingsmodel is mogelijk door het toevoegen van uitleg en voorbeelden voor het beoordelen van veel voorkomende antwoorden die geheel of gedeeltelijk fout zijn en die de 'eigen' docent ten onrechte goed zou kunnen rekenen.
- De onderzochte examens geschiedenis, tehatex en Nederlands bevatten relatief veel open vragen. De beoordelingsmodellen van deze examens zijn vaak principieel onvolledig. Voorbeelden zijn: 'Een voorbeeld van een juist antwoord is', 'Een goed antwoord moet de volgende strekking hebben', 'De kern van een juist antwoord is' en 'Uit het antwoord moet blijken dat'. Te overwegen valt volledig af te zien van dit type open vragen of een deel ervan te vervangen door objectief scoorbare gesloten vragen of door in- of aanvulvragen waarbij een kort antwoord volstaat. Een bijkomend voordeel is een aanzienlijke reductie van de correctielast.

Vervolgonderzoek

- Voor het onderzoek naar verschillen tussen docenten in soepelheid en naar de invloed van de tweede correctie op de uiteindelijke scores is alleen examenwerk van het eerste tijdvak gebruikt. Het tweede tijdvak is met name bedoeld als herkansing voor kandidaten die op basis van het eerste tijdvak gezakt zouden zijn. Omdat het de laatste kans is, staat er meer op het spel dan bij het eerste tijdvak. Voor naar soepelheid neigende docenten zou de verleiding om kandidaten te bevoordelen bij het tweede tijdvak wellicht nog groter kunnen zijn dan bij het eerste tijdvak. Mocht

het onderzoek naar de verschillen tussen eerste, tweede en derde correctie over enige tijd herhaald worden, lijkt het zinvol daarbij ook het tweede tijdvak te betrekken.

- Op basis van het panelonderzoek zijn aanbevelingen gedaan voor de aanpassing van de vraagstelling en de beoordelingsmodellen van toekomstige examens. Nader ontwikkelingswerk en onderzoek zal moeten uitwijzen in hoeverre deze aanpassingen leiden tot een vermindering van de verschillen tussen correctoren. Bij gebleken effectiviteit zal Cito deze wijzigingen doorvoeren in de reguliere examenproductie.
- De docentenpanels concluderen dat er eerste correctoren zijn die een aanzienlijk deel van de antwoorden onverklaarbaar toegeeflijk beoordelen (zonder dat dit kan worden toegeschreven aan een gebrek aan vakkennis, onvolkomenheden in het examen of de wijze waarop de kandidaat het antwoord geformuleerd heeft en dergelijke). De beperkte omvang van het beoordelaarsonderzoek laat het doen van een uitspraak over het exacte percentage strategisch-opportunistische beoordelingen niet toe. Daartoe is vervolgonderzoek nodig met meer examenwerken en correctoren.

Literatuur

Alberts, R., & Erens, B. (2012). *Verslag van de examencampagne 2011 voortgezet onderwijs*. Arnhem: Cito.

Algra, A. (2004). Eerste en tweede correctie examens: problemen en regels. S*choolmanagers_VO #6, 8-10.*

Bergh, H. van den & Kuhlemeier, H. (1997). Multiniveau modellen voor de analyse van leerwinst vergeleken. *Tijdschrift voor Onderwijsresearch, 22*, 2, 54-75.

Ministerie van OCW (2012). *Examenbesluit*. Online beschikbaar via www.examenblad.nl.

Gitsels, H., & Kuhlemeier, H. (in voorbereiding). Arnhem: Cito.

Groot, A.D. de, & Wijnen, W.H.F.W. (1983) *Vijven en zessen*, Groningen: Wolters-Noordhoff.

Kuhlemeier, H., & Dietvorst, P. (2009). *De praktijk van de beroepsgerichte examens voor het vmbo. Resultaten van een onderzoek naar de voorbereiding, afname, beoordeling, tweede correctie en herkansing van vier beroepsgerichte examens*. Arnhem: Cito. http://www.onderwijsinspectie.nl/binaries/content/assets/Actueel_publicaties/2009/Praktijk+beroepsgerichte+examens+vmbo+-+printversie.pdf

Kuhlemeier, H., Gitsels, H., Boom, S., Kerkhof, A. van de, & Sinkeldam, R. (2012). *Relaties tussen examenkenmerken en verschillen tussen correctoren in soepelheid bij het CSE geschiedenis, tehatex en Nederlands*. Arnhem: Cito.

Kuhlemeier, H., & Kremers, E. (2012). *De praktijk van de eerste en tweede correctie van de centraal schriftelijke examens*. Arnhem: Cito.

Kuhlemeier, H., Rijn, P. van, & Kremers, E. (2012). *Eerste, tweede en derde correctie van geannoteerde en blanco examenwerken in de centraal schriftelijke examens: Wat is het verschil?* Arnhem: Cito.

Onderwijsraad (2006). *Examinering: Draagvlak en toegankelijkheid*. Den Haag: Onderwijsraad. Online beschikbaar via: http://www.onderwijsraad.nl/upload/publicaties/316/documenten/examinering__draagvlak_en_toegankelijkheid.pdf

Platform VVVO (2008a). *Tweede correctie is gekkenwerk geworden* (persbericht van 29 mei 2008). Online beschikbaar via: http://www.platformvvvo.nl/brieven-archief/198-persbericht-integrale-tweede-correctie.html

Platform VVVO (2008b). *Tijd nodig voor tweede correctie*. Online beschikbaar via: http://www.platformvvvo.nl/brieven-archief/233-tijd-nodig-voor-integrale-tweede-correctie.html

Sanders, P. (1983). Objectieve beoordeling van open-vragen examens. In P. Weeda (red.). *Examens in discussie: Een bundel opstellen voor J.W. Solberg* (pp. 163-172). Groningen: Wolters-Noordhoff.

VO-raad (2011). *Protocol eerste en tweede correctie centrale examens vmbo, havo en vwo*. Online beschikbaar via www.vo-raad.nl