

Eerste, tweede en derde correctie van examenwerk: Wat is het verschil?

Inhoud

Samenvatting.....	3
1. Inleiding	7
2. Methode van onderzoek	14
2.1 Steekproef van vestigingen en kandidaten	14
2.2 Verzameling aanvullende gegevens	15
2.3 Werving van correctoren	16
2.4 Digitaliseren van de examenwerken.....	16
2.5 Toewijzing van kopieën B en C aan correctoren.....	17
2.6 Correctie van het examenwerk.....	17
2.7 Volledigheid van de verzamelde gegevens	18
2.8 Representativiteit van de responsegroep.....	19
2.9 Statistische analyse	20
2.10 Kwaliteit van de derde correctie	21
2.10.1 Nederlands VWO.....	21
2.10.2 Engels GT	22
2.10.3 Wiskunde GT	23
2.10.4 Biologie 1, 2 Vwo	23
2.10.5 Geschiedenis Havo.....	24
2.10.6 Tehatex Havo	25
3. Resultaten.....	26
3.1 Nederlands VWO.....	26
3.2 Engels GLTL.....	35
3.3 Wiskunde GLTL	44
3.4 Biologie VWO	52
3.5 Geschiedenis HAVO.....	60
3.6 Tehatex HAVO.....	68
4 Discussie	76
Bijlagen	86
Bijlage 1 Spreidingsdiagrammen Nederlands VWO	86
Bijlage 2 Spreidingsdiagrammen Engels GLT	88
Bijlage 3 Spreidingsdiagrammen wiskunde GLTL	90
Bijlage 4 Spreidingsdiagrammen biologie VWO	92
Bijlage 5 Spreidingsdiagrammen geschiedenis HAVO	94
Bijlage 6 Spreidingsdiagrammen tehatex HAVO	96

Samenvatting

In de centrale schriftelijke examens wordt het examenwerk eerst door de eigen docent gecorrigeerd en vervolgens door een docent van een andere school. Eerste correctoren krijgen daarbij het advies om het examenwerk van annotaties te voorzien. Volgens de richtlijnen van de VO-raad (2009) dient de eerste corrector duidelijk aan te geven waar in het werk de onvolkomenheden zitten en de toegekende punten per vraag voor de linker kantlijn van het examenwerk te noteren. Over de vraag of het wenselijk is dat de tweede corrector kennis neemt van de annotaties van de eerste corrector zijn de meningen verdeeld. Van de ene kant bieden de punten en aantekeningen van de eerste corrector de tweede corrector steun bij het beoordelen. Van de andere kant kunnen de annotaties van de eerste corrector de objectiviteit van de tweede correctie in gevaar brengen. In dit onderzoek is onder meer nagegaan in hoeverre het verwijderen van de annotaties van de eerste corrector van invloed is op de scores van de tweede corrector. Daartoe heeft een steekproef van correctoren een steekproef van examenwerken uit het schooljaar 2008-2009 opnieuw gecorrigeerd. Deze hernieuwde correctie is uitgevoerd door zogeheten derde correctoren. Daarbij is van elk examenwerk zowel een geannoteerde als een blanco versie nagekeken. Afgezien van het ontbreken van de annotaties van de eerste correctie op het examenwerk was de blanco versie volledig identiek aan de geannoteerde versie. Daarnaast zijn de scores van de derde correctoren vergeleken met de scores van de eerste corrector en met de gezamenlijk tussen eerste en tweede corrector overeengekomen scores (eerste tijdvak). Voor de vakken Nederlands, Engels, wiskunde, biologie, geschiedenis en tehatex hebben dertig derde correctoren in totaal 616 examenwerken opnieuw gecorrigeerd (waarvan 562 annotaties van de eerste corrector bevatten). Per vak waren vijf correctoren in het onderzoek betrokken. Een check op de kwaliteit van de derde correcties toonde aan dat de scores van de vijf correctoren in voldoende mate overeenstemmen. Hieronder vatten we de belangrijkste onderzoeksresultaten samen en doen we enkele aanbevelingen voor vervolgonderzoek.

Geannoteerde versus blanco derde correctie

Doel van het onderzoek was na te gaan in hoeverre het verwijderen van de annotaties van de eerste corrector van invloed is op de scores van een tweede corrector. De desbetreffende onderzoeksvraag luidde: 'In hoeverre is er een verschil tussen de scores van geannoteerde en blanco derde correctie? Het onderzoek maakt aannemelijk dat derde correctoren zich door de annotaties van de eerste corrector laten beïnvloeden. Hoewel er verschillen tussen examenvakken zijn, liggen de scores van geannoteerde en blanco derde correctie vaak behoorlijk ver uit elkaar. Voor de vakken biologie, geschiedenis en tehatex zou blanco derde correctie tot duidelijk hogere percentages onvoldoendes hebben geleid dan geannoteerde derde correctie (bij toepassing van de definitieve normeringstermen schooljaar 2008-2009).

Eerste correctie versus geannoteerde en blanco derde correctie

De derde correctie is ook vergeleken met de eerste correctie zoals deze door de 'eigen' docent tijdens het eerste tijdvak is uitgevoerd. De desbetreffende onderzoeksvraag luidde: 'In hoeverre is de samenhang tussen eerste en derde correctie voor geannoteerde werken anders dan voor blanco werken?' Ervan uitgaande dat derde correctoren zich door de annotaties van de eerste corrector laten beïnvloeden verwachtten we dat eerste correctie sterker zou samenhangen met geannoteerde dan met blanco derde correctie. Inderdaad blijken de scores van de eerste correctoren over het algemeen meer te lijken op die van geannoteerde dan van blanco derde correctie. Ook deze bevinding wijst erop dat derde correctoren rekening houden met de annotaties van de eerste corrector. Kennelijk is onafhankelijke tweede correctie het beste gegarandeerd als het examenwerk geen annotaties van de eerste corrector bevat. De derde correctoren in het onderzoek hadden de opdracht om het examenwerk zo volledig, objectief en zorgvuldig mogelijk na te kijken. Bovendien kregen zij ruim de tijd om het correctiewerk uit te voeren en ontvingen zij een vergoeding die vergelijkbaar is met hetgeen correctoren van de staatsexamens ontvangen. Hiermee kan de vraag of het verwijderen van

de annotaties van de eerste corrector nog nodig is als tweede correctoren voldoende tijd krijgen bevestigend beantwoord worden.

Het geconstateerde verschil tussen geannoteerde en blanco derde correctie lijkt op gespannen voet te staan met het uitgangspunt dat iedere leerling recht heeft op een zo onafhankelijke mogelijke correctie van zijn of haar examenwerk. Vanuit het perspectief van een rechtvaardige beoordeling valt te overwegen de huidige richtlijn voor het annoteren van examenwerk opnieuw te bezien (VO-raad, 2009). Een eventueel besluit ten faveure van blanco tweede correctie dient te worden genomen op basis van een zorgvuldige afweging van de voor- en nadelen van blanco versus geannoteerde correctie. Daarbij dienen ook de omstandigheden waaronder de correctie plaatsvindt in ogenschouw genomen te worden.

Het onderhavige onderzoek is niet onder reële examenomstandigheden uitgevoerd (onder meer omdat de derde correctoren de examenwerken buiten de examenperiode nakijken en zoals gezegd ruim de tijd kregen en een redelijke vergoeding ontvingen). Uit onderzoek van het Platform VVVO (2008a) blijkt dat integrale tweede correctie voor docenten een hoge taakbelasting met zich meebrengt; docenten blijken de integrale tweede correctie alleen ten koste van hun eigen vrije tijd en nachtrust te kunnen uitvoeren. Als de overheid aan een integrale tweede correctie wil vasthouden, aldus het Platform VVVO, zal de voor de correctie benodigde tijd 'zichtbaar en geormerkt in de taakbelasting van de betrokken docenten opgenomen moeten worden (p. 1)'. Vandaar dat wij pleiten voor een proef met 'onafhankelijke' tweede correctie van blanco examenwerken tijdens de examencampagne. Voor de uitvoerbaarheid is het van belang dat eerste en tweede correctoren voldoende tijd krijgen om het correctiewerk integraal uit te voeren. Een mogelijke uitwerking is dat de examenwerken na gescand te zijn digitaal naar de correctoren verzonden worden. Zo krijgen alle correctoren een blanco examen onder ogen en hoeven tweede correctoren niet meer te wachten tot de 'eigen' docent klaar is. Voor de onafhankelijkheid van de correctie is het van belang dat de tweede corrector geen inzage krijgt in de schoolexamencijfers van de kandidaten.

Eerste correctie versus derde correctie

Op grond van eerder onderzoek verwachtten we dat de 'eigen' docent over het algemeen hogere scores zou toekennen dan 'onafhankelijke' derde correctoren. Dit blijkt inderdaad het geval. Eerste correctoren blijken over het algemeen soepeler na te kijken dan derde correctoren. Dit wordt ook zichtbaar in de percentages onvoldoendes die voor eerste correctie duidelijk lager zijn dan voor derde correctie. De geconstateerde verschillen tussen eerste en derde correctie staan op gespannen voet met het uitgangspunt dat iedere leerling recht heeft op een rechtvaardige beoordeling. Voor het cijfer op het examen zou het niet uit mogen maken wie het examenwerk nakijkt. De opzet van het onderzoek laat geen honderd procent zekere uitspraken toe over mogelijke oorzaken. Daartoe is relatief kleinschalig vervolgonderzoek nodig waarbij een panel van intensief getrainde correctoren een beperkte steekproef uit de examenwerken onder gecontroleerde omstandigheden nogmaals beoordeelt en de resultaten daarvan vervolgens vergelijkt met de eerste en derde correcties uit dit onderzoek. Verder lijkt onderzoek geboden naar de wijze waarop docenten de eerste correctie uitvoeren. In allerlei beoordelingscontexten is namelijk aangetoond dat beoordelaars te soepel beoordelen en te weinig spreiden tussen meer en minder vaardige kandidaten als zij degenen die zij moeten beoordelen zelf hebben opgeleid. Het is daarom verleidelijk om het verschil in soepelheid tussen eerste en derde correctoren mede toe te schrijven aan het gegeven dat de derde corrector de kandidaat niet persoonlijk kent en geen persoonlijk belang heeft bij te soepele (of te strenge) correcties.

Verschillen tussen opgaventypen

In het onderzoek is ook gekeken naar de invloed van het opgaventype. De desbetreffende onderzoeksvraag luidde: 'In hoeverre is de overeenstemming tussen de scores van verschillende correctieprocedures afhankelijk van het opgaventype? De drie correctieprocedures waarvan we de scores per opgave konden vergelijken waren eerste correctie, geannoteerde derde correctie en blanco derde correctie. De verwachting was dat de verschillen tussen de scores van verschillende correctieprocedures voor gesloten vragen kleiner zijn dan voor open vragen. Daarnaast verwachtten

we dat dit verschil in overeenstemming tussen eerste en derde correctoren in het nadeel van open vragen voor geannoteerde examenwerken minder groot is dan voor blanco examenwerken. Bij de meerkeuzevragen maakt het niets uit of de opgaven door eerste of derde correctoren worden nagekeken of dat de derde correctie plaatsvindt op basis van geannoteerd dan wel blanco examenmateriaal (afgezien van sporadische slordigheidsfoutjes). Bij de lang-antwoordvragen leiden de correctieprocedures daarentegen tot duidelijk verschillende scores. Kennelijk zijn het vooral de annotaties bij de lang-antwoordvragen die de scores van de derde correctoren hebben beïnvloed. Een voor de hand liggende verklaring is dat de antwoordmodellen bij open vragen de correctoren minder steun bij het nakijken bieden dan de correctievoorschriften bij meerkeuzevragen (die objectief scoorbaar zijn). De ruimere mogelijkheden voor eigen interpretatie verhogen mogelijk de onzekerheid van de corrector waardoor deze eerder geneigd zal zijn zich door externe informatie te laten leiden, zoals de annotaties van de eerste corrector.

Correctietijd van geannoteerde en blanco derde correctie

Het onderzoek had ook tot doel om inzicht te verschaffen in de tijdsinvestering die gemoeid is met integrale tweede correctie en in de vraag of geannoteerde werken meer, evenveel of minder correctietijd kosten dan blanco werken. De tijd die met het nakijken van het examenwerk gemoeid is, blijkt sterk van vak tot vak te verschillen. Het varieert van gemiddeld ruim elf minuten voor een examenwerk Engels tot bijna dertig minuten voor een examenwerk tehate. Bij vier van de zes vakken blijken geannoteerde werken duidelijk minder correctietijd te kosten dan blanco werken. Ook dit lijkt erop te wijzen dat derde correctoren zich door de annotaties van de eerste corrector hebben laten beïnvloeden. De bevinding dat geannoteerde werken minder correctietijd kosten, staat op gespannen voet met het gegeven dat het lezen en interpreteren van de annotaties van de eerste corrector tijd kost. Deze extra tijdsinvestering zou zich echter weer terug kunnen verdienen, bijvoorbeeld als de annotaties de derde corrector in staat zouden stellen om sneller tot een gefundeerde beslissing over de score te komen. Een alternatieve verklaring stelt dat de annotaties de derde corrector ertoe verleiden te vroeg te kiezen voor de interpretatie en score van de eerste corrector. Annotaties zouden dan bijdragen tot het ontstaan van een tunnelvisie waarbij de eerste corrector ertoe neigt om belangrijke aanwijzingen die niet bij de score of interpretatie van de eerste corrector passen over het hoofd te zien. Nader onderzoek lijkt hier geboden, bijvoorbeeld in de vorm van cognitieve interviews met docenten tijdens de examenperiode.

De invloed van de tweede correctie op de score voor het centraal schriftelijk examen

Het onderzoek heeft ook gegevens opgeleverd over het functioneren van de tweede correctie. De desbetreffende onderzoeksvraag luidde: 'In hoeverre verschilt de gezamenlijk tussen de eerste en tweede corrector overeengekomen score van die van de eerste corrector?' We verwachtten dat de gezamenlijke score lager zou zijn dan de score van de eerste corrector. De redenering was dat eerste correctoren eerder geneigd zijn tot het geven van te soepele dan te strenge correcties en dat de huidige tweede correctie de neiging tot het geven van overwegend te soepele correcties bestrijdt. Geheel tegen de verwachtingen in blijkt de gezamenlijke score niet noemenswaard af te wijken van de score van de eerste corrector. Voor de ogenschijnlijk geringe invloed van de tweede correctie zijn ten minste twee verklaringen te geven. Een eerste verklaring stelt dat er weinig aanleiding is om de scores van de eerste corrector in twijfel te trekken (VO-raad, 2008a) en dat eerste en tweede corrector het daarom vrijwel altijd volledig met elkaar eens zijn. Een tweede verklaring is dat de eerste en/of tweede correctie tekort schieten (Inspectie van het Onderwijs, 2006, 2008a,b). De opzet van het onderzoek laat het niet toe om een onderscheid tussen beide verklaringen te maken. Daartoe is nader onderzoek nodig, bijvoorbeeld naar de wijze waarop de eerste en tweede correctie tijdens de examenperiode worden uitgevoerd.

De invloed van het schoolexamencijfer op de correctie

In het onderzoek is ook onderzocht in hoeverre eerste correctoren bij het nakijken rekening houden met het cijfer op het schoolexamen. Kandidaten met onvoldoendes op het schoolexamen zouden

meer van de soepelheid van de eerste corrector kunnen profiteren dan hun klasgenoten met hogere cijfers. Een voorbeeld hiervan is als eerste correctoren de interpretatieruimte die het correctievoorschrift biedt wat sterker zouden benutten bij kandidaten in de gevarenzone dan bij 'veilige' kandidaten. De desbetreffende onderzoeksvraag luidde 'In hoeverre is de samenhang tussen de correcties en het schoolexamencijfer voor kandidaten met cijfers onder of rond de grens tussen voldoende en onvoldoende anders dan voor degenen met hogere SE-cijfers?' Deze verwachting kon niet worden bevestigd. Van het veronderstelde kromlijnige verband tussen cijfers en correcties bleek geen sprake. Eerste correctoren blijken leerlingen onder of rond de grens tussen voldoende en voldoende niet soepeler te beoordelen dan hun klasgenoten met hogere cijfers voor het schoolexamen. Hierbij zij aangetekend dat de gegevens alleen betrekking hebben op het eerste tijdvak (waaraan nagenoeg alle kandidaten deelnemen). Bij het tweede tijdvak, dat alleen bestemd is voor inhalers en herkansers, staat er naar verhouding meer op het spel en zou het schoolexamencijfer wellicht wel een rol kunnen spelen.

1. Inleiding

In de centrale schriftelijke examens wordt het examenwerk eerst door de eigen docent gecorrigeerd en vervolgens door een docent van een andere school. Eerste correctoren krijgen daarbij het advies om het examenwerk van annotaties te voorzien. Volgens de richtlijnen van de VO-raad (2009) dient de eerste corrector duidelijk aan te geven waar in het werk de onvolkomenheden zitten en de toegekende punten per vraag voor de linker kantlijn van het examenwerk te noteren. Het corrigeren van examenwerk is een complex proces waarbij correctoren verschillende strategieën toepassen (Suto & Greateorex, 2008). Annotaties kunnen hierbij verschillende functies vervullen. Een belangrijk onderscheid is dat tussen persoonlijke annotaties die een corrector vooral voor zichzelf maakt en publieke annotaties die hij of zij maakt voor andere correctoren (Wolfe & Neuwirth, 2001). Experimenteel onderzoek naar het gebruik van annotaties voor persoonlijke doeleinden heeft aangetoond dat het annoteren tijdens het lezen het begrip van de tekst kan bevorderen (o.a. O'Hare & Sellen, 1997). De publieke functie van annotaties in een examencontext is communicatief van aard. Eerste correctoren annoteren om uit te leggen hoe de score tot stand is gekomen en de genomen beslissingen voor andere correctoren te rechtvaardigen. Inhoudsanalyse van examenwerk laat zien dat correctoren verschillen in het aantal en de aard van de aangebrachte annotaties, terwijl de annotatieprofielen ook verschillen tussen vraagtypen en examenvakken laten zien (Crisp & Johnson, 2005).

Over de vraag of het wenselijk is dat de tweede corrector kennis neemt van de annotaties van de eerste corrector zijn de meningen verdeeld. Annotatie maakt het corrigeren makkelijker omdat het de cognitieve belasting van de correctietaak vermindert. Correctoren hebben zelf het idee dat annoteren hen helpt bij het toepassen van de evaluatiecriteria en dat het de objectiviteit van de beoordeling ten goede komt (Bramley & Pollitt, 1996). Het beschikken over de annotaties van de eerste corrector kan tweede correctoren steun bieden, zeker in situaties waarin de vraagstelling van de examenopgave open is en het correctievoorschrift weinig steun biedt. Pilliner (1965) is daarentegen van mening dat kennis van de scores en aantekeningen van de eerste corrector de onafhankelijkheid van de tweede correctie in gevaar kan brengen. McVey (1975) stelt zelfs dat de scores van de tweede corrector 'are not really his own - they are those of the first scrutineer, slightly modified' (p. 212). Murphy (1979) is van mening dat annotaties van de eerste corrector omwille van de objectiviteit verwijderd moeten worden. Volgens Massey en Foulkes (1994) is onafhankelijke correctie echter geen deugd op zich, maar slechts een middel om een zo eerlijk mogelijke beoordeling voor iedereen te realiseren. Tweede correctoren kunnen hun voordeel doen met de extra informatie die annotaties bieden. Een score die via bemiddeling (Engels: *reconciling; consensus approach*) tot stand komt, is immers beter dan de gemiddelde score van twee volledig onafhankelijke correctoren.

In het vervolg van deze inleiding geven we eerst een overzicht van eerder verricht onderzoek naar de invloed van annotaties op de kwaliteit van de tweede correctie. Daarna gaan we in op de beleidsmatige context waarin het onderzoek is uitgevoerd. Vervolgens bespreken we de onderzoeksvragen, de verwachtingen met betrekking tot de resultaten en de potentiële relevantie van het onderzoek voor de praktijk van de centraal schriftelijke examinering.

Verricht onderzoek naar de invloed van annotaties

In het Verenigd Koninkrijk is onderzoek gedaan naar de invloed van annotaties op de kwaliteit van de tweede correctie (Murphy, 1979; Wilmot, 1984; WJEC, 2004; Fearnley, 2005; Meadows & Baird, 2005; Newton, 1996; Vidal Rodeiro, 2005; Shaw & Imam, 2008). Voor zover uit de doorgegaan summier rapportages valt op te maken geeft het verrichte onderzoek een relatief eenduidig beeld te zien. De resultaten van het onderzoek in het Verenigd Koninkrijk zijn echter niet zonder meer te vertalen naar de Nederlandse examencontext. Vandaar dat wij eerst twee wezenlijke verschillen tussen beide examensystemen bespreken alvorens de resultaten van de literatuurstudie te bespreken.

Verschillen tussen het Nederlandse en Engelse examensysteem

Een eerste verschil is dat de eerste correctie in Nederland wordt uitgevoerd door de 'eigen' docent, terwijl dat in het Verenigd Koninkrijk een anonieme externe corrector is die onbekend is met de school, de docent en de kandidaat (Kuhlemeier, 2007; Kuhlemeier, Van Weeren & Van der Werf, 2006). In diverse beoordelingscontexten is aangetoond dat beoordelaars ertoe neigen soepeler te beoordelen en minder spreiding aan te brengen als zij degenen die beoordeeld worden persoonlijk kennen (Albanese, 2000; Guilford, 1936; Landy & Farr, 1980; Murphy & Cleveland, 1995; Symonds, 1931; Thorndike, 1920). Zo blijken praktijkdocenten in het vmbo hun eigen leerlingen bij bijna tachtig procent van de centraal schriftelijke en praktische examens (CSPE) te soepel te beoordelen (Inspectie van het onderwijs, 2009a,b). Naar eigen zeggen doen de praktijkdocenten dat om kandidaten het voordeel van de twijfel te geven, om rekening te houden met persoonlijke omstandigheden (bijvoorbeeld in de thuissituatie), om rekening te houden met de manier waarop het onderwerp of de vaardigheid in de lessen behandeld is of om inzet of goede bedoelingen te belonen. In het Onderwijsverslag 2008-2009 concludeert de Inspectie dat de huidige examenpraktijk van het CSPE niet door de beugel kan, dat de regels onvoldoende worden nageleefd, dat scholen te sterk verschillen in de manier waarop zij het examen afnemen en beoordelen en dat de rechtsgelijkheid van kandidaten onvoldoende gewaarborgd is (Inspectie van het Onderwijs, 2008a,b). Ook bij de algemene vakken in het vmbo zijn er inmiddels aanwijzingen dat eerste correctoren hun eigen leerlingen vaak soepeler beoordelen dan correctoren die onbekend zijn met de te beoordelen kandidaten. Omdat de 'eigen' docent in het Verenigd Koninkrijk doorgaans niet bij de formele totstandkoming van de uiteindelijke score betrokken is, is de vraag naar de soepelheid of strengheid van de correctie daar veel minder beleidsrelevant dan bij ons. Het wekt dan ook geen verbazing dat verschillen tussen de eerste en tweede correctie in de soepelheid of strengheid van de correctie in het Verenigd Koninkrijk vaak niet gerapporteerd worden (men kijkt doorgaans alleen naar gemiddelde absolute verschillen).

Een tweede verschil is dat de tweede correctie in het Verenigd Koninkrijk wordt uitgevoerd door een getrainde *assistant examiner* (AE) die werkt onder supervisie van een *senior examiner* (SE). De SE is verantwoordelijk voor de standaardisatie van de AE's die daartoe een steekproef uit de eerste correcties van de AE's corrigeert en de scores zo nodig aanpast. Het score van de SE wordt opgevat als een 'ware' score waarmee de scores van de AE's worden vergeleken. In Nederland wordt de 'moderatie' uitgevoerd door een docent van een andere school. Beiden hebben geen training ontvangen en worden niet extern gesuperviseerd. Overigens wordt de tweede correctie door de SE in het Verenigd Koninkrijk geen tweede correctie (*double marking*) genoemd. De term tweede correctie is gereserveerd voor de situatie waarin een tweede AE het examenwerk van de eerste AE opnieuw corrigeert. In Engeland betreft de kwestie van de zinvolheid van tweede correctie dus vooral de vraag of het examenwerk behalve door een eerste AE (integraal) en een SE (steekproefsgewijs) ook nog eens door een tweede AE integraal gecorrigeerd zou moeten worden.

Overzicht van onderzoek naar de invloed van annotaties

In een experimentele opzet onderzocht Murphy (1979) het effect van annotaties op de betrouwbaarheid van de tweede correctie. Hij liet twee ervaren GCE-examinatoren een steekproef van tweehonderd GCE essays nogmaals corrigeren. Van honderd van deze tweehonderd examenwerken waren de punten en aantekeningen van de eerste examiner verwijderd. De correlaties met de eerste corrector bleken voor de geannoteerde examenwerken aanzienlijk hoger dan voor de geschoonde examenwerken (namelijk .94 versus .86 voor de eerste tweede corrector en .96 versus .85 voor de tweede tweede corrector). De correlaties tussen de beide tweede correctoren gaven een vergelijkbaar patroon te zien (namelijk .94 voor geannoteerde werken versus .87 voor geschoonde werken). Murphy concludeert hieruit dat de punten en aantekeningen van de eerste corrector tot een hogere overeenstemming tussen correctoren hebben geleid (zowel tussen de eerste en tweede corrector als tussen de tweede correctoren onderling). Het gemiddelde absolute verschil tussen de scores was kleiner voor geannoteerde examenwerken dan voor geschoonde examenwerken (waarbij het gemiddelde absolute verschil werd berekend door het absolute verschil in punten over alle kandidaten op te tellen en de uitkomst daarvan te delen door het aantal kandidaten). Volgens de auteur neigen tweede correctoren ertoe hun scores aan te passen aan die van de eerste corrector, ook al doen ze

nog zo hun best om de annotaties van de eerste corrector te negeren en hun eigen oordeel te vellen. Als de tweede correctie een onafhankelijk en onbevooroordeeld oordeel over de kwaliteit van een examenwerk moet opleveren, dient men de examenwerken volgens Murphy dan ook te zuiveren van de annotaties van de eerste corrector.

Annotaties bij examenvragen kunnen bestaan uit getalsmatige en/of tekstuele annotaties. Wilmut (1984) veronderstelde dat getalsmatige annotaties een ander effect op de tweede corrector hebben dan tekstuele annotaties. Kennis van de door de eerste corrector toegekende punten zou tweede correctoren ertoe kunnen bewegen de eigen scores meer in overeenstemming met die van de ander te brengen. Tekstuele annotaties bepalen vooral op welke aspecten van het examenwerk een tweede corrector zal letten. Een opmerking van de eerste corrector kan de tweede corrector bijvoorbeeld attent maken op een kenmerk van het examenwerk dat hij of zij anders gemist zou hebben. Als de eerste corrector iets belangrijks over het hoofd ziet, is de kans groter dat de tweede corrector dat ook doet. In een kleinschalige studie vergeleek Wilmut (1984) het effect van drie soorten annotaties: examenwerk met de punten en opmerkingen, hetzelfde werk met alleen opmerkingen en hetzelfde examenwerk helemaal zonder annotaties. De verwachting was dat de samenhang tussen eerste en tweede correctoren voor volledig geannoteerde examenwerken hoger zou zijn dan voor gedeeltelijk geannoteerde werken die op zijn beurt weer hoger zou zijn dan voor blanco examenwerken. Het verschil tussen de drie condities in de gemiddelde score was niet significant. De correlaties tussen enerzijds de originele scores en anderzijds de condities 'punten en opmerkingen', 'alleen opmerkingen' en 'noch punten noch opmerkingen' vertoonden echter wel het verwachte patroon, te weten respectievelijk .97, .95 en .93. Dit suggereerde volgens Wilmut dat getalsmatige en tekstuele annotaties ieder een afzonderlijke bijdrage leveren.

Meadows & Baird (2005; aangehaald in Fearnley, 2005) onderzochten verschillen tussen eerste correcties van *assistant examiners* (AE) en tweede correcties van *senior examiners* (SE). Voor geannoteerde werken waren de discrepanties tussen de scores van SE's en AE's kleiner dan voor de blanco werken. Hieruit blijkt volgens de auteurs dat de annotaties van de eerste corrector de correcties van tweede correctoren kunnen vertroebelen.

In een eerste experiment met elektronische correctie van volledig geschoonde examenwerken vond het *Welsh Joint Education Committee (WJEC)* (2004, aangehaald in Meadows & Billington, 2005) uitzonderlijk grote verschillen tussen *principal examiners* en *assistant examiners*, ondanks dat het ging om een relatief recht-toe-recht-aan examen met een tamelijk objectief correctievoorschrift. De onderzoekers schrijven deze grote verschillen toe aan het gegeven dat de PE's bij wijze van uitzondering blind corrigeerden (dit wil zeggen: zonder de annotaties van de *assistant examiners* in de correctie te kunnen betrekken).

Fearnley (2005) deed onderzoek naar het effect van geannoteerde versus blanco examenwerken bij de GCSE vakken Engels en *Business Studies*. Per vak waren er 32 AE's waarvan er 16 geannoteerde en 16 blanco examenwerken beoordeelden. Daarnaast was er een SE die alle geannoteerde werken corrigeerde en een andere SE die de blanco werken corrigeerde (de zogeheten ware scores waarmee de scores van de AE's werden vergeleken). Per vak werden honderd examenwerken in het onderzoek betrokken. Voor geannoteerde examenwerken Engels bleken de scores van twee van de acht paren correctoren significant van elkaar te verschillen, terwijl dat voor de blanco werken voor vijf van de acht paren gold. Voor het examenvak *Business Studies* waren de overeenkomstige aantallen één van de zeven paren voor geannoteerde werken en vijf van de zeven voor blanco werken. Voor beide vakken waren de correlaties tussen de scores van correctoren voor geannoteerde werken hoger dan voor blanco werken. Dit gold zowel voor de SE's als de AE's. Gezien de opzet van het onderzoek ligt het voor de hand de kleinere verschillen en de hogere correlaties tussen correctoren in de 'geannoteerde' conditie toe te schrijven aan de kennis die de 'second instance markers' hadden van de punten van de eerdere correctie. Fearnley concludeert dat tweede correctie van geannoteerde examenwerken een bias in de richting van de scores van de eerste corrector veroorzaakt die afbreuk doet aan de effectiviteit van de tweede correctie en dat de onderzoeksresultaten reden geven tot twijfel aan de validiteit van het gebruik van geannoteerde examenwerken in de context van de tweede correctie.

Newton (1996) onderzocht het effect van het gedeeltelijk verwijderen van de annotaties van de eerste corrector op de betrouwbaarheid van de tweede correctie van examenwerk (dit wil zeggen: alle

getalsmatige annotaties werden verwijderd, maar de overige aantekeningen bleven intact). Hoewel de bevindingen statistisch gezien niet significant waren, was het patroon wel vergelijkbaar met dat van Wilmut (1984). Het gemiddelde verschil tussen de scores van de eerste en tweede correctie was groter voor de volledig geschoonde werken dan voor de werken waarvan alleen de punten verwijderd waren.

Vidal Rodeiro (2005) deed onderzoek naar het effect van annotaties op de tweede correctie van de GCSE examens Engels en Grieks als klassieke taal. Het ging om tweehonderd examenwerken per examen waarvan honderd geannoteerde en honderd blanco werken. Voor elk vak namen een *principal examiner* (PE), twee *senior examiners* (SE) en twee *assistant examiners* (AE) aan het onderzoek deel. De SE's en AE's corrigeerden dezelfde examenwerken waarvan honderd geannoteerd en honderd blanco. Als het verschil tussen de scores van de eerste en de tweede corrector groter was dan tien procent van de maximumscore, werd een *reconciliation procedure* uitgevoerd onder leiding van de PE (resultierend in een zogeheten *adjudication mark*). De effectmaten waren de correlatiecoëfficiënt, de proportie overeenstemming en coëfficiënt *Kappa*. Het gemiddelde absolute verschil tussen de scores van de eerste en tweede correctie was kleiner voor de geannoteerde dan voor de blanco examenwerken. Bij het examen Grieks was het gemiddelde absolute verschil tussen eerste corrector en PE slechts .69 punt voor blanco werken versus 1.97 punten voor geannoteerde werken (op een schaal van 1 t/m 60); bij het examen Engels ging het om respectievelijk 1.98 voor geannoteerde werken en 5.64 punten voor blanco werken (op een schaal van 1 t/m 40). De percentages exacte overeenstemming tussen eerste corrector en SE waren bij Grieks 17% voor blanco werken en 50% voor geannoteerde werken en bij Engels waren de overeenkomstige percentages 3% en 38%.

Verder bleek de correctietijd van vijf à zes werken per uur voor beide vakken gelijk. Daarbij maakte het niet uit of de werken nog de punten en aantekeningen van de eerste corrector bevatten. De onderzoeker merkt op dat de hogere overeenstemming voor geannoteerde dan voor blanco werken op twee manieren geïnterpreteerd kan worden. Volgens de negatieve interpretatie laten tweede examinatoren zich onvermijdelijk door de annotaties van de eerste corrector beïnvloeden, hoezeer zij ook hun best doen om deze annotaties te negeren en hun eigen oordeel te vellen (waardoor zij fouten van de eerste corrector ten onrechte over het hoofd zien). Volgens de positieve interpretatie bieden annotaties tweede correctoren informatie over de totstandkoming van de eerdere score en zouden zij ook als zij blind gecorrigeerd hadden tot dezelfde score gekomen zijn.

In een experimentele opzet vergeleken Shaw en Imam (2008) het effect van een uitgebreid en een vereenvoudigd annotatiesysteem. Beide vormen van annotatie werden zowel op papieren als elektronische examenwerken toegepast. In alle vier condities werd van geschoonde examenwerken gebruik gemaakt. De uitgebreidheid van de annotaties bleek van belang voor de examenscores, terwijl het medium (papier versus elektronisch) er niet toe deed. De overeenstemming tussen de correctoren was iets hoger voor uitgebreide dan voor vereenvoudigde annotatie.

Beleidsmatige context

In de Nederlandse centrale schriftelijke examens wordt eerste correctoren geadviseerd het examenwerk van de kandidaat van annotaties te voorzien. Volgens de richtlijnen voor de eerste en tweede corrector dient de eerste corrector duidelijk aan te geven waar in het werk de onvolkomenheden zitten en de toegekende punten per vraag voor de linker kantlijn van het examenwerk te noteren (VO-raad, 2009). Net als in het Verenigd Koninkrijk kan niet iedereen zich echter in dit advies vinden. In de jaren zeventig van de vorige eeuw toonde het Cito zich al een voorstander van 'onafhankelijke' correctie (Luijten, 1997). De eerste en tweede docenten mochten niet op de hoogte zijn van elkaars bevindingen. De correctie moest zo anoniem mogelijk plaatsvinden in de zin dat van de leerling alleen het kandidaatnummer op het examenwerk vermeld mocht worden. Kremers (in Douma, 2008) is van mening dat de annotaties van de eerste corrector de objectiviteit van de tweede correctie in gevaar kunnen brengen. Vandaar dat hij pleit voor 'onafhankelijke' tweede correctie van blanco in plaats van geannoteerde examenwerken. Hij verwijst daarbij naar het Engelse examensysteem waarbij de examenwerken eerst naar een 'kopieercentrum' worden verzonden. Daar worden de werken gescand en digitaal naar de correctoren gestuurd. Zo krijgen alle correctoren een

blanco examen onder ogen. Het Cito wil dat hiermee voor enkele vakken een proef wordt gedaan. De VO-raad heeft laten weten deze werkwijze interessant te vinden. Tweede correctoren zouden namelijk eerder aan het werk kunnen omdat ze niet meer hoeven te wachten tot de eigen docent klaar is. In haar brief aan de Tweede Kamer d.d. 23 oktober 2008 zegt de staatssecretaris te willen nagaan of het systeem van eerste en tweede correctie effectiever kan worden ingericht. Tevens zegt zij in deze brief blij te zijn met het initiatief van de VO-raad en Cito om een pilot te starten met het scannen van blanco examenwerk, waardoor eerste en tweede corrector niet op de hoogte zijn van elkaars correcties. Alvorens definitief tot een dergelijke pilot te besluiten, wordt echter eerst een onderzoek uitgevoerd naar de veronderstelde meerwaarde van tweede correctie van blanco examenwerk ten opzichte van geannoteerd examenwerk. Van dit onderzoek wordt in deze publicatie verslag gedaan. In het onderzoek heeft een steekproef van correctoren een steekproef van examenwerken buiten de examenperiode opnieuw nagekeken. De zogeheten derde correctoren keken zowel geannoteerde als blanco examenwerken na. De vergelijking tussen geannoteerde en blanco derde correctie biedt inzicht in de toegevoegde waarde van een tweede correctie zonder de annotaties van de eerste corrector.

In de Nederlandse centrale schriftelijke examens wordt het werk van de kandidaten eerst door de eigen docent nagekeken en vervolgens door een docent van een andere school. Deze tweede corrector is aangewezen door de IBG (Informatie Beheer Groep). Kandidaten hebben recht op een rechtvaardige beoordeling. Idealiter zou het niet uit mogen maken wie het examen nakijkt. In de praktijk zijn er echter verschillen tussen eerste en tweede correctoren (Algra, 2004). 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 zwakgebod 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. Volgens het Examenbesluit moeten tweede correctoren het werk van de kandidaten integraal nakijken, alsof het gaat om een eerste correctie. Dit wil zeggen dat de tweede corrector alle werken nakijkt en per kandidaat het volledige examenwerk met alle opgaven. In het Onderwijsverslag 2004-2005 (Inspectie van het Onderwijs, 2006) constateert de Inspectie dat lang niet alle correctoren het examenwerk zorgvuldig genoeg nakijken. Een op de drie tweede correctoren is van mening dat de eerste corrector tekortschiet. Veertig procent van de tweede correctoren ontdekt dat er fouten zijn gemaakt tijdens de eerste correctieronde. Maar ook de tweede correctoren laten steken vallen, ook zij kijken het werk niet zorgvuldig genoeg na. Als een eerste corrector wordt gecorrigeerd door een tweede corrector die onvoldoende nauwkeurig nakijkt dan is de kans groot dat het examenwerk niet naar behoren wordt beoordeeld. De kandidaten krijgen in dat geval niet het cijfer dat ze verdienen. De Inspectie concludeert dat 'gezien het percentage correctoren dat fouten maakt en onvoldoende nauwkeurig alle opgaven controleert, kan het niet anders dan dat een aantal kandidaten een niet correcte eindscore krijgt (p. 339)'. In de brief aan de Tweede Kamer d.d. 23 oktober 2008 onderstreept de staatssecretaris nogmaals het belang van een volledige tweede correctie (met als argument dat de leerling recht heeft op een objectieve en professionele correctie). Voor lang niet iedereen staat de noodzaak van integrale tweede correctie echter vast. De VO-raad is bijvoorbeeld van mening dat docenten de examens tijdens de eerste correctieronde goed nakijken en dat het zeer goed mogelijk is dat een goed uitgevoerde steekproef voldoende garantie biedt (VO-raad, 2008a). Ook volgens het Platform Vakinhoudelijke Verenigingen Voortgezet Onderwijs (VVVO) is steekproefsgewijze tweede correctie in de gegeven situatie de beste oplossing (Platform VVVO, 2008a). Op basis van eigen onderzoek wijst zij op de hoge taakbelasting die integrale tweede correctie voor docenten met zich meebrengt; docenten blijken de integrale tweede correctie alleen ten koste van hun eigen vrije tijd en nachtrust te kunnen uitvoeren (Platform VVVO, 2008b). Als de overheid aan een integrale tweede

correctie wil vasthouden, aldus het Platform VVVO, zal de voor de correctie benodigde tijd 'zichtbaar en geormerkt in de taakbelasting van de betrokken docenten opgenomen moeten worden (p. 1)'. In het onderhavige onderzoek keken de derde correctoren het examenwerk integraal na, dit wil zeggen dat zij alle opgaven van alle kandidaten corrigeerden. Behalve een redelijke vergoeding kregen de derde correctoren ruim de tijd om de correctie uit te voeren. De vergelijking van de eerste met de derde correctie geeft meer informatie over overeenkomsten en verschillen tussen de huidige eerste correctie en een integrale tweede correctie (op zowel geannoteerd als blanco materiaal). In het onderhavige onderzoek is de correctoren ook gevraagd de correctietijd per examenwerk nauwgezet te registreren. Hiermee kan worden vastgesteld welke tijdsinvestering gemoeid is met integrale tweede correctie en of geannoteerde werken meer, evenveel of minder correctietijd vergen dan blanco werken.

Onderzoeksvragen en verwachtingen

In het onderzoek is een aantal al door de eerste en tweede corrector nagekeken examenwerken opnieuw gecorrigeerd. Deze hernieuwde correctie is uitgevoerd door zogeheten derde correctoren. Daarbij is van elk examenwerk zowel een geannoteerde als een blanco versie nagekeken. Deze opzet stelt ons in staat de volgende vier correctieprocedures met elkaar te vergelijken:

- A. Derde correctie van geannoteerde examenwerken
- B. Derde correctie van blanco examenwerken
- C. Eerste correctie door de eigen docent
- D. De gezamenlijk tussen eerste en tweede corrector overeengekomen correctie

Het onderzoek heeft tot doel een antwoord te geven op zes hieronder gepresenteerde onderzoeksvragen. Op basis van de in de inleiding besproken literatuur formuleren we daarbij ook enkele verwachtingen met betrekking tot de uitkomsten van het onderzoek.

- 1 In hoeverre is er een verschil tussen de scores van geannoteerde en blanco derde correctie?
Verwachtingen. We verwachten een hoge samenhang tussen geannoteerde en blanco derde correcties, maar niet een zodanig hoge samenhang dat de scores volledig uitwisselbaar zijn. Correctoren kunnen hetzelfde antwoord immers verschillend interpreteren en met punten waarderen. De toegekende scores zullen ook verschillen als de ene derde corrector zich zou laten beïnvloeden door de annotaties op het examenwerk (terwijl dat voor de andere corrector die de blanco versie nakijkt natuurlijk niet mogelijk is). De annotaties op het examenwerk kunnen er immers toe bijdragen dat derde correctoren hun scores aanpassen aan die van de eerste corrector. Daarnaast verwachten we dat geannoteerde derde correctie tot hogere scores zal leiden dan blanco derde correctie. De annotaties kunnen derde correctoren er immers toe verleiden mee te gaan met de wellicht wat toegeeflijke correcties van de eerste corrector.
- 2 In hoeverre is de samenhang tussen eerste en derde correctie voor geannoteerde werken anders dan voor blanco werken?
Verwachtingen. De verwachting is dat eerste correctie sterker zal samenhangen met geannoteerde dan met blanco derde correctie. De motivering is analoog aan die bij de eerste onderzoeksvraag. Op grond van het in de inleiding besproken onderzoek verwachten we tevens dat derde correctie over het algemeen tot lagere scores zal leiden dan eerste correctie. Immers, anders dan eerste correctoren kennen derde correctoren de kandidaten niet persoonlijk en hebben zij geen persoonlijk belang bij te soepele (of te strenge) correcties.
- 3 In hoeverre is de overeenstemming tussen de scores van verschillende correctieprocedures afhankelijk van het opgaventype?
Verwachtingen. De antwoordmodellen bij open vragen bieden de correctoren doorgaans minder steun bij het nakijken dan de correctievoorschriften bij meerkeuzevragen (die objectief scorebaar zijn). Open vragen leiden naar verwachting tot meer onzekerheid bij de corrector dan gesloten vragen. Daardoor zal hij of zij eerder geneigd zijn zich door externe informatie te laten leiden, zoals

de annotaties van de eerste corrector. We verwachten dan ook dat de overeenstemming tussen eerste en derde correctie bij open vragen kleiner zal zijn dan bij gesloten vragen. Daarnaast verwachten we dat dit verschil in overeenstemming tussen eerste en derde correctoren in het nadeel van open vragen voor geannoteerde examenwerken minder groot zal zijn dan voor blanco examenwerken. De annotaties van de eerste corrector kunnen immers tot een hogere overeenstemming tussen eerste en derde corrector leiden.

- 4 In hoeverre is er een verschil in de gemiddelde correctietijd tussen geannoteerde en blanco examenwerken?

Verwachtingen. Ten aanzien van de correctietijd van geannoteerde en blanco werken staan wij in dubio. Van de ene kant zou het nakijken van geannoteerd werk meer tijd kunnen kosten dan dat van blanco werk. Met het lezen en interpreteren van de annotaties van de eerste corrector is nu eenmaal tijd gemoeid. Van de andere kant zou deze extra tijdsinvestering zich weer kunnen terugverdienen, bijvoorbeeld als de derde corrector door kennis te nemen van de annotaties sneller tot een gefundeerde beslissing over de score zou komen.

- 5 In hoeverre verschilt de gezamenlijk tussen de eerste en tweede corrector overeengekomen score van die van de eerste corrector?

Verwachtingen. In het onderzoek hebben de scholen ook gegevens over de tweede correctie verstrekt, te weten de totaalscore van de tweede corrector en de gezamenlijk tussen de eerste en tweede corrector overeengekomen totaalscore. Anders dan de score van de tweede corrector bleek de gezamenlijk overeengekomen score voor vrijwel alle kandidaten beschikbaar. Met betrekking tot de gezamenlijke score verwachten we dat deze lager zal zijn dan de score van de eerste corrector, maar hoger dan de beide derde correcties. De redenering is dat eerste correctoren eerder geneigd zijn tot het geven van te soepele dan te strenge correcties, dat de huidige tweede correctie de neiging tot het geven van overwegend te soepele correcties bestrijdt en dat integrale derde correctie met voldoende tijd dat in nog sterkere mate zal doen.

- 6 In hoeverre is de samenhang tussen de correcties en het schoolexamencijfer voor kandidaten met cijfers onder of rond de grens tussen voldoende en onvoldoende anders dan voor degenen met hogere SE-cijfers?

Verwachtingen. De verwachting is dat kandidaten die op het schoolexamen zwak of op de grens tussen voldoende en onvoldoende presteren meer profiteren van het veronderstelde soepelheideffect dan hun hoog presterende klasgenoten. Een voorbeeld hiervan is als eerste correctoren de interpretatieruimte die het correctievoorschrift biedt wat sterker zouden benutten bij kandidaten in de gevarenzone dan bij 'veilige' kandidaten.

Potentiële relevantie

De potentiële relevantie van het onderzoek voor de huidige examenpraktijk kan als volgt worden aangeduid. De vergelijking tussen geannoteerde en blanco derde correctie biedt inzicht in de toegevoegde waarde van een tweede correctie zonder de annotaties van de eerste corrector. Het geeft een antwoord op de onderzoeksvraag of het schonen van examenwerk nog nodig is als tweede correctoren voldoende tijd zouden krijgen. De vergelijking van de eerste met de derde correctie geeft meer informatie over overeenkomsten en verschillen tussen de huidige eerste correctie en een integrale tweede correctie (op zowel geannoteerd als blanco materiaal). De vergelijking tussen eerste en gezamenlijke correctie geeft inzicht in de invloed van de huidige tweede correctie op de scores en cijfers voor het centraal schriftelijk examen.

2. Methode van onderzoek

2.1 Steekproef van vestigingen en kandidaten

Populatie

In het onderzoek zijn zes examenvakken betrokken, te weten Nederlands VWO, Engels GT, Wiskunde GT, Biologie VWO, Geschiedenis HAVO en Tehatex HAVO. In alle gevallen ging het om examens voor het schooljaar 2008-2009. Bij de steekproeftrekking is gebruik gemaakt van het populatiebestand van vestigingen voor voortgezet onderwijs die zich met een of meer kandidaten voor de genoemde examens hadden ingeschreven. Tabel 1 toont de verdeling van de vestigingen en het aantal ingeschreven kandidaten per examen in de populatie (na verwijdering van een klein aantal vestigingen met tien of minder ingeschreven kandidaten).

Tabel 1
Populatie, steekproef en response

Examen	Populatie		Steekproef		Response fase 1 en 2		Uiteindelijk gebruikt	
	Aantal vesti- gingen	Aantal kandi- daten	Aantal vesti- gingen	Aantal kandi- daten	Aantal vesti- gingen	Aantal werken	Aantal vesti- gingen	Aantal werken
Nederlands VWO	514	38674	80	6838	21	147	21	105
Engels GT	836	52682	80	4972	22	154	22	110
Wiskunde GT	819	44676	80	4263	37	259	22	110
Biologie VWO	513	16633	80	3033	25	174	23	115
Geschiedenis HAVO	502	33560	80	5341	19	130	19	95
Tehatex HAVO	216	5136	80	2137	20	138	20	100
Totaal		191361	480	26584	144	1002	127	635

Steekproef en response

De steekproef van examenkandidaten i.c. examenwerken is in twee fasen getrokken: eerst zijn vestigingen getrokken (fase 1) en vervolgens kandidaten i.c. examenwerken binnen vestigingen (fase 2). In beide gevallen was de trekking aselekt.

De verdeling van het aantal eindexamenkandidaten in de populatie is weergegeven in de eerste en tweede kolom van Tabel 1. Uit deze examenpopulatie is per examenvak een random steekproef van 80 vestigingen getrokken (voor de verdeling van de steekproef, zie kolom 3 en 4). De examensecretarissen van de 480 getrokken vestigingen is gevraagd het volledige overzicht met alle kandidaten van het desbetreffende examen op te sturen. Van de 480 benaderde examensecretarissen zegden 159 (33%) hun medewerking aan het onderzoek toe.

In fase 2 is uit de verstrekte lijst met alle kandidaten een aselechte steekproef van zeven kandidaten getrokken. De examenwerken van deze zeven kandidaten zijn vervolgens bij de vestigingen opgevraagd, samen met enkele aanvullende gegevens zoals de totaalscore van de eerste corrector, de totaalscore van de tweede corrector, de gezamenlijk overeengekomen totaalscore, de scores van de eerste corrector per opgave, het cijfer op het centraal schriftelijk examen en het cijfer op het schoolexamen. Van de 159 ingeschreven vestigingen hebben er 144 in totaal 1002 examenwerken daadwerkelijk toegestuurd (zie kolom 5 en 6 van Tabel 1).

Van de zeven opgevraagde examenwerken per vestiging zijn er telkens vijf daadwerkelijk in het onderzoek gebruikt. Van de zeven examenwerken werden steeds de eerste vijf gekozen, tenzij het handschrift van de kandidaat en/of de annotaties van de eerste corrector slecht leesbaar waren (zie voor meer informatie paragraaf 2.4).

Voor wiskunde en biologie zijn er meer examenwerken ontvangen dan voor het onderzoek strikt noodzakelijk was. Om budgettaire redenen zijn niet al deze examenwerken in het onderzoek betrokken. De verdeling van de uiteindelijk gebruikte examenwerken over de examenvakken en vestigingen is weergegeven in kolom 7 en 8 van Tabel 1. In totaal zijn 635 examenwerken van 127 vestigingen in het onderzoek betrokken.

2.2 Verzameling aanvullende gegevens

Behalve de zeven originele examenwerken is de vestiging gevraagd voor elke getrokken kandidaat de volgende aanvullende gegevens te verstrekken:

- De totaalscore van de eerste corrector;
- De totaalscore van de tweede corrector;
- De gezamenlijk tussen de eerste en tweede corrector overeengekomen totaalscore;
- Het overzicht met de scores per kandidaat per opgave;
- Het cijfer voor het centraal schriftelijk examen in het desbetreffende vak (eerste tijdvak);
- Het cijfer voor het schoolexamen in het desbetreffende vak.

Bovenstaande gegevens zijn overgenomen uit het door de vestigingen verstrekte materiaal. Dat bestond doorgaans uit de volgende overzichten:

- de Overzichtsscorelijst met de scores per opgave;
- het Formulier voor 1e en 2e correctie CSE en bepaling cijfer CE per lesgroep;
- door de scholen zelf ontwikkelde formulieren en overzichten, deels ter vervanging van en deels ter aanvulling op de *Overzichtsscorelijst* en/of het *Formulier 1e en 2e correctie*.

De *Overzichtsscorelijst* bevat de door de eerste corrector toegekende scores per kandidaat per opgave. Daarnaast bevat dit overzicht in veel gevallen de neerslag van het werk van de tweede corrector. Dat wil zeggen: de met de hand geschreven wijzigingen van de totaalscore van de kandidaten en/of de scores per opgave.

Het *Formulier voor de 1e en 2e correctie CSE en bepaling cijfer CE per lesgroep* biedt ruimte voor het noteren van de volgende gegevens:

- het schoolexamencijfer (kolom 3)
- de totaalscore van de eerste corrector (kolom 4a)
- de totaalscore van de tweede corrector (kolom 4b)
- de gezamenlijk tussen eerste en tweede corrector overeengekomen totaalscore (kolom 4c)
- het eindcijfer op het centraal schriftelijk examen (kolom 4d)
- het eindcijfer op het centraal examen (kolom 6)
- het eindcijfer voor het eerste tijdvak (kolom 7)

De voor deze studie gebruikte aanvullende gegevens zijn op de volgende wijze uit de genoemde formulieren overgenomen:

- De totaalscore van de eerste corrector is overgenomen uit kolom 4a van het Formulier 1e en 2e correctie. Was kolom 4a niet ingevuld, dan is de totaalscore van de eerste corrector overgenomen uit de *Overzichtsscorelijst* per opgave.
- De gezamenlijk tussen de eerste en tweede corrector overeengekomen score is overgenomen uit kolom 4c van het Formulier 1e en 2e correctie.
- Het cijfer voor het schoolexamen is overgenomen uit kolom 3 van het Formulier 1e en 2e correctie.

2.3 Werving van correctoren

De steekproef voor het werven van de correctoren is getrokken uit het bestand van vestigingen voor voortgezet onderwijs die zich hadden ingeschreven voor een of meer van de examens Nederlands VWO, Engels GT, Wiskunde GT, Biologie VWO, Geschiedenis HAVO en Tehatex HAVO voor het schooljaar 2008-2009 (zie kolom 1 en 2 van Tabel 1). De werving van correctoren verliep uitermate moeizaam.

Medio oktober zijn in eerste instantie in totaal zestig vestigingen benaderd (vijftien per examen). De wervingsbrief met bijbehorend inschrijfformulier is verzonden aan de sectieleider in het desbetreffende vak en opleidingstype. Gevraagd is naar docenten die het desbetreffende examen als eerste corrector hebben nagekeken en die bereid zijn een set van ongeveer veertig examenwerken voor onderzoeksdoeleinden te corrigeren. Deze eerste wervingsfase resulteert in slechts acht inschrijvingen, terwijl er in totaal dertig correctoren nodig waren.

Begin november is daarom een tweede steekproef van 225 vestigingen aangeschreven (samengesteld uit 15 vestigingen Nederlands, 50 Engels, 50 wiskunde, 50 biologie, 50 geschiedenis en 30 tehatex). Eind november resulteerde dit in voldoende correctoren voor Nederlands, Engels en wiskunde, maar voor biologie, geschiedenis en tehatex kwamen we er nog respectievelijk twee, drie en twee tekort. Voor deze drie vakken is nog een derde steekproef aangeschreven van 100 (biologie), 150 (geschiedenis) en 100 (tehatex) vestigingen. Deze derde werving resulteerde uiteindelijk in vijf correctoren per vak die bereid waren de examenwerken te corrigeren.

De te corrigeren examenwerken zijn een week voor de kerstvakantie 2009 aan de derde correctoren verzonden (samen met onder meer een handleiding voor het invullen van de scoreformulieren). Begin januari bleken vier correctoren zich om persoonlijke redenen te hebben teruggetrokken (twee voor wiskunde en twee voor tehatex). Voor wiskunde waren er vier "reservecorrectoren" beschikbaar. Geen van hen was echter in staat of bereid voor hun verhinderde collega in te springen. Via interne contacten bleek het mogelijk om zowel voor wiskunde als tehatex één vervangende corrector te vinden. Ten behoeve van de ontbrekende corrector wiskunde en tehatex moest echter opnieuw geworven worden. De inmiddels vierde wervingsbrief werd medio januari 2010 verzonden aan 135 sectieleiders tehatex en vijftig sectieleiders wiskunde. In de derde week van januari konden ook de laatste twee correctoren aan het werk.

Op een uitzondering hadden alle derde correctoren als eerste en/of tweede corrector ervaring met het toepassen van het correctievoorschrift in kwestie. De uitzondering was een gepensioneerd toetsdeskundige wiskunde die bereid was de plaats in te nemen van een van de correctoren wiskunde die zich had teruggetrokken.

2.4 Digitaliseren van de examenwerken

Per examenvak zijn van iedere deelnemende vestiging zoals gezegd zeven examenwerken opgevraagd. Van elk van deze zeven originele examenwerken is voor doeleinden van archivering een digitale kopie gemaakt nadat deze voorzien was van een viercijferige identificatiecode en een paginering. Deze kopie wordt hierna kopie A genoemd.

Van de zeven originele examenwerken zijn er vijf daadwerkelijk in het onderzoek gebruikt. Van de zeven examenwerken werden steeds de eerste vijf gekozen, tenzij het handschrift van de kandidaat en/of de annotaties van de eerste corrector slecht leesbaar waren. In dat geval is telkens het eerstvolgende examenwerk dat wel goed leesbaar was gekozen. Waren er drie of meer examenwerken met leesbaarheidsproblemen, dan werden de vijf best kopieerbare examenwerken gekozen.

Van de vijf geselecteerde gedigitaliseerde originelen is vervolgens een geanonimiseerde kopie gemaakt. Hierbij werd de originele kopie A als uitgangspunt genomen. Van deze digitale kopie werden de gegevens van de vestiging en de voor- en achternaam van de kandidaat digitaal verwijderd. De annotaties van de eerste corrector bleven staan. Deze kopie wordt hierna kopie B genoemd.

Van de geanonimiseerde kopieën B werden vervolgens de scorepunten en aantekeningen van de eerste corrector digitaal verwijderd. Deze geschoonde kopie noemen we hierna kopie C.

2.5 Toewijzing van kopieën B en C aan correctoren

De toewijzing is zodanig opgezet dat elke corrector van elke school evenveel werken nakijkt en van eenzelfde kandidaat nooit zowel kopie B als C corrigeert. Tabel 2 laat zien hoe de kopieën B en C aan de vijf correctoren zijn toegewezen (voor een ideaaltypische vestiging met vijf geselecteerde kandidaten).

Tabel 2
Toewijzing van kopieën aan correctoren

		Correctoren				
		1	2	3	4	5
School 1						
Kopie B						
	Leerling 1	X				
	Leerling 2		X			
	Leerling 3			X		
	Leerling 4				X	
	Leerling 5					X
Kopie C						
	Leerling 1		X			
	Leerling 2			X		
	Leerling 3				X	
	Leerling 4					X
	Leerling 5	X				

2.6 Correctie van het examenwerk

De correctie vond plaats aan de hand van de *Handleiding voor de corrector*. Behalve algemene informatie over het onderzoek bevatte deze handleiding aanwijzingen voor het corrigeren en het invullen van het scoreformulier.

In de sectie “Algemene informatie over het onderzoek” is de correctoren verteld dat het onderzoek tot doel heeft meer informatie te verkrijgen over de kwaliteit van de correctie en de tijd die met integrale correctie gemoeid is (d.w.z. het nakijken van alle opgaven van alle kandidaten). Daarnaast is hun verteld dat zij tot een steekproef van correctoren behoren die een steekproef van examenwerken zo objectief en zorgvuldig mogelijk maar zonder tijdsdruk corrigeren.

In de sectie “Aanwijzingen voor het corrigeren” is de docenten gevraagd de examenwerken met behulp van het bijgevoegde correctievoorschrift na te kijken. Verder is hun verteld dat dit voorschrift volledig identiek is aan het “officiële” correctievoorschrift dat zij vorig schooljaar gebruikten bij de correctie van het examen 2008-2009. De correctoren is gevraagd het correctievoorschrift voor het corrigeren nog een keer nauwkeurig door te kijken. De tijd die hiermee gemoeid is, konden zij bij het Cito declareren (met een maximum van twee uur).

De docenten is er expliciet op gewezen dat zij het examenwerk integraal moeten nakijken, dit wil zeggen alle opgaven van alle kandidaten. Verder is aangegeven dat het werk zo objectief en nauwkeurig mogelijk nagekeken dient te worden en dat iedere afwijking van het correctievoorschrift de

scores onbruikbaar maakt voor het onderzoek. Ook is hun verteld dat zij het examenwerk desgewenst van annotaties konden voorzien (als zij dat ook tijdens de correctie van het vorige examen gedaan hadden).

Met betrekking tot de correctietijd per examenwerk is de correctoren erop gewezen dat de registratie van de begin- en eindtijd niet bedoeld is om hen te controleren. Het noteren van de begin- en eindtijd heeft uitsluitend tot doel de zuivere correctietijd te bepalen, dit wil zeggen de correctietijd exclusief allerlei andere werkzaamheden (zoals het vooraf bestuderen van de handleiding en het correctievoorschrift en het retourneren van het materiaal).

Anders dan in de reguliere examenpraktijk hadden de correctoren niet de mogelijkheid om de cijfers van de kandidaten op het schoolexamen in te zien. Evenmin hadden ze de beschikking over het overzicht met de door de eerste corrector toegekende totaalscores en de scores per opgave.

2.7 Volledigheid van de verzamelde gegevens

De volgende tabel geeft inzicht in de volledigheid van de belangrijkste gegevens die verzameld zijn.

Tabel 3

Volledigheid van examenwerken en aanvullende gegevens (in vergelijking met het totale aantal examenwerken)

	Nederlands	Engels	Wiskunde	Biologie	Geschiedenis	Tehatex	Totaal
Examenwerken	105	110	110	100	95	100	620
Examenwerken met annotaties	105	101	105	85	75	95	566
Totaalscore eerste corrector	105	105	102	100	95	94	601
Totaalscore eerste en tweede corrector gezamenlijk	104	100	102	100	90	89	585
Cijfer schoolexamen	105	110	102	100	95	99	611
Geannoteerde derde correctie	105	110	107	100	95	99	616
Blanco derde correctie	105	110	107	100	95	99	616

De eerste rij van Tabel 3 vermeldt het aantal verzamelde examenwerken per vak. Bij biologie dient opgemerkt te worden dat het aantal van 100 werken lager is dan het eerder gerapporteerde aantal van 115 werken. De verklaring is dat drie vestigingen met in totaal vijftien kandidaten per abuis het examenwerk van het beeldschermexamen complex biologie 1,2 VWO hadden opgestuurd. Anders dan in het papieren examen biologie gebruiken de kandidaten bij ongeveer een derde van de opgaven de computer als hulpmiddel. De opgaven zelf staan echter op papier en ook de antwoorden worden op papier gegeven. Vanwege verschillen in het aantal en de aard van de vragen waren de vijftien complex-werken niet bruikbaar voor het onderzoek en zijn ze op het totale aantal verzamelde examenwerken in mindering gebracht. De derde correctoren hebben alle 620 verzamelde examenwerken nagekeken en 99% van de ingeleverde scoreformulieren bleek bruikbaar voor de statistische analyse.

Bij alle vakken behalve Nederlands waren er examenwerken die geen annotaties van de eerste corrector bevatten. In totaal ging het om negen werken Engels (8%), vijf werken wiskunde (5%), vijftien werken biologie (15%), twintig werken geschiedenis (22%) en vijf werken tehatex (5%). In totaal was 9% van de examenwerken 'blanco'. De derde correctoren ook de niet-geannoteerde

examenwerken nagekeken, maar de gegevens zijn niet gebruikt voor de vergelijking van de correctieprocedures. Ter controle is via een weg variantieanalyse nagegaan of de gemiddelden van de niet-geannoteerde werken verschillen van de overige werken. Deze analyse is afzonderlijk uitgevoerd voor de totaalscore van de eerste corrector, de gezamenlijke score en de scores van de beide derde correcties. Bij de vakken Engels, wiskunde, biologie en tehatex gaf geen van de uitgevoerde toetsingen significantie te zien. Bij deze vakken mogen we er dus vanuit gaan dat de voor de vergelijking van de correctieprocedures gebruikte examenwerken een representatieve afspiegeling vormen van de volledige set examenwerken. Bij geschiedenis bleken de niet-geannoteerde werken een hogere gemiddelde score te hebben voor de eerste correctie ($p = .019$) en de gezamenlijke correctie ($p = .030$), maar bij de geannoteerde en blanco derde correctie was het verschil niet significant ($p > .05$).

De totaalscore van de eerste corrector, de gezamenlijk tussen eerste en tweede corrector overeengekomen totaalscore en het cijfer op het schoolexamen konden na herhaaldelijk rappelleren in vrijwel alle gevallen verkregen worden (respectievelijk van 97%, 94% en 99% van de verzamelde examenwerken). Van een groot aantal examenkandidaten bleek de totaalscore van de tweede corrector te ontbreken. De daarvoor bestemde kolom op het *Formulier 1e en 2e correctie* was dan leeg. We hebben in dat geval geprobeerd de totaalscore van de tweede corrector over te nemen uit de *Overzichtsscorelijst per opgave*. In veel gevallen bood ook deze lijst geen soelaas. Er is dan eerst schriftelijk en daarna zo nodig telefonisch contact opgenomen met de desbetreffende vestiging met het verzoek de totaalscores van de tweede corrector alsnog te verstrekken (uiteraard alleen voor de zeven aan het onderzoek deelnemende kandidaten). Ook dit leidde helaas niet altijd tot het gewenste resultaat. Al met al ontbreekt de totaalscore van de tweede corrector voor 25% van de verzamelde examenwerken. Om deze reden is besloten deze score in het onderzoek verder niet te gebruiken.

Van de 30 correctoren hebben er 26 de correctietijd per examenwerk volgens de aanwijzingen in de handleiding geregistreerd. Uit het gegeven dat de ingevulde begin- en eindtijden in bijna alle gevallen logisch op elkaar aansloten en vrijwel alle registraties bruikbaar bleken, kan men afleiden dat de correctoren de correctietijd zorgvuldig hebben geregistreerd. Vier docenten hebben de examenwerken horizontaal in plaats van verticaal nagekeken (dit wil zeggen voor alle kandidaten eerst opgave 1, dan opgave 2, et cetera versus eerst alle opgaven van kandidaat 1, dan alle opgaven van kandidaat 2, enzovoorts). Omdat deze vier docenten horizontaal nakeken, konden zij de correctietijd natuurlijk niet per werk registreren. Aangezien horizontale correctie de voorkeur verdient boven verticale correctie, mochten zij volstaan met een schatting van de gemiddelde correctietijd. Een van hen heeft de tijdsbesteding daarbij waarschijnlijk behoorlijk overschat. De gerapporteerde gemiddelde correctietijd per examenwerk tehatex moet daarom met enige terughoudendheid geïnterpreteerd worden.

2.8 Representativiteit van de responsegroep

Het aantal kandidaten in het onderzoek is klein in vergelijking met dat in de examenpopulatie. Het is mogelijk dat vooral vestigingen met relatief hoge prestaties op het centraal examen zich voor het onderzoek hebben opgegeven. Vandaar dat is nagegaan in hoeverre de gemiddelde scores in de responsegroep vergelijkbaar zijn met die in de examenpopulatie. Een overzicht van de scores in de volledige examenpopulatie was op het moment van het schrijven van dit rapport niet beschikbaar. Vandaar dat we gebruik hebben gemaakt van de resultaten per opleidingstype en leerweg uit het Examenverslag (Alberts & Erens, 2009). Deze gegevens zijn afkomstig van de alfabetisch eerste vijf of soms tien kandidaten per vestiging. Aangezien het Cito deze gegevens gebruikt voor de normering mag worden aangenomen dat de representativiteit ten opzichte van de examenpopulatie gegarandeerd is. Tabel 4 toont de gemiddelde score en de standaarddeviatie in de normeringsgroep en de groep kandidaten i.c. examenwerken die uiteindelijk voor de rapportage in hoofdstuk 4 gebruikt zijn (d.w.z. na verwijdering van examenwerken zonder annotaties van de eerste corrector). Daarnaast is het gemiddelde cijfer op het centraal schriftelijk examen en het percentage onvoldoendes

weergegeven (voor het eerste tijdvak). Voor de responsegroep zijn deze laatste twee gegevens berekend door de scores van de eerste correctoren om te zetten in cijfers. Daarbij is rekening gehouden met de definitieve normeringstermen zoals die zijn vastgesteld door het College van Examens (publicatiedatum woensdag 17 juni 2009). Voor Nederlands, Engels en tehatex liggen de gemiddelden van de eerste correctoren in de responsegroep en de normeringsgroep niet ver uit elkaar. Voor wiskunde, biologie en geschiedenis zijn de gemiddelden van de eerste correctoren in het onderzoek duidelijk hoger dan in de normeringsgroep. Bij deze drie vakken lijken vestigingen met hoge scores op het centraal schriftelijk examen in de responsegroep oververtegenwoordigd.

Tabel 4

Gemiddelde prestatie en standaarddeviatie in de normeringsgroep en de responsegroep (eerste corrector eerste tijdvak)

Examen	Aantal kandidaten	Gemiddelde score	Standaard deviatie	Gemiddeld cijfer	Percentage onvoldoendes
Nederlands VWO					
Normeringsgroep	2369	29.10	6.20	6.30	23
Responsegroep	105	29.78	6.21	6.46	20
Engels GT					
Normeringsgroep	3908	28.90	6.50	6.40	21
Responsegroep	96	28.47	6.47	6.34	27
Wiskunde GT					
Normeringsgroep	3848	47.30	11.80	6.30	28
Responsegroep	102	49.65	10.70	6.56	23
Biologie VWO					
Normeringsgroep	2233	41.50	8.60	6.40	17
Responsegroep	85	44.32	8.64	6.76	15
Geschiedenis HAVO					
Normeringsgroep	2290	41.70	9.50	6.20	25
Responsegroep	75	44.31	10.33	6.54	23
Tehatex HAVO					
Normeringsgroep	4977	40.50	8.00	6.20	22
Responsegroep	89	41.94	8.55	6.83	22

2.9 Statistische analyse

De gegevens zijn geanalyseerd met behulp van het statistische pakket SPSS 17.0. Voor de analyse van de overeenstemming zijn, afhankelijk van het meetniveau van de variabele, verschillende analysetechnieken gebruikt.

Een eenvoudige maat voor de mate waarin verschillende correctoren of correctieprocedures tot dezelfde scores leiden is het percentage exacte overeenstemming. Deze maat houdt er echter geen rekening mee dat correctoren alleen op basis van toeval al tot op zekere hoogte met elkaar zullen overeenstemmen. Denk bijvoorbeeld aan een vier-keuzeopgave waarbij de kans dat een examenkandidaat het juiste alternatief kiest louter op basis van toeval al een kwart is. Een bekende maat die corrigeert voor toevallige overeenstemming is *Kappa* (Cohen, 1960). De mate van overeenstemming over de scores per opgave is bepaald met *Cohen's Kappa*. Deze maat is 1 als de correctoren perfect overeenstemmen en 0 als de overeenstemming niet groter is dan men op basis van toeval mag verwachten. Landis en Koch (1997) geven de volgende vuistregel voor het interpreteren van de hoogte van *Kappa*: .00 tot .20 is *slight*, .21 tot .40 is *fair*, .41 tot .60 is *moderate*, .61 tot .80 is *substantial* en .81 tot 1.00 is *almost perfect*. Als check op de kwaliteit van de

derde correctie hebben de correctoren per vak een set van vijf examenwerken onafhankelijk van elkaar gecorrigeerd. De overeenstemming tussen de vijf correctoren is bepaald met *Cohen's multi-rater Kappa*.

Een maat die laat zien

hoe sterk correctoren of correctieprocedures van elkaar afwijken is de gemiddelde absolute afwijking. Deze spreidingsmaat is in ons geval het gemiddelde van de absolute waarde van het verschil tussen de scores van de beide correctieprocedures (dat wil zeggen dat indien het verschil negatief is het minteken is weggelaten).

De mate van overeenstemming over de totaalscore op het examen is bepaald met behulp van de *intraclass correlation* (ICC). De ICC is geschat volgens een *two-way random effects model* waarbij zowel correctoren als examenwerken als random effecten zijn beschouwd (Shrout & Fleiss, 1979; Yaffee, 1998). Omdat ook de systematische verschillen tussen correctoren in ons geval van belang zijn (dit wil zeggen: verschillen in gemiddelden) en niet alleen de rangorde, is gekozen voor de *absolute agreement* definitie (naast *consistency*). Zoals te doen gebruikelijk rapporteren we twee schattingen van de ICC: een voor één individuele corrector (of correctieprocedure) en een voor het gemiddelde of de som van de scores van de correctoren (of correctieprocedures).

De dataset bevat per examenwerk i.c. kandidaat vier metingen, namelijk de totaalscores voor de vier correctieprocedures. De verschillen tussen de gemiddelde examenscores van de vier correctieprocedures zijn op significantie getoetst met een multivariate variantieanalyse (MVA) voor herhaalde metingen. De variabele correctieprocedure is daarbij gedefinieerd als een binnen-subjecten factor. Bij de significantietoetsing van verschillen tussen specifieke correctieprocedures is de eerste correctie telkens als referentiecategorie gekozen. Een vergelijkbare analyse is uitgevoerd voor de verschillen in correctietijd tussen geannoteerde en blanco derde correctie.

De mate waarin en wijze waarop de correcties samenhangen met het cijfer voor het schoolexamen is geanalyseerd met multiple regressie-analyse. Omdat we een kromlijng verband veronderstellen, is voor de regressie van het SE-cijfer op de correcties naast een lineaire ook een kwadratische term geschat.

2.10 Kwaliteit van de derde correctie

In het kader van een check op de kwaliteit van de derde correcties is voor elk van de zes vakken een aselecte steekproef van vijf examenwerken door alle vijf correctoren gecorrigeerd. De correctoren keken de vijf werken geheel onafhankelijk van elkaar na.

De overeenstemming over de totaalscore op het examen is bepaald aan de hand van de *intraclass correlation* (ICC).

De overeenstemming voor de scores op de individuele opgaven is bepaald aan de hand van de correlaties tussen de vijf correctoren, het percentage volledig identieke scores en *Cohen's Kappa* voor correctorparen. Daarnaast is *Cohen's multi-rater Kappa* voor de vijf correctoren berekend.

2.10.1 Nederlands VWO

Het examen Nederlands VWO bevat 32 opgaven met een maximumscore van 50.

Overeenstemming over de totaalscores

De totaalscores van de vijf correctoren op het examen Nederlands stemmen in voldoende mate overeen. De gemiddelde Pearson product-moment correlatie voor de totaalscores van de vijf correctoren is .87. Kennelijk rangordenen de correctoren de kandidaten op vergelijkbare wijze. De

intraclassecorrelatie (ICC), die rekening houdt met verschillen in gemiddelde scores, is iets lager, maar bedraagt voor de individuele corrector nog steeds .80 en voor het gezamenlijke oordeel van de vijf correctoren .95. Kennelijk stemmen de totaalscores van de correctoren in hoge mate met elkaar overeen.

Overeenstemming over de scores per opgave

De analyse van de samenhang tussen de vijf correctoren per opgave is uitgevoerd op in totaal 800 scores (d.w.z. 32 opgaven * 5 kandidaten * 5 correctoren). Het percentage exacte overeenstemming voor deze 800 scores bedraagt 79%. *Cohen's multirater Kappa* voor de vijf correctoren bedraagt .69 (met een standaardfout van .02) hetgeen volgens de vuistregel van Landis en Koch (1997) *substantial* is. De voor toevallige samenhang gecorrigeerde overeenstemming tussen de correctorparen is weergegeven in Tabel 5. Geconstateerd kan worden dat de ene corrector qua overeenstemming niet duidelijk voor de ander onderdoet.

Tabel 5

Overeenstemming tussen de vijf paren correctoren Nederlands (*Cohen's Kappa*)

Corrector	Corrector			
	1	2	3	4
2	.68			
3	.67	.71		
4	.64	.74	.71	
5	.74	.69	.73	.60

2.10.2 Engels GT

Het examen Engels GT bestaat uit 32 opgaven met een maximumscore van 48.

Overeenstemming over de totaalscores

De totaalscores van de vijf correctoren op het examen Engels GT stemmen in voldoende mate overeen. De gemiddelde Pearson product-moment correlatie voor de totaalscores van de vijf correctoren is .84. Kennelijk rangordenen de correctoren de kandidaten op vergelijkbare wijze. De intraclassecorrelatie (ICC), die rekening houdt met eventuele verschillen in gemiddelde scores, is iets lager, maar bedraagt voor de individuele corrector nog steeds .81 en voor het gezamenlijke oordeel van de vijf correctoren .96. Kennelijk stemmen de totaalscores van de correctoren in hoge mate met elkaar overeen.

Overeenstemming over de scores per opgave

De analyse van de samenhang tussen de vijf correctoren voor de eerste 31 opgaven is uitgevoerd op in totaal 775 scores (d.w.z. 31 opgaven * 5 kandidaten * 5 correctoren). Het percentage exacte overeenstemming voor deze 800 scores bedraagt 94%. *Cohen's multirater Kappa* voor de vijf correctoren bedraagt .89 (met een standaardfout van .02) hetgeen volgens de vuistregel van Landis en Koch (1997) *almost perfect* is. De voor toevallige samenhang gecorrigeerde overeenstemming tussen de correctorparen is weergegeven in Tabel 6. Geconstateerd kan worden dat de ene corrector qua overeenstemming niet veel voor de ander onderdoet.

Tabel 6

Overeenstemming tussen de vijf paren correctoren Engels (*Cohen's Kappa*)

Corrector	Corrector			
	1	2	3	4
2	.88			
3	.89	.90		
4	.84	.88	.91	
5	.88	.92	.94	.94

Voor de overeenstemming over opgave 32 - de schrijfpdracht (maximumscore 13) - is de intraclasscorrelatie berekend. De ICC bedroeg .70 voor één corrector en .92 voor alle vijf.

2.10.3 Wiskunde GT

Het examen wiskunde GL bestaat uit 24 opgaven met een maximumscore van 75.

Overeenstemming over de totaalscores

De totaalscores van de vijf correctoren op het examen wiskunde stemmen in voldoende mate overeen. De gemiddelde Pearson product-moment correlatie voor de totaalscores van de vijf correctoren is .98. Kennelijk rangordenen de correctoren de kandidaten op vergelijkbare wijze. De intraclasscorrelatie (ICC), die rekening houdt met verschillen in gemiddelde scores, bedraagt voor een individuele corrector .95 en voor het gezamenlijke oordeel van de vijf correctoren .99. Kennelijk stemmen de totaalscores van de correctoren in hoge mate met elkaar overeen.

Overeenstemming over de scores per opgave

De analyse van de samenhang tussen de vijf correctoren per opgave is uitgevoerd op in totaal 600 scores (d.w.z. 24 opgaven * 5 kandidaten * 5 correctoren). Het percentage exacte overeenstemming voor deze 600 scores bedraagt 74%. *Cohen's multirater Kappa* voor de vijf correctoren bedraagt .66 (met een standaardfout van .01) hetgeen volgens de vuistregel van Landis en Koch (1997) *substantial* is. De voor toevallige samenhang gecorrigeerde overeenstemming tussen de correctorparen is weergegeven in Tabel 7. Zo blijken corrector 2 en 5 het minder met elkaar eens dan corrector 1 en 3.

Tabel 7

Overeenstemming tussen de vijf paren correctoren wiskunde (*Cohen's Kappa*)

Corrector	Corrector			
	1	2	3	4
2	.67			
3	.79	.67		
4	.64	.63	.71	
5	.65	.57	.68	.61

2.10.4 Biologie 1, 2 Vwo

Het examen Biologie VWO bestaat uit 35 opgaven met een maximumscore van 73.

Overeenstemming over de totaalscores

Voor biologie waren voor de check op de kwaliteit van de derde correctie slechts drie examenwerken beschikbaar (zie paragraaf 2.7). De totaalscores van de vijf correctoren stemmen in voldoende mate

overeen. De gemiddelde product-moment correlatie voor de totaalscores van de vijf correctoren is .97. Kennelijk rangordenen de correctoren de kandidaten op nagenoeg dezelfde wijze. De intraclasscorrelatie (ICC), die rekening houdt met verschillen in gemiddelde scores, is iets lager, maar bedraagt voor de individuele corrector nog steeds .81 en voor het gezamenlijke oordeel van de vijf correctoren .95.

Overeenstemming over de scores per opgave

De analyse van de samenhang tussen de vijf correctoren per opgave is uitgevoerd op in totaal 525 scores (d.w.z. 35 opgaven * 3 examenwerken * 5 correctoren). Het percentage exacte overeenstemming voor deze 525 scores bedraagt 82%. *Cohen's multirater Kappa* bedraagt .73 (met een standaardfout van .03) hetgeen volgens de vuistregel van Landis en Koch (1997) *substantial* is. De voor toevallige samenhang gecorrigeerde overeenstemming tussen de correctorparen is weergegeven in Tabel 8. Geconstateerd kan worden dat de ene corrector qua overeenstemming niet veel voor de ander onderdoet.

Tabel 8

Overeenstemming tussen de vijf paren correctoren biologie (*Cohen's Kappa*)

Corrector	Corrector			
	1	2	3	4
2	.69			
3	.76	.75		
4	.73	.74	.71	
5	.73	.74	.66	.76

2.10.5 Geschiedenis Havo

Het examen geschiedenis havo bestaat uit 30 opgaven met een maximumscore van 76.

Overeenstemming over de totaalscores

De overeenstemming tussen de totaalscores van de vijf correctoren op het examen geschiedenis is lager dan bij Nederlands, Engels, wiskunde en biologie. De gemiddelde correlatie tussen totaalscores van de vijf correctoren is .84. Kennelijk rangordenen de correctoren de kandidaten op vergelijkbare wijze. De intraclasscorrelatie (ICC), die gevoelig is voor verschillen in gemiddelde, bedraagt voor de individuele corrector .62 en voor het gezamenlijke oordeel van de vijf correctoren .89.

Overeenstemming over de scores per opgave

De analyse van de samenhang tussen de vijf correctoren per opgave is uitgevoerd op in totaal 750 scores (d.w.z. 30 opgaven * 5 kandidaten * 5 correctoren). Het percentage exacte overeenstemming voor deze 750 scores bedraagt 67%. *Cohen's multirater Kappa* voor de vijf correctoren bedraagt .52 (met een standaardfout van .02) hetgeen volgens de vuistregel van Landis en Koch (1997) *moderate* is. De voor toevallige samenhang gecorrigeerde overeenstemming tussen de correctorparen is weergegeven in Tabel 9. Geconstateerd moet worden dat de mate overeenstemming weinig van paar tot paar verschilt. De laagste *Kappa* bedraagt .46 en de hoogste .60.

Tabel 9

Overeenstemming tussen de vijf paren correctoren geschiedenis (*Cohen's Kappa*)

Corrector	Corrector			
	1	2	3	4
2	.49			
3	.51	.47		
4	.57	.52	.56	
5	.60	.46	.52	.48

2.10.6 Tehatex Havo

Het examen tehatex Havo bestaat uit 36 opgaven met een maximumscore van 75.

Overeenstemming over de totaalscores

De overeenstemming tussen de totaalscores van de vijf correctoren op het examen tehatex is lager dan bij Nederlands, Engels, wiskunde en biologie. De gemiddelde Pearson product-moment correlatie voor de totaalscores van de vijf correctoren is .69. Kennelijk rangordenen de correctoren de kandidaten lang niet altijd op dezelfde wijze. De intraclasscorrelatie (ICC), die rekening houdt met verschillen in gemiddelde scores, is duidelijk lager en bedraagt voor de individuele corrector .34 en voor het gezamenlijke oordeel van de vijf correctoren .72. Kennelijk verschillen de correctoren nogal eens van mening over de hoogte van de totaalscore op het examen tehatex.

Overeenstemming over de scores per opgave

De analyse van de samenhang tussen de vijf correctoren per opgave is uitgevoerd op in totaal 900 scores (d.w.z. 36 opgaven * 5 kandidaten * 5 correctoren). Het percentage exacte overeenstemming voor deze 900 scores bedraagt 66%. *Cohen's multirater Kappa* voor de vijf correctoren bedraagt .49 (met een standaardfout van .02) hetgeen volgens de vuistregel van Landis en Koch (1997) *moderate* is. De voor toevallige overeenstemming gecorrigeerde overeenstemming tussen de correctorparen is weergegeven in Tabel 10. Geconstateerd kan worden dat corrector 3 en 4 het minder met elkaar eens zijn dan bijvoorbeeld corrector 1 en 5.

Tabel 10

Overeenstemming tussen de vijf paren correctoren tehatex (*Cohen's Kappa*)

Corrector	Corrector			
	1	2	3	4
2	.47			
3	.50	.54		
4	.46	.45	.38	
5	.58	.47	.49	.49

3. Resultaten

3.1 Nederlands VWO

Het examen Nederlands VWO bevat 32 opgaven met een maximumscore van 50. Het examen bestaat uit vier meerkeuzevragen, vijf classificatievragen, negen kort-antwoordvragen (waarvan een beweringenvraag) en veertien lang-antwoordvragen (waarbij de vragen bij de samenvatting in overleg met de toetsdeskundige die het examen gemaakt heeft alle als lang-antwoordvraag zijn gecodeerd). In totaal zijn 105 examenwerken Nederlands in het onderzoek betrokken. Alle examenwerken waren van annotaties voorzien.

Tabel 11 toont enkele beschrijvende kenmerken van vier correctieprocedures, te weten eerste correctie, gezamenlijke eerste en tweede correctie, derde correctie van geannoteerde examenwerken en derde correctie van blanco examenwerken. De vier correctieprocedures zijn vergeleken in termen van de totaalscore op het examen, het gemiddeld verschil tussen de totaalscores, de gemiddelde absolute afwijking (d.w.z. het aantal scorepunten dat de correctieprocedures gemiddeld van elkaar verschillen waarbij het er niet toe doet of het verschil positief of negatief is), het cijfer voor het eerste tijdvak en het percentage onvoldoendes waartoe de correctieprocedure geleid zou hebben. De spreidingsdiagrammen zijn weergegeven in Bijlage 1.

Tabel 11

Enkele beschrijvende gegevens voor de vier correctieprocedures Nederlands

	Aantal examenwerken	Minimum	Maximum	Gemiddelde/ percentage	Standaard deviatie
Totaalscore					
Eerste correctie	105	16	43	29.78	6.21
Eerste en tweede correctie gezamenlijk	104	14	43	29.54	5.93
Derde correctie geannoteerd	105	13	42	27.68	5.98
Derde correctie blanco	105	12	40	27.11	6.37
Verschil tussen de totaalscores					
Eerste correctie minus gezamenlijke score	104	-7	6	.27	1.59
Eerste correctie minus derde correctie geannoteerd	105	-7	11	2.10	3.73
Eerste correctie minus derde correctie blanco	105	-12	12	2.67	4.76
Derde correctie geannoteerd minus blanco	105	-11	11	-.56	4.59
Gemiddelde absolute afwijking					
Eerste correctie versus gezamenlijke score	104	0	7	.87	1.36
Eerste correctie versus derde correctie geannoteerd	105	0	11	3.38	2.61
Eerste correctie minus derde correctie blanco	105	0	12	4.57	2.95
Derde correctie geannoteerd versus blanco	105	0	11	3.55	2.94
Cijfer eerste tijdvak					
Eerste correctie	105	4.00	8.80	6.46	1.11
Gezamenlijke correctie	104	3.60	8.80	6.42	1.07
Derde correctie geannoteerd	105	3.40	8.70	6.08	1.08
Derde correctie blanco	105	3.30	8.30	5.98	1.14
Percentage onvoldoendes					
Eerste correctie	105			20%	
Gezamenlijke correctie	104			20%	
Derde correctie geannoteerd	105			31%	
Derde correctie blanco	105			30%	

Verschillen tussen de vier correctieprocedures in gemiddelde Nederlands

Kijken we eerst naar de totaalscore op het examen Nederlands VWO, dan zien we dat de examenkandidaten volgens de eerste correctoren gemiddeld bijna 30 punten hebben behaald (van maximaal 50 punten). De verschillen tussen gemiddelden van de vier correctieprocedures zijn getoetst met behulp van een multivariate variantieanalyse. De belangrijkste resultaten kunnen als volgt worden samengevat:

- Het verschil tussen de gemiddelden van geannoteerde en blanco derde correctie is niet significant ($p = .208$). Het verschil is niet veel meer dan een half scorepunt (oftewel .10 cijferpunt).
- Het gemiddelde van eerste correctie verschilt significant van dat van geannoteerde derde correctie ($p < .001$) en blanco derde correctie ($p < .001$). Eerste correctie kent in vergelijking met

geannoteerde en blanco derde correctie een respectievelijk 2.1 en 2.7 punten hogere score (oftewel .38 en .48 cijferpunt).

- Het verschil tussen de gemiddelden van de eerste corrector en de gezamenlijk overeengekomen score is niet significant ($p = .087$). De gezamenlijke scores zijn gemiddeld slechts een kwart punt lager dan de scores van de eerste corrector (oftewel .04 cijferpunt).

Verschillen tussen vier correctieprocedures in spreiding Nederlands

De gemiddelde absolute afwijking voor de vier correctieprocedures zijn weergegeven in het middelste gedeelte van Tabel 11. Een maat die laat zien hoe sterk correctoren of correctieprocedures van elkaar afwijken is de gemiddelde absolute afwijking. Deze spreidingsmaat is in ons geval het gemiddelde van de absolute waarde van het verschil tussen de scores van de beide correctieprocedures (dat wil zeggen dat indien het verschil negatief is het minteken is weggelaten). De belangrijkste bevindingen beschrijven we als volgt:

- De gemiddelde absolute afwijking tussen de scores van geannoteerde en blanco derde correctie bedraagt 3.6. Dit wil zeggen dat deze twee scores gemiddeld 3.6 scorepunten uit elkaar liggen (negatief dan wel positief). Dit komt overeen met 7% van de maximumscore van 50 oftewel .64 cijferpunt.
- De eerste correctie wijkt gemiddeld 3.4 punten af van geannoteerde derde correctie en gemiddeld 4.6 punten van blanco derde correcties (respectievelijk 7% en 9% van de maximumscore (oftewel .61 en .82 cijferpunt).
- De score van de eerste en gezamenlijke correctie liggen gemiddeld minder dan één scorepunt uit elkaar, oftewel .16 cijferpunt (positief dan wel negatief).

Tabel 12 laat zien hoe de GAA's verdeeld zijn (waarbij waarden van vijf en hoger zijn samengenomen). Kijken we bijvoorbeeld naar de kolom voor absolute afwijkingen van nul, dan blijkt dat de totaalscores van de eerste en de gezamenlijke correctie bij 55% van de nagekeken examenwerken volledig identiek zijn. Daarentegen bedraagt het percentage exacte overeenstemming tussen eerste en derde corrector slechts 8% (geannoteerd) en 5% (blanco). De kolom met de GAA's van vijf of hoger vertoont het omgekeerd beeld: voor eerste en gezamenlijke correctie heeft 4% van de GAA's een waarde van vijf of hoger, voor eerste versus derde geannoteerd is het 28% en voor eerste versus blanco derde correctie gaat het om 50%. Het patroon van uitkomsten geeft aanleiding tot de veronderstelling dat de eerste corrector het doorgaans sterker eens is met de tweede dan met de derde corrector en dat het verschil tussen eerste correctie en geannoteerde derde correctie kleiner is dan dat tussen eerste correctie en blanco derde correctie.

Tabel 12

Verdeling van de gemiddelde absolute afwijking voor vier correctieprocedures Nederlands

		0	1	2	3	4	5 of hoger	Totaal
Eerste correctie minus gezamenlijke score	Aantal	57	26	12	3	2	4	104
	Percentage	55%	25%	12%	3%	2%	4%	100%
Eerste correctie minus derde correctie geannoteerd	Aantal	8	24	18	9	17	29	105
	Percentage	8%	23%	17%	9%	16%	28%	100%
Eerste correctie minus derde correctie blanco	Aantal	5	12	19	9	8	52	105
	Percentage	5%	11%	18%	9%	8%	50%	100%
Derde correctie geannoteerd minus blanco	Aantal	11	21	20	12	3	38	105
	Percentage	11%	20%	18%	12%	3%	37%	100%

Cijfers en percentage onvoldoendes centraal examen Nederlands

In het voorgaande is vastgesteld dat de correctieprocedures niet alle vier tot dezelfde totaalscores op het centraal schriftelijk examen leiden. Het onderste gedeelte van Tabel 11 laat zien tot welke cijfers en percentages onvoldoendes de vier correctieprocedures geleid zouden hebben. In de omzetting van scores naar cijfers is de definitieve normeringsterm 1.1 van het eerste tijdvak gebruikt. De inzet van de tweede corrector leidt niet tot een verhoging van het percentage onvoldoendes: het blijft 20%. Bij geannoteerde en blanco correctie zou respectievelijk 31% en 30% van de kandidaten een onvoldoende hebben behaald.

Overeenstemming tussen de vier correctieprocedures: totaalscore Nederlands

De mate waarin de vier correctieprocedures tot vergelijkbare totaalscores op het examen Nederlands leiden, is ook bepaald aan de hand van de intraclasscorrelatie ICC (zie Tabel 13):

- De samenhang tussen geannoteerde en blanco derde correctie is hoog (voor één correctieprocedure is de ICC .72 en voor beide samen .84). Omdat de hier niet gerapporteerde 95%-betrouwbaarheidsintervallen elkaar overlappen, moet dit verschil als betekenisloos worden beschouwd.
- De voor toevallige overeenstemming gecorrigeerde samenhang tussen eerste en gezamenlijke correctie is aanzienlijk hoger dan die tussen eerste en geannoteerde blanco derde correctie (.97 versus .77). Daarnaast hangt eerste correctie sterker samen met geannoteerde dan met blanco derde correctie (.77 versus .66).
- Voor eerste en gezamenlijk overeengekomen correcties is de ICC zeer hoog (voor een correctieprocedure .96 en voor beide .98). Kennelijk zijn deze twee correctieprocedures in statistisch opzicht niet of nauwelijks van elkaar te onderscheiden en voegt de gezamenlijk overeengekomen score weinig toe aan de totaalscore van de eerste corrector.

Tabel 13

Overeenstemming over de totaalscores (ICC) Nederlands

Overeenstemming totaalscores	Aantal examenwerken	Correlatie	Intraclasscorrelatie	
			Individueel	Gezamenlijk
Eerste correctie en gezamenlijke correctie	104	.97	.97	.98
Eerste correctie en derde correctie geannoteerd	105	.81	.77	.87
Eerste correctie en derde correctie blanco	105	.72	.66	.79
Derde correctie geannoteerd en blanco	105	.73	.72	.84

Overeenstemming tussen de vier correctieprocedures: scores per opgave Nederlands

De mate waarin de onderscheiden correctieprocedures tot vergelijkbare itemscores leiden, is bepaald aan de hand van *Cohen's Kappa* (zie Tabel 14). Omdat de scholen de gezamenlijk overeengekomen score vrijwel nooit per opgave registreren, kunnen deze scores niet met de beide derde correcties in verband worden gebracht en ontbreken deze gegevens in de tabel. Te zien is dat de overeenstemming tussen de drie correctieprocedures ook voor de scores op de individuele opgaven zeer hoog is. De eerste correctie hangt zoals verwacht hoger samen met geannoteerde derde correctie dan met blanco derde correctie (.73 versus .67).

Tabel 14

Overeenstemming over de scores per opgave Nederlands

	Aantal scores	Percentage exacte overeenstemming	Cohen's Kappa	Standaardfout
Eerste correctie en derde correctie geannoteerd	3342	81	.73	.010
Eerste correctie en derde correctie blanco	3342	77	.67	.011
Derde correctie geannoteerd en blanco	3356	78	.68	.010

Overeenstemming tussen drie correctieprocedures per type opgave Nederlands

In de hiervoor besproken analyse is geen onderscheid gemaakt naar type opgave. Het is goed denkbaar dat de samenhang tussen de correctieprocedures voor het ene opgaventype anders is dan voor het andere opgaventype. Zo zou de samenhang tussen de scores van de drie correctieprocedures bij objectief scorebare meerkeuzevragen hoger kunnen zijn dan bij open opgaven. Vandaar dat de mate waarin de drie correcties tot dezelfde itemscores leiden ook per opgaventype is gerapporteerd (zie Tabel 15). De resultaten vatten we als volgt samen:

- De percentages exacte overeenstemming en de voor toevallige samenhang gecorrigeerde overeenstemming zijn het hoogst voor de meerkeuze- en de classificatievragen, gevolgd door de kort-antwoordvragen en ten slotte de lang-antwoordvragen.
- De overeenstemming tussen de drie correctieprocedures is groter voor de gesloten vragen (meerkeuze- en classificatievragen) dan voor de lang-antwoordvragen. Voor de gesloten vragen is de overeenstemming vrijwel perfect. Kennelijk maakt het bij dit opgaventype weinig uit of de opgaven door eerste of derde correctoren worden nagekeken of dat de derde correctie plaatsvindt op basis van geannoteerd dan wel blanco examenmateriaal.
- Bij kort-antwoordvragen en lang-antwoordvragen is het percentage exacte overeenstemming tussen eerste en geannoteerde derde correctie wat hoger dan tussen eerste en blanco derde correctie.
- Bij de kort-antwoordvragen lijkt de voor toevallige overeenstemming gecorrigeerde samenhang tussen eerste en derde correcties wat hoger voor geannoteerde dan voor blanco derde correcties (.73 versus .66). Gezien de grootte van de standaardfouten moet dit verschil echter als niet significant worden beschouwd ($p < .05$).
- Bij de lang-antwoordvragen hangt eerste correctie ook na correctie voor toevallige samenhang hoger samen met geannoteerde dan met blanco derde correctie (.54 versus .46).

Tabel 15

Samenhang tussen drie correctieprocedures per opgaventype Nederlands

	Aantal Scores	Percentage exacte overeenstemming	Cohen's Kappa	Standaardfout
Meerkeuzevragen				
Eerste correctie en derde correctie geannoteerd	420	98	.95	.015
Eerste correctie en derde correctie blanco	420	98	.95	.015
Derde correctie geannoteerd en blanco	420	99	.97	.012
Classificatievragen				
Eerste correctie en derde correctie geannoteerd	525	99	.98	.009
Eerste correctie en derde correctie blanco	525	99	.98	.010
Derde correctie geannoteerd en blanco	525	100	1.00	.004
Kort-antwoordvragen				
Eerste correctie en derde correctie geannoteerd	945	83	.73	.020
Eerste correctie en derde correctie blanco	944	79	.66	.022
Derde correctie geannoteerd en blanco	944	81	.70	.021
Lang-antwoordvragen				
Eerste correctie en derde correctie geannoteerd	1452	68	.54	.017
Eerste correctie en derde correctie blanco	1453	62	.46	.018
Derde correctie geannoteerd en blanco	1467	62	.46	.018

Verskil in correctietijd Nederlands

De gemiddelde correctietijd per corrector per versie is weergegeven in Tabel 16. Onder meer is te zien dat corrector 1 voor het nakijken van een geannoteerde examenwerk gemiddeld 33 minuten nodig had voor een blanco werk gemiddeld 34 minuten. De vijf correctoren hadden per geannoteerde examenwerk gemiddeld 23.36 minuten nodig en per blanco werken gemiddeld 23.99 minuten. Met behulp van multivariate variantie-analyse (MVA) is nagegaan of dit verschil in correctietijd tussen geannoteerd en blanco examenwerk statistisch significant is en dat bleek niet het geval ($p = .674$). Een geannoteerd werk kost gemiddeld slechts 38 seconden minder correctietijd dan een blanco examenwerk.

Tabel 16

Correctietijd per corrector voor geannoteerde en blanco examenwerken Nederlands

Corrector	Geannoteerde examenwerken		Blanco examenwerken	
	Gemiddelde	Standaard deviatie	Gemiddelde	Standaard deviatie
1	33.29	4.64	34.38	8.14
2	40.14	3.15	40.29	2.85
3	16.57	2.80	15.71	1.59
4	11.48	2.29	15.10	2.90
5	15.33	1.68	14.48	2.25
Gemiddelde	23.36	11.69	23.99	11.87

Tabel 16 doet vermoeden dat de correctoren verschillen in de gemiddelde correctietijd en dat het verschil tussen de correctietijd van geannoteerd en blanco werk voor de ene corrector groter is dan voor de andere. Vandaar dat een tweede MVA is uitgevoerd waarbij we corrector als tussen-subjecten factor aan het model toevoegen. De MVA stelt ons in staat om het hoofdeffect van corrector (tussen-subjecten) en de interactie tussen tijd en corrector (binnen-subjecten) te bestuderen. Zowel het hoofdeffect van corrector als het interactie-effect blijkt statistisch significant (beide $p < .001$). Kennelijk verschillen correctoren in de hoeveelheid geïnvesteerde correctietijd en is het verschil tussen de correctietijd van geannoteerde en blanco examenwerken voor de ene corrector groter dan voor de andere. Ten aanzien van het hoofdeffect van corrector laat de tabel zien dat corrector 1 en 2 ruim twee keer zoveel tijd investeren als corrector 3, 4 en 5. De interactie tussen corrector en versie zit hem vooral in corrector 4. Bij corrector 4 kosten geannoteerde werken duidelijk minder tijd dan blanco werken (11 versus 15 minuten). De correctoren 1, 2, 3 en 5 besteden ongeveer evenveel tijd aan geannoteerde als aan blanco examenwerken.

Vanwege de grote verschillen in tijdsbesteding tussen correctoren en de interactie met de versie van het examenwerk mag aan het eerder gerapporteerde verschil in gemiddelde correctietijd tussen geannoteerde en blanco derde correctie geen grote betekenis worden gehecht.

Cijfer op het schoolexamen Nederlands

Nagegaan is of de correcties samenhangen met het cijfer voor het schoolexamen en zo ja in hoeverre deze relatie verschilt per correctieprocedure. Tabel 17 geeft informatie over de verdeling van de correcties en het schoolexamencijfer.

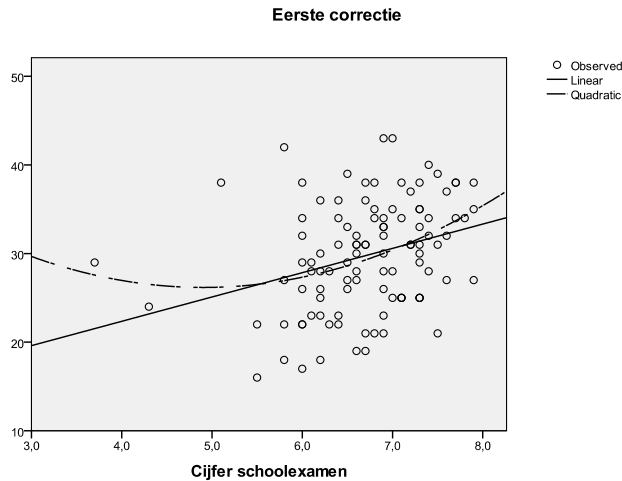
Tabel 17

Karakterisering van de scoreverdeling voor drie correctieprocedures en het schoolexamencijfer Nederlands

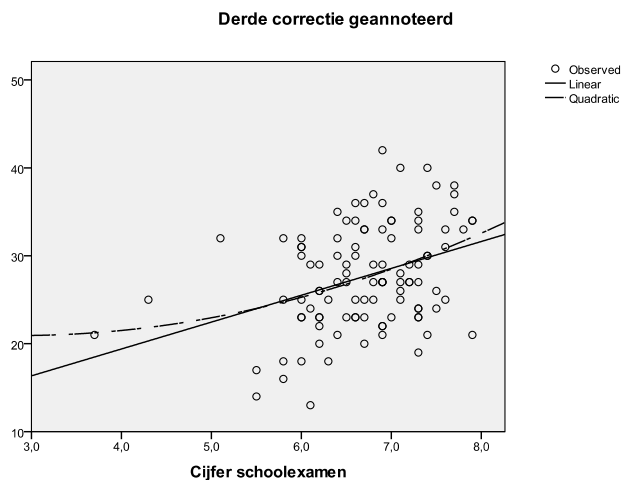
	Aantal examenwerken	Min	Max	Gemid- delde	Standaard deviatie	Scheefheid	
						Statistic	SE
Eerste correctie	105	16	43	29.78	6.21	-.07	.24
Derde correctie geannoteerd	105	13	42	27.68	5.98	.02	.24
Derde correctie blanco	105	12	40	27.11	6.37	-.16	.24
Cijfer schoolexamen	105	3.7	7.9	6.71	.71	-1.09	.24

De scores voor de drie correctieprocedures lijken redelijk normaal verdeeld. De verdeling van de schoolexamencijfers is sterk negatief scheef: lage cijfers en cijfers rond de grens tussen voldoende en

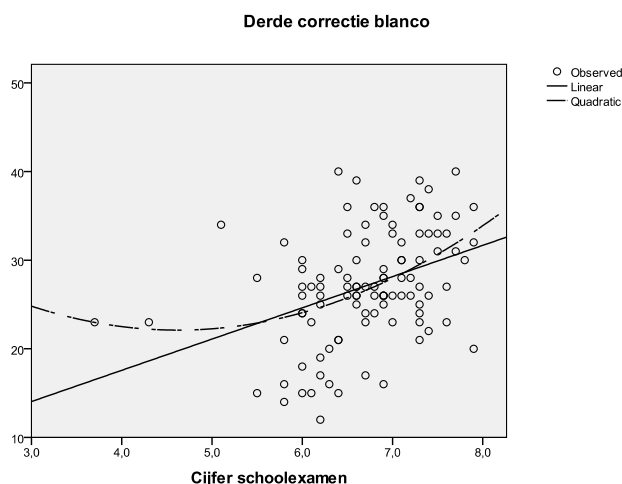
onvoldoende worden veel minder gegeven dan hogere cijfers. De spreidingsdiagrammen in Figuur 1a, 1b en 1c bevestigen dit beeld (waarbij de schoolexamencijfers zijn weergegeven op de horizontale as en de drie soorten correcties op de verticale as). In de figuur geeft de doorgetrokken rechte lijn de sterkte en richting van het lineaire verband weer en staat de onderbroken curve voor het kwadratische verband.



Figuur 1a Spreidingsdiagram eerste correctie versus cijfer schoolexamen Nederlands



Figuur 1b Spreidingsdiagram geannoteerde derde correctie versus cijfer schoolexamen Nederlands



Figuur 1c Spreidingsdiagram geannoteerde blanco correctie versus cijfer schoolexamen Nederlands

De spreidingsdiagrammen doen een kromlijng verband tussen correctieprocedures en schoolexamencijfer vermoeden. De resultaten van de multiple regressieanalyse zijn weergegeven in Tabel 18. Weergegeven zijn de regressiegewichten met bijbehorende standaardfouten van model 1 met alleen een lineaire term voor het schoolexamencijfer en van model 2 met zowel een lineaire als een kwadratische term. De tabel geeft ook R, de multiple correlatie, die in het geval van model 1 gelijk is aan de product-moment correlatie. De lineaire samenhang tussen de prestaties op het centraal examen en het schoolexamen is in alle drie gevallen zwak (een toename van het schoolexamencijfer met één punt gaat gepaard met een toename van de score op het centraal schriftelijk examen met respectievelijk 2.7 (eerste correctie), 3.1 (derde correctie geannoteerd) en 3.5 scorepunten (derde correctie blanco). Hoewel de plotjes anders doen vermoeden lijkt er weinig aanleiding om een kromlijng verband tussen de correcties en het schoolexamencijfer te veronderstellen. In model 2 is noch de lineaire noch de kwadratische term significant, terwijl de multiple correlatie gelijk of niet veel hoger is dan die van model 1.

Tabel 18

Uitkomsten van de analyse van de regressie van de totaalscore op het schoolexamencijfer Nederlands

		Ongestandaardiseerde regressiegewichten		Gestandaardiseerde regressiegewichten	T	Sig.	R
		b	se	Beta			
Eerste correctie							
Model 1	Lineair	2.74	.82	.31	3.33	<.001	.31
Model 2	Lineair	-9.45	7.80	-1.08	-1.21	.228	.35
	Kwadratisch	.96	.61	1.40	1.57	.119	
Derde correctie geannoteerd							
Model 1	Lineair	3.06	.78	.36	3.94	<.001	.36
Model 2	Lineair	-2.51	7.43	-.30	-.34	.736	.37
	Kwadratisch	.44	.58	.66	.75	.453	
Derde correctie blanco							
Model 1	Lineair	3.52	.82	.39	4.31	<.001	.39
Model 2	Lineair	-9.52	7.73	-1.06	-1.23	.221	.42
	Kwadratisch	1.03	.61	1.46	1.70	.093	

3.2 Engels GLTL

Het examen Engels GT bevat 32 opgaven en kent een maximumscore van 48 (waarbij opgave 32 de totaalscore op de schrijfpdracht is). Het examen bestaat uit tien gewone meerkeuzevragen, één meerkeuzevraag met doorloopstam, elf meerkeuze-invulvragen, twee bewerkingsvragen, één kort-antwoordvraag, drie citeervragen en vier lang-antwoordvragen (waarvan er een de totaalscore voor de schrijfpdracht was). In totaal hebben de derde correctoren de geannoteerde en blanco versie van 110 examenwerken gecorrigeerd. Negen examenwerken bleken geen annotaties van de eerste corrector te bevatten. Voor de vergelijking van de vier correctieprocedures zijn derhalve maximaal 101 examenwerken beschikbaar.

Tabel 19 toont enkele beschrijvende kenmerken van vier correctieprocedures, te weten eerste correctie, gezamenlijke eerste en tweede correctie, derde correctie van geannoteerde examenwerk en derde correctie van blanco examenwerken. De vier correctieprocedures zijn vergeleken in termen van de totaalscore op het examen, het gemiddeld verschil tussen de totaalscores, de gemiddelde absolute afwijking (d.w.z. het aantal scorepunten dat de correctieprocedures gemiddeld van elkaar verschillen waarbij het er niet toe doet of het verschil positief of negatief is), het cijfer voor het eerste tijdvak en het percentage onvoldoendes waartoe de correctieprocedure geleid zou hebben. De spreidingsdiagrammen zijn weergegeven in Bijlage 2.

Tabel 19

Enkele beschrijvende gegevens voor de vier correctieprocedures Engels

	Aantal examenwerken	Minimum	Maximum	Gemiddelde/ percentage	Standaard deviatie
Totaalscore					
Eerste correctie	96	13	43	28.47	6.47
Eerste en tweede correctie gezamenlijk	91	10	42	27.85	6.37
Derde correctie geannoteerd	101	9	42	27.17	6.60
Derde correctie blanco	101	11	41	27.43	6.59
Verskil tussen de totaalscores					
Eerste correctie minus gezamenlijke score	91	-4.00	4.00	.33	1.12
Eerste correctie minus derde correctie geannoteerd	96	-4.00	8.00	1.34	2.37
Eerste correctie minus derde correctie blanco	96	-8.00	8.00	1.07	2.79
Derde correctie geannoteerd minus blanco	101	-7.00	7.00	-.26	2.57
Gemiddelde absolute afwijking					
Eerste correctie versus gezamenlijke score	91	.00	4.00	.57	1.01
Eerste correctie versus derde correctie geannoteerd	96	.00	8.00	2.20	1.59
Eerste correctie minus derde correctie blanco	96	.00	8.00	2.39	1.79
Derde correctie geannoteerd versus blanco	101	.00	7.00	1.90	1.73
Cijfer eerste tijdvak					
Eerste correctie	96	3.40	9.10	6.34	1.22
Gezamenlijke correctie	91	2.90	8.90	6.23	1.20
Derde correctie geannoteerd	101	2.70	8.90	6.10	1.24
Derde correctie blanco	101	3.10	8.70	6.16	1.24
Percentage onvoldoendes					
Eerste correctie	96			27%	
Gezamenlijke correctie	91			30%	
Derde correctie geannoteerd	101			36%	
Derde correctie blanco	101			34%	

Verschillen tussen de vier correctieprocedures in gemiddelden Engels

Kijken we eerst naar de totaalscore op het examen Engels GT dan zien we dat de examenkandidaten volgens de eerste correctoren gemiddeld ruim 28 punten hebben behaald (van maximaal 48 punten). De verschillen tussen de gemiddelden van de vier correctieprocedures zijn getoetst met behulp van een multivariate variantieanalyse. De belangrijkste resultaten kunnen als volgt worden samengevat:

- Het verschil tussen de gemiddelden van geannoteerde en blanco derde correctie is niet significant ($p = .208$). Het verschil bedraagt slechts een kwart scorepunt (oftewel .06 cijferpunt).
- Het gemiddelde van eerste correctie verschilt significant van dat van geannoteerde derde correctie ($p = .006$) en blanco derde correctie ($p < .001$). Eerste correctie leidt in vergelijking met geannoteerde en blanco derde correctie tot een respectievelijk 1.3 en 1.1 punten hogere score (oftewel .24 en .18 cijferpunt).

- Het verschil tussen de gemiddelden van de eerste correctie en de gezamenlijk tussen eerste en tweede corrector overeengekomen score is niet significant ($p = .208$). De gezamenlijke score is gemiddeld slechts een derde punt lager dan die van de eerste correctie (oftewel .11 cijferpunt).

Verschillen tussen vier correctieprocedures in spreiding Engels

De gemiddelde absolute afwijking (GAA) voor de vier correctieprocedures zijn weergegeven in het middelste gedeelte van Tabel 19. Een maat die laat zien hoe sterk correctoren of correctieprocedures van elkaar afwijken is de gemiddelde absolute afwijking. Deze spreidingsmaat is in ons geval het gemiddelde van de absolute waarde van het verschil tussen de scores van de beide correctieprocedures (dat wil zeggen dat indien het verschil negatief is het minteken is weggelaten). De belangrijkste resultaten kunnen als volgt worden beschreven:

- Het gemiddelde absolute verschil tussen geannoteerde en blanco derde correctie bedraagt 1.9 scorepunten. Dit wil zeggen dat deze twee scores gemiddeld bijna twee punten uit elkaar liggen (negatief dan wel positief). Dit komt overeen met 4% van de maximumscore (oftewel .36 cijferpunt).
- De scores van de eerste correctie wijken dan gemiddeld 2.2 punten af van geannoteerde derde correctie en gemiddeld 2.4 punten van blanco derde correctie. Dit komt overeen met respectievelijk 5% en 5% van de maximumscore (oftewel .41 en .45 cijferpunt).
- Voor eerste correctie en gezamenlijke eerste en tweede correctie bedraagt de gemiddelde absolute afwijking slechts zestiende punt (oftewel .11 cijferpunt).

Tabel 20 laat zie hoe de GAA's verdeeld zijn (waarbij waarden van vijf en hoger zijn samengenomen). Kijken we bijvoorbeeld naar de kolom voor absolute afwijkingen van nul, dan blijkt dat de totaalscores van de eerste en de gezamenlijke correctie bij 68% van de nagekeken examenwerken volledig identiek zijn. Daarentegen bedraagt het percentage exacte overeenstemming tussen eerste en derde corrector slechts 11% (geannoteerd) en 13% (blanco). De kolom met de GAA's van vijf of hoger vertoont het omgekeerd beeld: voor eerste en gezamenlijke correctie heeft 0% van de GAA's een waarde van vijf of hoger, voor eerste versus derde geannoteerd is het 7% en voor eerste versus blanco derde correctie gaat het om 13%. Het patroon van uitkomsten doet vermoeden dat de eerste corrector het doorgaans sterker eens is met de tweede dan met de derde corrector en dat het verschil tussen eerste correctie en geannoteerde derde correctie kleiner is dan dat tussen eerste correctie en blanco derde correctie.

Tabel 20

Verdeling van de gemiddelde absolute afwijking voor vier correctieprocedures Engels

		0	1	2	3	4	5 of hoger	Totaal
Eerste correctie minus gezamenlijke score	Aantal	62	15	8	3	3	0	91
	Percentage	68%	16%	9%	3%	3%	0%	100%
Eerste correctie minus derde correctie geannoteerd	Aantal	11	27	21	19	11	7	96
	Percentage	11%	28%	22%	20%	11%	7%	100%
Eerste correctie minus derde correctie blanco	Aantal	12	22	24	15	11	12	96
	Percentage	13%	23%	25%	16%	11%	13%	100%
Derde correctie geannoteerd minus blanco	Aantal	23	29	15	19	7	8	101
	Percentage	23%	29%	15%	19%	7%	8%	100%

Cijfers en percentage onvoldoendes centraal examen Engels

In het voorgaande is vastgesteld dat de correctieprocedures niet alle vier tot dezelfde totaalscores op het centraal examen leiden. Het onderste gedeelte van Tabel 19 laat zien tot welke cijfers en percentages onvoldoende de vier correctieprocedures geleid zouden hebben. In de omzetting van scores naar cijfers is de normeringsterm 1.0 van het eerste tijdvak gebruikt. De inzet van de tweede

corrector leidt tot een kleine verhoging van het percentage onvoldoendes: van 27% naar 30%. Bij geannoteerde en blanco correctie zou respectievelijk 36% en 34% van de kandidaten voor het eerste tijdvak een onvoldoende hebben behaald.

Overeenstemming tussen de vier correctieprocedures: totaalscore Engels

De mate waarin de vier correctieprocedures tot vergelijkbare totaalscores op het examen leiden, is ook bepaald aan de hand van de intraclasscorrelatie ICC (zie Tabel 20):

- Geannoteerde en blanco derde correctie lijken tot in hoge mate vergelijkbare examenscores te leiden (voor één correctieprocedure is de ICC .92 en voor beide samen .96). Aangezien de hier niet gerapporteerde 95%-betrouwbaarheidsintervallen elkaar overlappen, moet dit verschil echter als betekenisloos worden beschouwd.
- Eerst correctie hangt niet sterker samen met geannoteerde derde correctie dan met blanco derde correctie (.92 versus .90).
- De samenhang tussen eerste en gezamenlijke correctie is hoger dan die tussen eerste en geannoteerde derde correctie (.98 versus .92) en die tussen eerste en blanco derde correctie (.98 versus .90).
- De eerste en de gezamenlijk overeengekomen correcties hangen nagenoeg perfect samen (voor één correctieprocedure .98 en voor beide .99). Kennelijk zijn deze twee correctieprocedures in statistisch opzicht niet of nauwelijks van elkaar te onderscheiden en voegt de een zeer weinig toe aan de ander.

Tabel 20

Overeenstemming over de totaalscores (ICC) Engels

Overeenstemming totaalscores	Aantal examenwerken	Correlatie	Intraclasscorrelatie	
			Individueel	Gezamenlijk
Eerste correctie en gezamenlijke correctie	91	.99	.98	.99
Eerste correctie en derde correctie geannoteerd	96	.93	.92	.96
Eerste correctie en derde correctie blanco	96	.90	.90	.95
Derde correctie geannoteerd en blanco	101	.92	.92	.96

Overeenstemming tussen de vier correctieprocedures: scores per opgave Engels

De mate waarin de onderscheiden correctieprocedures tot vergelijkbare scores voor de individuele opgaven leiden, is bepaald aan de hand van *Cohen's Kappa* (zie Tabel 21). De scores voor de schrijfofopdracht zijn minimaal van intervalniveau (maximumscore 13) en zijn om deze reden apart geanalyseerd. Omdat vestigingen de scores van de tweede corrector en de gezamenlijk overeengekomen score vrijwel nooit per opgave registreren, kunnen de derde correcties niet met deze beide correcties in verband worden gebracht en ontbreken deze gegevens in de tabel. Te zien is dat de overeenstemming tussen de drie correctieprocedures ook voor de scores op de individuele opgaven zeer hoog is. Gezien de grootte van de standaardfouten hangt eerste correctie niet hoger samen met geannoteerde dan met blanco derde correctie (.93 versus .91).

Tabel 21

Overeenstemming over de scores per opgave Engels

	Aantal scores	Percentage exacte overeenstemming	Cohen's Kappa	Standaardfout
Eerste correctie en derde correctie geannoteerd	2976	96	.93	.006
Eerste correctie en derde correctie blanco	2976	95	.91	.007
Derde correctie geannoteerd en blanco	3131	95	.91	.007

Overeenstemming tussen drie correctieprocedures per type opgave Engels

In de hiervoor besproken analyse is geen onderscheid gemaakt naar type opgave. Het is goed denkbaar dat de samenhang tussen de correctieprocedures voor het ene opgavetype anders is dan voor het andere opgavetype. Zo zou de samenhang tussen de drie correctieprocedures bij objectief scoorbare gesloten opgaven hoger kunnen zijn dan bij open opgaven. Vandaar dat de mate waarin de drie correcties tot dezelfde itemscores leiden ook per opgavetype is gerapporteerd (zie Tabel 22):

- Bij de drie soorten meerkeuzevragen leiden de drie correctieprocedures tot nagenoeg identieke scores. Kennelijk maakt het voor deze meerkeuzevragen weinig uit of de opgaven door eerste of derde correctoren worden nagekeken of dat de derde correctie plaatsvindt op basis van geannoteerd dan wel blanco examenmateriaal (afgezien van sporadische slordigheidsfoutjes).
- Bij de drie half open opgavetypen - bewerkingsvragen, kort-antwoordvragen en citeervragen - zijn de verschillen tussen de drie correctieprocedures over het algemeen wat groter dan bij de meerkeuzevragen,
- Bij de lang-antwoordvragen lopen de scores voor de drie correctieprocedures het sterkst uiteen.
- Eerste correctie hangt bij geen van de onderzochte opgavetypen sterker samen met geannoteerde dan met blanco derde correctie. Althans, in geen van de gevallen overlappen de 95%-betrouwbaarheidsintervallen van twee maal de standaardfout elkaar. Dit geldt ook voor lang-antwoordvragen, waarbij men moet bedenken dat de standaardfouten hier zeer groot zijn omdat er slechts drie van deze vragen in deze analyse betrokken zijn (de vierde lang-antwoordvraag was de totaalscore op de schrijfpodracht die afzonderlijk geanalyseerd is).

Tabel 22

Samenhang tussen drie correctieprocedures per opgaventype Engels

	Aantal scores	Percentage exacte overeen- stemming	Cohen's Kappa	Standaard- fout
Meerkeuzevragen				
Eerste correctie en derde correctie geannoteerd	960	99	.98	.006
Eerste correctie en derde correctie blanco	960	99	.97	.008
Derde correctie geannoteerd en blanco	1010	99	.98	.006
Meerkeuzevraag met doorloopstam				
Eerste correctie en derde correctie geannoteerd	96	99	.98	.021
Eerste correctie en derde correctie blanco	96	99	.94	.036
Derde correctie geannoteerd en blanco	101	99	.96	.028
Meerkeuze-invulvragen				
Eerste correctie en derde correctie geannoteerd	1111	99	.97	.007
Eerste correctie en derde correctie blanco	1111	98	.97	.008
Derde correctie geannoteerd en blanco	1111	99	.99	.004
Bewerkingvragen				
Eerste correctie en derde correctie geannoteerd	192	95	.92	.027
Eerste correctie en derde correctie blanco	192	93	.87	.034
Derde correctie geannoteerd en blanco	202	96	.93	.024
Kort-antwoordvragen				
Eerste correctie en derde correctie geannoteerd	96	95	.77	.098
Eerste correctie en derde correctie blanco	96	96	.82	.085
Derde correctie geannoteerd en blanco	101	95	.81	.082
Citeervragen				
Eerste correctie en derde correctie geannoteerd	288	89	.84	.029
Eerste correctie en derde correctie blanco	288	86	.77	.033
Derde correctie geannoteerd en blanco	303	85	.76	.033
Lang-antwoordvragen				
Eerste correctie en derde correctie geannoteerd	288	86	.73	.040
Eerste correctie en derde correctie blanco	288	80	.59	.047
Derde correctie geannoteerd en blanco	303	77	.54	.048

De samenhang tussen de drie correctieprocedures bij de lang-antwoordvraag voor de schrijfpdracht is geanalyseerd met behulp van de intraclasscorrelatie (zie Tabel 23). Gezien de grootte van de hier niet gerapporteerde 95%-betrouwbaarheidsintervallen zijn de gerapporteerde ICC's in statistisch opzicht niet onderscheidbaar.

Tabel 23

Overeenstemming (ICC) over de lang-antwoordvraag voor de schrijfpdracht Engels

Overeenstemming	Aantal examenwerken	Intraclasscorrelatie	
		Individueel	Gezamenlijk
Eerste correctie en derde correctie geannoteerd	96	.65	.79
Eerste correctie en derde correctie blanco	96	.62	.76
Derde correctie geannoteerd en blanco	101	.72	.84

Verskil in correctietijd Engels

De gemiddelde correctietijd per corrector per versie is weergegeven in Tabel 24. Te zien is onder meer dat corrector 4 voor het nakijken van een geannoteerde examenwerk gemiddeld bijna 6 minuten nodig had voor een blanco werk gemiddeld ruim 7 minuten. De vijf correctoren hadden gemiddeld 11.47 minuten nodig en voor de blanco examenwerken 11.36 minuten. Met behulp van multivariate variantie-analyse (MVA) is nagegaan of het tijdsverschil significant is en dat bleek niet het geval ($p = .868$). Het verschil blijkt slechts zeven seconden te zijn.

Tabel 24

Correctietijd per corrector voor geannoteerde en blanco examenwerken Engels

Corrector	Geannoteerde examenwerken		Blanco examenwerken	
	Gemiddelde	Standaard deviatie	Gemiddelde	Standaard deviatie
1	11.50	2.67	10.57	2.09
2	10.55	1.61	10.35	1.53
3	15.15	3.47	15.00	3.93
4	5.85	1.27	7.10	1.68
5	14.14	2.50	13.80	2.07
Gemiddelde	11.47	4.05	11.36	3.66

Met behulp van MVA is ook nagegaan of de ene corrector meer tijd aan de correctie besteedde dan de andere en of het tijdsverschil tussen geannoteerde en blanco werken stabiel is over correctoren. Zowel het hoofdeffect van corrector als het interactie-effect van corrector met versie blijkt statistisch significant (beide $p < .001$). Kennelijk verschillen correctoren in de hoeveelheid geïnvesteerde correctietijd en is het verschil tussen de correctietijd van geannoteerde en blanco examenwerken voor de ene corrector groter dan voor de andere. Ten aanzien van het hoofdeffect van corrector laat de tabel zien dat corrector 4 veel sneller nakijkt dan de overige correctoren. Ook de interactie tussen corrector en versie lijkt deels aan corrector 4 toe te schrijven: hij of zij besteedt aan geannoteerde werken wat minder correctietijd dan aan blanco werken (6 versus 7 minuten).

Vanwege de verschillen in tijdsbesteding tussen correctoren en de interactie met de versie van het examenwerk mag aan het eerder gerapporteerde verschil in gemiddelde correctietijd tussen geannoteerde en blanco derde correctie geen al te grote betekenis worden gehecht.

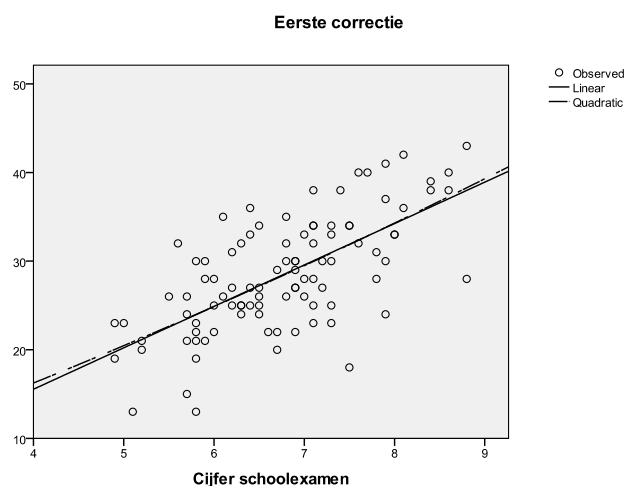
Cijfer op het schoolexamen Engels

Nagegaan is in hoeverre de correcties samenhangen met het cijfer voor het schoolexamen en zo ja in hoeverre deze relatie verschilt per correctieprocedure. Tabel 25 geeft informatie over de verdeling van de drie correctieprocedures en het schoolexamencijfer. Tabel 25

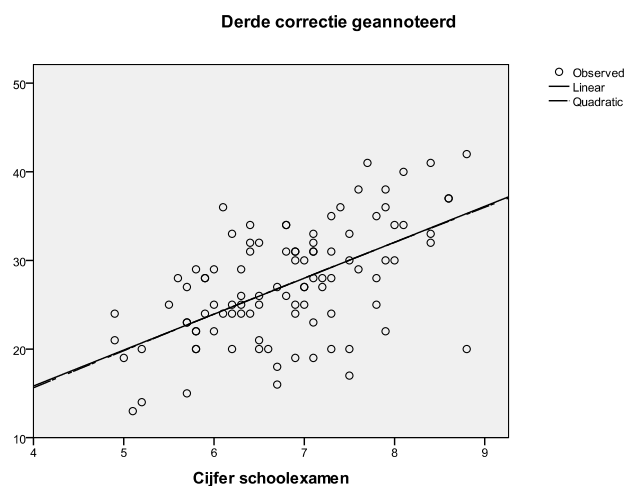
Karakterisering van de scoreverdeling voor drie correctieprocedures en het schoolexamencijfer Engels

	Aantal examenwerken	Min	Max	Gemid- delde	Standaard deviatie	Scheefheid	
						Statistic	SE
Eerste correctie	96	13	43	28.47	6.47	.08	.25
Derde correctie geannoteerd	101	9	42	27.17	6.60	-.04	.24
Derde correctie blanco	101	11	41	27.43	6.59	.03	.24
Cijfer schoolexamen	101	4.90	8.80	6.80	.92	.10	.24

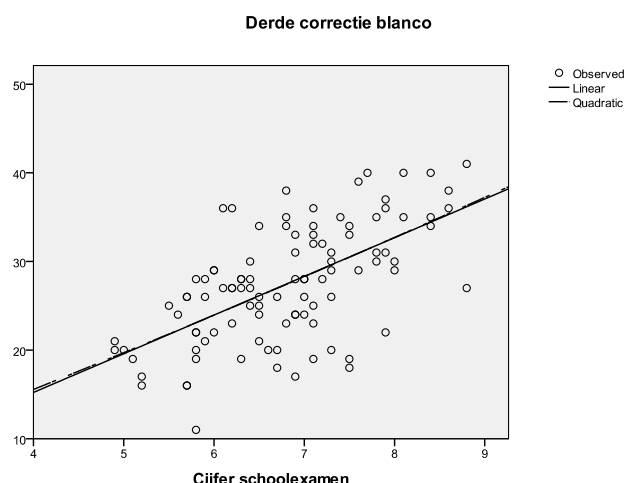
De scores voor de drie correctieprocedures lijken redelijk normaal verdeeld terwijl de verdeling van het schoolexamencijfer negatief scheef is met relatief weinig lage cijfers en relatief veel middelmatige en hoge cijfers. De spreidingsdiagrammen in Figuur 2a, 2b en 2c bevestigen dit beeld.



Figuur 2a Spreidingsdiagram eerste correctie versus cijfer schoolexamen Engels



Figuur 2b Spreidingsdiagram geannoteerde derde correctie versus cijfer schoolexamen Engels



Figuur 2c Spreidingsdiagram geannoteerde blanco correctie versus cijfer schoolexamen Engels

De resultaten van de multiple regressieanalyse zijn weergegeven in Tabel 26. Weergegeven zijn de regressiegewichten met bijbehorende standaardfouten van model 1 met alleen een lineaire term voor het schoolexamencijfer en van model 2 met zowel een lineaire als een kwadratische term. De tabel geeft ook R, de multiple correlatie, die in het geval van model 1 gelijk is aan de product-moment correlatie. De lineaire samenhang tussen de prestaties op het centraal examen en het schoolexamen is in alle drie gevallen relatief sterk. Een toename van het schoolexamencijfer met één punt gaat gepaard met een toename van de score op het centraal schriftelijk examen met respectievelijk 4.7 (eerste correctie), 4.1 (derde correctie geannoteerd) en 4.4 scorepunten (derde correctie blanco). Er zijn geen aanwijzingen voor het veronderstelde kromlijng verband tussen de correcties en het schoolexamencijfer. De multiële correlaties voor model 1 zijn telkens gelijk aan dan die voor model 2. De relatie tussen correcties en schoolexamencijfer kan met één lineaire term afdoende worden beschreven.

Tabel 26

Uitkomsten van de analyse van de regressie van de totaalscore op het schoolexamencijfer Engels

		Ongestandaardiseerde regressiegewichten		Gestandaardiseerde regressiegewichten	T	Sig.	R
		b	se	Beta			
Eerste correctie							
Model 1	Lineair	4.67	.55	.66	8.57	<.001	.66
Model 2	Lineair	3.33	6.74	.47	.49	.623	.66
	Kwadratisch	.10	.49	.19	.20	.842	
Derde correctie geannoteerd							
Model 1	Lineair	4.05	.60	.56	6.80	<.001	.56
Model 2	Lineair	4.43	7.46	.62	.59	.554	.56
	Kwadratisch	-.03	.54	-.05	-.05	.960	
Derde correctie blanco							
Model 1	Lineair	4.37	.57	.61	7.63	<.001	.61
Model 2	Lineair	3.73	7.16	.52	.52	.604	.61
	Kwadratisch	.05	.52	.09	.09	.929	

3.3 Wiskunde GLTL

Het examen wiskunde GT bevat 24 opgaven met een maximumscore van 75. Het examen bestaat uit één classificatievraag, twee in- of aanvulvragen en 21 lang-antwoordvragen. Van de 110 examenwerken bleken er vijf geen annotaties van de eerste corrector te bevatten. Van de 105 resterende examenwerken waren er drie onbruikbaar voor analyse omdat de desbetreffende scoreformulieren onvolledig waren ingevuld.

Tabel 27 toont enkele beschrijvende kenmerken van vier correctieprocedures, te weten eerste correctie, gezamenlijke eerste en tweede correctie, derde correctie van geannoteerde examenwerk en derde correctie van blanco examenwerken. De vier correctieprocedures zijn vergeleken in termen van de totaalscore op het examen, het gemiddeld verschil tussen de totaalscores, de gemiddelde absolute afwijking (d.w.z. het aantal scorepunten dat de correctieprocedures gemiddeld van elkaar verschillen waarbij het er niet toe doet of het verschil positief of negatief is), het cijfer voor het eerste tijdvak en het percentage onvoldoendes waartoe de correctieprocedure geleid zou hebben. De spreidingsdiagrammen zijn weergegeven in Bijlage 3.

Tabel 27 Enkele beschrijvende gegevens voor de vier correctieprocedures wiskunde

	Aantal examenwerken	Minimum	Maximum	Gemiddelde	Standaard deviatie
Totaalscore					
Eerste correctie	97	25	74	49.65	10.70
Eerste en tweede correctie gezamenlijk	97	23	74	49.39	10.74
Derde correctie geannoteerd	102	22	73	47.76	11.21
Derde correctie blanco	102	23	73	47.19	11.14
Verschil tussen de totaalscores					
Eerste correctie minus gezamenlijke score	97	-2	5	.26	1.09
Eerste correctie minus derde correctie geannoteerd	97	-7	10	1.43	3.45
Eerste correctie minus derde correctie blanco	97	-9	13	1.99	3.96
Derde correctie geannoteerd minus blanco	102	-10	10	.58	3.99
Gemiddelde absolute afwijking					
Eerste correctie versus gezamenlijke score	97	0	5	.53	.99
Eerste correctie versus derde correctie geannoteerd	97	0	10	2.88	2.38
Eerste correctie minus derde correctie blanco	97	0	13	3.62	2.53
Derde correctie geannoteerd versus blanco	102	0	10	3.26	2.34
Cijfer eerste tijdvak					
Eerste correctie	97	3.60	9.50	6.56	1.28
Gezamenlijke correctie	97	3.40	9.50	6.53	1.29
Derde correctie geannoteerd	102	3.20	9.40	6.33	1.35
Derde correctie blanco	102	3.40	9.40	6.26	1.34
Percentage onvoldoendes					
Eerste correctie	97			23%	
Gezamenlijke correctie	97			26%	
Derde correctie geannoteerd	102			27%	
Derde correctie blanco	102			29%	

Verschillen tussen de vier correctieprocedures in gemiddelde wiskunde

Kijken we eerst naar de totaalscore op het examen wiskunde, dan zien we dat de examenkandidaten volgens de eerste correctoren gemiddeld bijna vijftig punten hebben behaald (van maximaal 75 punten). De verschillen tussen de gemiddelden van de vier correctieprocedures zijn getoetst met behulp van een multivariate variantieanalyse. De belangrijkste uitkomsten vatten we als volgt samen:

- Het verschil tussen de gemiddelden van geannoteerde en blanco derde correctie is significant ($p = .022$). Het verschil is echter slechts zes tiende scorepunt (oftewel .07 cijferpunt).
- Het gemiddelde van eerste correctie verschilt significant van dat van geannoteerde derde correctie ($p < .001$) en blanco derde correctie ($p < .001$). Eerste correctie leidt in vergelijking met geannoteerde en blanco derde correctie tot een respectievelijk 1.4 en 2.0 punten hogere score (oftewel .23 en .30 cijferpunt).
- Het verschil tussen de gemiddelden van eerste en gezamenlijke correctie is significant ($p = .022$). De gezamenlijke score is echter gemiddeld slechts een kwart scorepunt lager dan de eerste correctie (oftewel .03 cijferpunt).

Verschillen tussen vier correctieprocedures in spreiding wiskunde

De gemiddelde absolute afwijking (GAA) voor de vier correctieprocedures zijn weergegeven in het middelste gedeelte van Tabel 27:

- Het gemiddelde absolute verschil tussen geannoteerde en blanco derde correctie bedraagt 3.3 scorepunten (negatief dan wel positief). Dit komt overeen met 4% van de maximumscore van 75 (oftewel .39 cijferpunt).
- De scores van de eerste correctie wijken gemiddeld 2.9 punten af van geannoteerde derde correcties en gemiddeld 3.6 punten van blanco derde correcties. Dit komt overeen met respectievelijk 4% en 5% van de maximumscore (oftewel .35 en .43 cijferpunt).
- Voor eerste correctie en gezamenlijke eerste en tweede correctie bedraagt de gemiddelde absolute afwijking .5. Dit wil zeggen dat deze twee scores slechts een halve scorepunt uit elkaar liggen (oftewel .06 cijferpunt).

Tabel 28 laat zien hoe de GAA's verdeeld zijn (waarbij waarden van vijf en hoger zijn samengenomen). Kijken we bijvoorbeeld naar de kolom voor absolute afwijkingen van nul, dan blijkt dat de totaalscores van de eerste en de gezamenlijke correctie bij 71% van de nagekeken examenwerken volledig identiek zijn. Daarentegen bedraagt het percentage exacte overeenstemming tussen eerste en derde corrector slechts 8% (geannoteerd) en 7% (blanco). De kolom met de GAA's van vijf of hoger vertoont het omgekeerd beeld: voor eerste en gezamenlijke correctie heeft 1% van de GAA's een waarde van vijf of hoger, voor eerste versus derde geannoteerd is het 22% en voor eerste versus blanco derde correctie gaat het om 31%.

Tabel 28

Verdeling van de gemiddelde absolute afwijking voor vier correctieprocedures wiskunde

		0	1	2	3	4	5 of hoger	Totaal
Eerste correctie minus gezamenlijke score	Aantal	69	13	10	3	1	1	97
	Percentage	71%	13%	10%	3%	1%	1%	100%
Eerste correctie minus derde correctie geannoteerd	Aantal	8	26	21	15	6	21	97
	Percentage	8%	27%	22%	15%	6%	22%	100%
Eerste correctie minus derde correctie blanco	Aantal	7	13	16	17	14	30	97
	Percentage	7%	13%	16%	18%	14%	31%	100%
Derde correctie geannoteerd minus blanco	Aantal	9	21	12	16	17	27	102
	Percentage	9%	21%	12%	16%	17%	26%	100%

Cijfers en percentage onvoldoendes centraal schriftelijk examen wiskunde

In het voorgaande is vastgesteld dat de correctieprocedures niet alle vier tot dezelfde totaalscores op het centraal examen leiden. In het onderste gedeelte van Tabel 27 is te zien tot welke cijfers en percentages onvoldoende de vier correctieprocedures geleid zouden hebben. In de omzetting van scores naar cijfers is de normeringsterm 0,6 van het eerste tijdvak gebruikt. De inzet van de tweede corrector leidt tot een kleine verhoging van het percentage onvoldoendes van 23% naar 26%. Bij geannoteerde en blanco correctie zou respectievelijk 27% en 29% van de kandidaten een onvoldoende hebben behaald.

Overeenstemming tussen de vier correctieprocedures: totaalscore wiskunde

De mate waarin de vier correctieprocedures tot vergelijkbare totaalscores op het examen wiskunde leiden, is ook bepaald aan de hand van de intraclasscorrelatie ICC (zie Tabel 29):

- Geannoteerde en blanco derde correctie lijken tot in hoge mate vergelijkbare examenscores te leiden. De correlatie is met .94 zeer hoog.
- De ICC voor de totaalscore of gemiddelde score van beide correctieprocedures is niet veel hoger dan voor de individuele score (.94 versus .97). Gezien de grootte van de hier niet gerapporteerde standaardfouten moet het verschil inderdaad als betekenisloos worden beschouwd.
- Eerste correctie hangt ongeveer even sterk samen met geannoteerde derde correctie dan met blanco derde correctie (.94 versus .92).
- Eerst correctie hangt niet sterker samen met geannoteerde derde correctie dan met de eerste en de gezamenlijk overeengekomen correcties hangen nagenoeg perfect samen (na afronding voor één correctieprocedure 1.00 en voor beide eveneens 1.00). Kennelijk zijn deze twee correctieprocedures in statistisch opzicht niet of nauwelijks van elkaar te onderscheiden en voegt de een zeer weinig toe aan de ander.

Tabel 29
Overeenstemming over de totaalscores (ICC) wiskunde

Overeenstemming totaalscores	Aantal examenwerken	Correlatie	Intraclasscorrelatie	
			Individueel	Gezamenlijk
Eerste correctie en gezamenlijke correctie	97	1.00	1.00	1.00
Eerste correctie en derde correctie geannoteerd	97	.95	.94	.97
Eerste correctie en derde correctie blanco	97	.93	.92	.96
Derde correctie geannoteerd en blanco	102	.94	.94	.97

Overeenstemming tussen de vier correctieprocedures: scores per opgave wiskunde

De mate waarin de onderscheiden correctieprocedures tot vergelijkbare itemscores leiden, is bepaald aan de hand van *Cohen's Kappa* (zie Tabel 30). Omdat de scholen de scores van de tweede corrector en de gezamenlijk overeengekomen score vrijwel nooit per opgave registreren, kan de eerste correctie niet met deze beide correcties in verband worden gebracht en ontbreken deze gegevens in de tabel. Te zien is dat de overeenstemming tussen de drie correctieprocedures ook voor de scores op de individuele opgaven zeer hoog is. De eerste correctie hangt wat hoger samen met geannoteerde derde correctie dan met blanco derde correctie (.79 versus .72).

Tabel 30
Overeenstemming over de scores per opgave wiskunde

	Aantal scores	Percentage exacte overeenstemming	Cohen's Kappa	Standaardfout
Eerste correctie en derde correctie geannoteerd	2325	84	.79	.010
Eerste correctie en derde correctie blanco	2328	78	.72	.011
Derde correctie geannoteerd en blanco	2445	78	.72	.011

Overeenstemming tussen drie correctieprocedures per type opgave wiskunde

In de hiervoor besproken analyse is geen onderscheid gemaakt naar type opgave. Het is goed denkbaar dat de samenhang tussen de correctieprocedures voor het ene opgaventype anders is dan voor het andere opgaventype. Zo zou de samenhang tussen de drie correctieprocedures bij objectief scoorbare gesloten opgaven hoger kunnen zijn dan bij open opgaven. Vandaar dat de mate waarin de drie correcties tot dezelfde itemscores leiden ook per opgaventype is gerapporteerd (zie Tabel 31):

- Bij alle vraagtypen is het percentage exacte overeenstemming tussen eerste en geannoteerde derde correctie hoger dan tussen eerste en blanco derde correctie.
- Alleen bij de lang-antwoordvragen is de voor toevallige overeenstemming gecorrigeerde samenhang tussen eerste en geannoteerde derde correctie significant hoger dan tussen eerste en blanco derde correctie (aangezien de 95%-betrouwbaarheidsintervallen elkaar niet overlappen).

Tabel 31

Samenhang tussen drie correctieprocedures per opgaventype wiskunde

	Aantal scores	Percentage exacte overeenstemming	Cohen's Kappa	Standaardfout
Classificatievragen				
Eerste correctie en derde correctie geannoteerd	97	86	.81	.046
Eerste correctie en derde correctie blanco	97	82	.77	.049
Derde correctie geannoteerd en blanco	102	76	.70	.054
In- of aanvulvragen				
Eerste correctie en derde correctie geannoteerd	194	71	.61	.042
Eerste correctie en derde correctie blanco	194	64	.54	.044
Derde correctie geannoteerd en blanco	204	61	.49	.043
Lang-antwoordvragen				
Eerste correctie en derde correctie geannoteerd	2034	85	.80	.010
Eerste correctie en derde correctie blanco	2037	79	.73	.011
Derde correctie geannoteerd en blanco	2139	80	.72	.011

Verskil in correctietijd wiskunde

De gemiddelde correctietijd per corrector per versie is weergegeven in Tabel 32. De standaarddeviaties voor corrector 2 en 3 zijn aanzienlijk kleiner dan die van de overige correctoren. Beide correctoren blijken te hebben volstaan met het registreren van de gemiddelde tijdsbesteding in plaats van de tijdsbesteding gedetailleerd per werk te noteren. De reden is dat zij de set examenwerken horizontaal in plaats van verticaal hebben nagekeken (dit wil zeggen voor alle kandidaten eerst opgave 1, dan opgave 2 et cetera in plaats van eerst alle opgaven van kandidaat 1, dan alle opgaven van kandidaat 2 enzovoorts). Onder meer is te zien dat corrector 1 voor het nakijken van een geannoteerd examenwerk gemiddeld veertien minuten nodig had en voor een blanco werk gemiddeld zestien minuten. De vijf correctoren hadden per geannoteerde examenwerk gemiddeld 12.29 minuten nodig en per blanco werken gemiddeld 14.35 minuten. Met behulp van multivariate variantie-analyse (MVA) is nagegaan of dit verschil statistisch significant is en dat bleek het geval ($p < .001$). Een geannoteerd werk kost gemiddeld twee minuten en vier seconden minder correctietijd dan een blanco examenwerk.

Tabel 32

Correctietijd per corrector voor geannoteerde en blanco examenwerken wiskunde

Corrector	Geannoteerde examenwerken		Blanco examenwerken	
	Gemiddelde	Standaard deviatie	Gemiddelde	Standaard deviatie
1	13.65	1.57	16.00	4.21
2	9.00	.00	9.30	1.34
3	10.00	.00	10.14	.65
4	12.29	1.95	14.35	2.80
5	16.70	3.06	21.81	5.55
Gemiddelde	12.29	3.24	14.35	5.67

Tabel 32 doet vermoeden dat de correctoren verschillen in de gemiddelde correctietijd en dat het verschil tussen de correctietijd van geannoteerd en blanco werk voor de ene corrector groter is dan voor de andere. Vandaar dat een tweede MVA is uitgevoerd waarbij we corrector als tussen-subjecten factor aan het model toevoegen. De MVA stelt ons in staat om het hoofdeffect van corrector (tussen-subjecten) en de interactie tussen tijd en corrector (binnen-subjecten) te bestuderen. Zowel het hoofdeffect van corrector als het interactie-effect blijkt statistisch significant (beide $p < .001$). Kennelijk verschillen correctoren in de hoeveelheid geïnvesteerde correctietijd en is het verschil tussen de correctietijd van geannoteerde en blanco examenwerken voor de ene corrector groter dan voor de andere. Ten aanzien van het hoofdeffect van corrector laat de tabel zien dat corrector 2 en 3 minder tijd nodig hadden dan de overige correctoren. De interactie tussen corrector en versie zit hem vooral in de corrector 1 en 5. Zij besteden naar verhouding weinig tijd aan geannoteerde werken.

Vanwege de grote verschillen in tijdsbesteding tussen correctoren en de interactie met de versie van het examenwerk mag aan het eerder gerapporteerde verschil in gemiddelde correctietijd tussen geannoteerde en blanco derde correctie geen grote betekenis worden gehecht.

Cijfer op het schoolexamen wiskunde

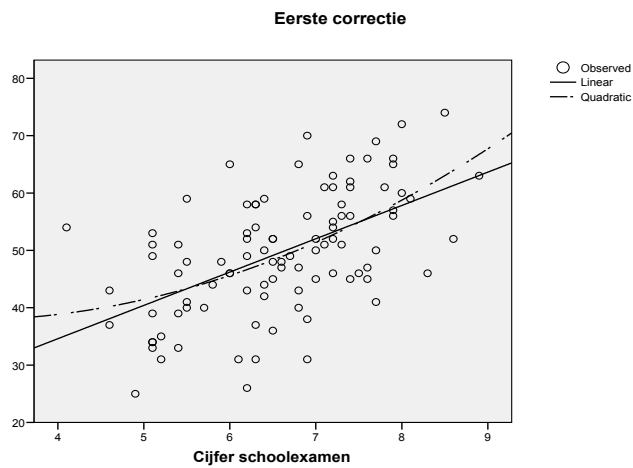
Nagegaan is of de correcties samenhangen met het cijfer voor het schoolexamen en zo ja in hoeverre deze relatie verschilt per correctieprocedure. Tabel 33 geeft informatie over de verdeling van de drie correctieprocedures en het schoolexamencijfer.

Tabel 33

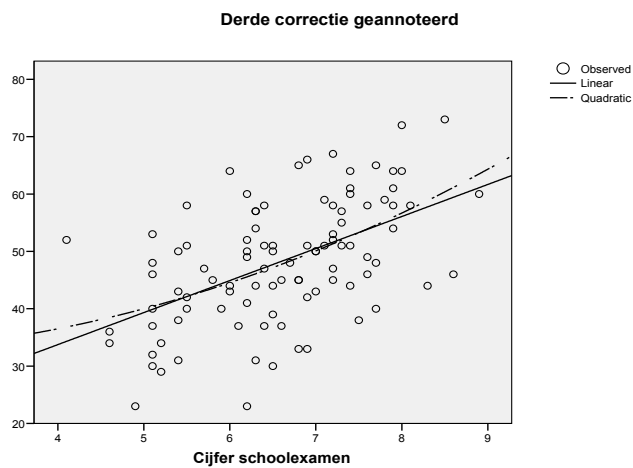
Karakterisering van de scoreverdeling voor drie correctieprocedures en het schoolexamencijfer

	Aantal examen- werken	Min	Max	Gemid- delde	Standaard deviatie	Scheefheid	
						Statistic	SE
Eerste correctie	97	25	74	49.65	10.70	-.05	.24
Derde correctie geannoteerd	102	22	73	47.76	11.21	-.17	.24
Derde correctie blanco	102	23	73	47.19	11.14	-.05	.24
Cijfer schoolexamen	97	4.1	8.9	6.59	1.01	-.14	.24

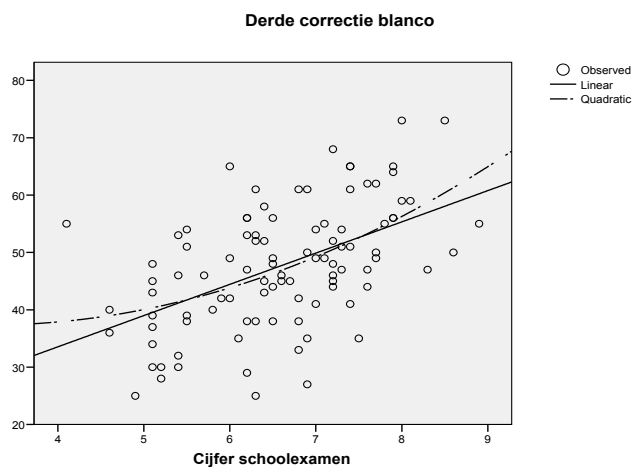
De scores voor de drie correctieprocedures en het cijfer op het schoolexamen lijken redelijk normaal verdeeld. De spreidingsdiagrammen in Figuur 3a, 3b en 3c bevestigen dit beeld.



Figuur 3a Spreidingsdiagram eerste correctie versus cijfer schoolexamen wiskunde



Figuur 3b Spreidingsdiagram geannoteerde derde correctie versus cijfer schoolexamen wiskunde



Figuur 3c Spreidingsdiagram geannoteerde blanco correctie versus cijfer schoolexamen wiskunde

De resultaten van de multiple regressieanalyse zijn weergegeven in Tabel 34. Weergegeven zijn de regressiegewichten met bijbehorende standaardfouten van model 1 met alleen een lineaire term voor het schoolexamencijfer en van model 2 met zowel een lineaire als een kwadratische term. De tabel geeft ook R, de multiple correlatie, die in het geval van model 1 gelijk is aan de product-moment correlatie. De lineaire samenhang tussen de prestaties op het centraal examen en het schoolexamen is in alle drie gevallen zwak (een toename van het schoolexamencijfer met één punt gaat gepaard met een toename van de score op het centraal schriftelijk examen met respectievelijk 5.8 (eerste correctie), 5.6 (derde correctie geannoteerd) en 5.5 scorepunten (derde correctie blanco). Er zijn geen aanwijzingen voor het veronderstelde kromlijng verband tussen de correcties en het schoolexamencijfer. In model 2 is noch de lineaire noch de kwadratische term significant, terwijl de multiple correlatie gelijk of niet veel hoger is dan die van model 1.

Tabel 34

Uitkomsten van de analyse van de regressie van de totaalscore op het schoolexamencijfer wiskunde

		Ongestandaardiseerde regressiegewichten		Gestandaardiseerde regressiegewichten	T	Sig.	R
		b	se	Beta			
Eerste correctie							
Model 1	Lineair	5.80	.91	.55	6.40	<.001	.55
Model 2	Lineair	-4.54	9.95	-.43	-.46	.649	.56
	Kwadratisch	.79	.76	.98	1.04	.299	
Derde correctie geannoteerd							
Model 1	Lineair	5.58	.93	.53	6.02	<.001	.53
Model 2	Lineair	-1.16	1.19	-.11	-.11	.909	.53
	Kwadratisch	.52	.78	.64	.66	.508	
Derde correctie blanco							
Model 1	Lineair	5.45	.94	.51	5.82	<.001	.51
Model 2	Lineair	-5.16	1.26	-.49	-.50	.616	.52
	Kwadratisch	.81	.78	1.00	1.04	.302	

3.4 Biologie VWO

Het examen biologie VWO bevat 35 opgaven en kent een maximumscore van 73. Het examen bestaat uit veertien meerkeuzevragen, één in- of aanvulvraag, vijf kort-antwoordvragen en vijftien lang-antwoordvragen. In totaal hebben de derde correctoren beide versies van 115 examenwerken **gecorrigeerd**. Vijftien waren onbruikbaar omdat ze van een ander biologie-examen afkomstig waren. Van de honderd resterende examenwerken bleken er vijftien geen annotaties van de eerste corrector te bevatten. Voor de vergelijking van geannoteerde en blanco derde correctie zijn derhalve 85 examenwerken beschikbaar.

Tabel 35 toont enkele beschrijvende kenmerken van vier correctieprocedures, te weten eerste correctie, gezamenlijke eerste en tweede correctie, derde correctie van geannoteerd examenwerk en derde correctie van blanco examenwerk. De vier correctieprocedures zijn vergeleken in termen van de totaalscore op het examen, het gemiddeld verschil tussen de totaalscores, de gemiddelde absolute afwijking (d.w.z. het aantal scorepunten dat de correctieprocedures gemiddeld van elkaar verschillen waarbij het er niet toe doet of het verschil positief of negatief is), het cijfer voor het eerste tijdvak en het percentage onvoldoendes waartoe de correctieprocedure geleid zou hebben. De spreidingsdiagrammen zijn weergegeven in Bijlage 4.

Tabel 35

Enkele beschrijvende gegevens voor de vier correctieprocedures biologie

	Aantal examenwerken	Minimum	Maximum	Gemiddelde/ percentage	Standaard deviatie
Totaalscore					
Eerste correctie	85	23	61	44.32	8.64
Eerste en tweede correctie gezamenlijk	85	23	61	43.94	8.65
Derde correctie geannoteerd	85	22	60	40.33	8.30
Derde correctie blanco	85	22	54	40.13	8.18
Verskil tussen de totaalscores					
Eerste correctie minus gezamenlijke score	85	-2	7	.38	1.27
Eerste correctie minus derde correctie geannoteerd	85	-3	11	3.99	2.91
Eerste correctie minus derde correctie blanco	85	-4	16	4.19	4.17
Derde correctie geannoteerd minus blanco	85	-8	8	.20	3.16
Gemiddelde absolute afwijking					
Eerste correctie versus gezamenlijke score	85	0	7	.68	1.14
Eerste correctie versus derde correctie geannoteerd	85	0	11	4.20	2.59
Eerste correctie minus derde correctie blanco	85	1	16	4.82	3.41
Derde correctie geannoteerd versus blanco	85	0	8	2.46	1.98
Cijfer eerste tijdvak					
Eerste correctie	85	4.10	8.80	6.76	1.07
Gezamenlijke correctie	85	4.10	8.80	6.72	1.07
Derde correctie geannoteerd	85	4.00	8.70	6.27	1.02
Derde correctie blanco	85	4.00	8.00	6.24	1.01
Percentage onvoldoendes					
Eerste correctie	85			15%	
Gezamenlijke correctie	85			16%	
Derde correctie geannoteerd	85			22%	
Derde correctie blanco	85			26%	

Verschillen tussen de vier correctieprocedures in gemiddelde biologie

Kijken we eerst naar de totaalscore op het examen biologie, dan zien we dat de examenkandidaten volgens de eerste correctoren gemiddeld ruim 44 punten hebben behaald (van maximaal 73 punten). De verschillen tussen de gemiddelden van de vier correctieprocedures zijn getoetst met behulp van een multivariate variantieanalyse. De belangrijkste uitkomsten vatten we als volgt samen:

- Het verschil tussen de gemiddelden van geannoteerde en blanco derde correctie is met een vijfde scorepunt miniem en dan ook niet significant ($p = .561$). Uitgedrukt in cijfers gaat het om .03 punt.
- Het gemiddelde van eerste correctie verschilt significant van dat van geannoteerde derde correctie ($p < .001$) en blanco derde correctie ($p < .001$). Eerste correctie leidt in vergelijking met geannoteerde en blanco derde correctie tot een respectievelijk 4.0 en 4.2 punten hogere score (hetgeen overeenkomt met .49 en .52 cijferpunt).

- De gemiddelden van de eerste correctie en de gezamenlijk door de eerste en tweede corrector toegekende score wijken niet significant van elkaar af ($p = .208$). Het verschil is zeer klein: de gezamenlijke score is gemiddeld slechts vier tiende lager dan die van de eerste correctie (hetgeen overeenkomt met .04 cijferpunt).

Verschillen tussen vier correctieprocedures in spreiding biologie

De gemiddelde absolute afwijking (GAA) voor de vier correctieprocedures zijn weergegeven in het middelste gedeelte van Tabel 35. Een maat die laat zien hoe sterk correctoren of correctieprocedures van elkaar afwijken is de gemiddelde absolute afwijking. Deze spreidingsmaat is in ons geval het gemiddelde van de absolute waarde van het verschil tussen de scores van de beide correctieprocedures (dat wil zeggen dat indien het verschil negatief is het minteken is weggelaten). De belangrijkste uitkomsten zijn:

- De scores voor geannoteerde en blanco derde correctie wijken gemiddeld 2.5 punten van elkaar af (negatief dan wel positief). Dit komt overeen met 3% van de maximumscore van 73 (oftewel .30 cijferpunt).
- De totaalscores van de eerste correctie wijken gemiddeld 4.2 punten af van geannoteerde derde correctie en gemiddeld 4.8 punten van blanco derde correctie. Dit komt overeen met respectievelijk 6% en 7% van de maximumscore (oftewel .52 en .59 cijferpunt).
- Voor eerste correctie en gezamenlijke eerste en tweede correctie bedraagt de gemiddelde absolute afwijking .70. Dit wil zeggen dat deze twee scores gemiddeld minder dan één scorepunt uit elkaar liggen (negatief dan wel positief). Dit komt overeen met .08 cijferpunt.

Tabel 36 laat zie hoe de GAA's verdeeld zijn (waarbij waarden van vijf en hoger zijn samengenomen). Kijken we bijvoorbeeld naar de kolom voor absolute afwijkingen van nul, dan blijkt dat de totaalscores van de eerste en de gezamenlijke correctie bij 58% van de nagekeken examenwerken volledig identiek zijn. Daarentegen bedraagt het percentage exacte overeenstemming tussen eerste en derde corrector slechts 4% (geannoteerd) en 0% (blanco). De kolom met de GAA's van vijf of hoger vertoont het omgekeerd beeld: voor eerste en gezamenlijke correctie heeft 2% van de GAA's een waarde van vijf of hoger, voor eerste versus geannoteerde derde correctie is het 38% en voor eerste versus blanco derde correctie gaat het om 48%. Het patroon van uitkomsten maakt aannemelijk dat de eerste corrector het doorgaans sterker eens is met de tweede dan met de derde corrector en dat het verschil tussen eerste correctie en geannoteerde derde correctie wat kleiner is dan dat tussen eerste correctie en blanco derde correctie.

Tabel 36

Verdeling van de gemiddelde absolute afwijking voor vier correctieprocedures biologie

		0	1	2	3	4	5 of hoger	Totaal
Eerste correctie minus gezamenlijke score	Aantal	49	24	8	2	0	2	85
	Percentage	58%	28%	9%	2%	0%	2%	100%
Eerste correctie minus derde correctie geannoteerd	Aantal	3	10	12	13	15	32	85
	Percentage	4%	12%	14%	15%	18%	38%	100%
Eerste correctie minus derde correctie blanco	Aantal	0	15	12	11	6	41	85
	Percentage	0%	18%	14%	13%	7%	48%	100%
Derde correctie geannoteerd minus blanco	Aantal	9	26	16	12	11	11	85
	Percentage	11%	31%	19%	14%	13%	13%	100%

Cijfers en percentage onvoldoendes centraal examen biologie

In het voorgaande is vastgesteld dat de vier correctieprocedures niet alle tot dezelfde examenscores leiden. In het onderste gedeelte van Tabel 35 is te zien tot welke cijfers en percentages onvoldoende

de vier correctieprocedures geleid zouden hebben. In de omzetting van scores naar cijfers is de normeringsterm 1.3 van het eerste tijdvak gebruikt. De inbreng van de tweede corrector leidt tot een verhoging van het percentage onvoldoendes van 15% naar 16% en bij geannoteerde en blanco correctie zou respectievelijk 22% en 26% van de kandidaten een onvoldoende hebben behaald.

Overeenstemming tussen de vier correctieprocedures: totaalscore biologie

De mate waarin de vier correctieprocedures tot vergelijkbare totaalscores op het examen leiden, is ook bepaald aan de hand van de intraclasscorrelatie ICC (zie Tabel 37).

- Geannoteerde en blanco derde correctie hangen zeer sterk samen (de ICC voor één en beide procedures is respectievelijk .93 en .96).
- Eerste correctie lijkt sterker samen met geannoteerde dan met blanco derde correctie (.85 versus .78).
- De samenhang tussen eerste en gezamenlijke correctie lijkt hoger dan die tussen eerste en geannoteerde derde correctie (.99 versus .85) en die tussen eerste en blanco derde correctie (.99 versus .78).
- Eerste en gezamenlijke correcties hangen nagenoeg perfect samen (.99 voor één en .99 voor beide). Kennelijk zijn deze twee correctieprocedures in statistisch opzicht niet of nauwelijks van elkaar te onderscheiden en voegt de een niets toe aan de ander.

Tabel 37

Overeenstemming over de totaalscores (ICC) biologie

Overeenstemming totaalscores	Aantal examenwerken	Correlatie	Intraclasscorrelatie	
			Individueel	Gezamenlijk
Eerste correctie en gezamenlijke correctie	85	,99	.99	.99
Eerste correctie en derde correctie geannoteerd	85	,94	.85	.92
Eerste correctie en derde correctie blanco	85	,88	.78	.88
Derde correctie geannoteerd en blanco	85	,93	.93	.96

Overeenstemming tussen de vier correctieprocedures: scores per opgave biologie

De mate waarin de correctieprocedures tot vergelijkbare itemscores leiden, is bepaald aan de hand van *Cohen's Kappa* (zie Tabel 38). Omdat de scholen de scores van de tweede corrector en de gezamenlijk overeengekomen score vrijwel nooit per opgave registreren, kan de eerste correctie niet met deze beide correcties in verband worden gebracht en ontbreken deze gegevens in de tabel. Te zien is dat de percentages exacte overeenstemming voor de scores op de individuele opgaven voor alle drie correctieprocedures hoog zijn. Van *Cohen's Kappa*, een maat die corrigeert voor toevallige overeenstemming, zijn de waarden hoog. Eerste correcties hangen zoals verwacht hoger samen met geannoteerde dan met blanco derde correcties (.78 versus .71).

Tabel 38
Overeenstemming over de scores per opgave

	Aantal scores	Percentage exacte overeenstemmi ng	Cohen's Kappa	Standaardfout
Eerste correctie en derde correctie geannoteerd	2975	85	.78	.010
Eerste correctie en derde correctie blanco	2975	81	.71	.011
Derde correctie geannoteerd en blanco	2975	86	.79	.009

Overeenstemming tussen drie correctieprocedures per type opgave biologie

In de hiervoor besproken analyse is geen onderscheid gemaakt naar type opgave. Het is goed denkbaar dat de samenhang tussen de correctieprocedures voor het ene opgaventype anders is dan voor het andere opgaventype. Zo zou de samenhang tussen de drie correctieprocedures bij objectief scoorbare meerkeuzeopgaven hoger kunnen zijn dan bij open opgaven. Vandaar dat de mate waarin de drie correcties tot dezelfde itemscores leiden ook per opgaventype wordt gerapporteerd (zie Tabel 39):

- De percentages exacte overeenstemming en de voor toevallige overeenstemming gecorrigeerde samenhang (*Kappa*) is hoger voor meerkeuzevragen dan voor kort-antwoordvragen dan voor lang-antwoordvragen.
- Voor meerkeuzevragen leiden de drie correctieprocedures tot nagenoeg identieke scores. Kennelijk maakt het voor dit opgaventype weinig uit of de opgaven door eerste of derde correctoren worden nagekeken of dat de derde correctie plaatsvindt op basis van geannoteerd dan wel blanco examenmateriaal (afgezien van sporadische slordigheidsfoutjes).
- Bij de kort-antwoordvragen lijkt de voor toevallige overeenstemming gecorrigeerde samenhang tussen eerste en derde correcties wat hoger voor geannoteerde dan voor blanco derde correcties (.86 versus .78). Gezien de grootte van de standaardfouten kan dit verschil echter als niet significant worden beschouwd ($p < .05$).
- Alleen bij de lang-antwoordvragen hangt eerste correctie na correctie voor toevallige overeenstemming hoger samen met geannoteerde dan met blanco derde correctie (.59 versus .46).
- De overeenstemming bij de in- en aanvulvragen is relatief laag. Dit opgaventype was echter slechts met één opgave in het onderzoek vertegenwoordigd, zodat de zeggingskracht van deze bevinding gering is.

Tabel 39
Samenhang tussen drie correctieprocedures per opgaventype biologie

	Aantal scores	Percentage exacte overeen- stemming	Cohen's Kappa	Standaard- fout
Meerkeuzevragen				
Eerste correctie en derde correctie geannoteerd	1190	98	.96	.008
Eerste correctie en derde correctie blanco	1190	98	.96	.008
Derde correctie geannoteerd en blanco	1190	1.00	1.00	.003
In- en aanvulvragen				
Eerste correctie en derde correctie geannoteerd	85	83	.68	.064
Eerste correctie en derde correctie blanco	85	80	.63	.065
Derde correctie geannoteerd en blanco	85	89	.81	.058
Kort-antwoordvragen				
Eerste correctie en derde correctie geannoteerd	425	91	.85	.023
Eerste correctie en derde correctie blanco	425	87	.78	.027
Derde correctie geannoteerd en blanco	425	92	.86	.023
Lang-antwoordvragen				
Eerste correctie en derde correctie geannoteerd	1275	72	.59	.018
Eerste correctie en derde correctie blanco	1275	63	.46	.019
Derde correctie geannoteerd en blanco	1275	71	.58	.019

Verschil in correctietijd biologie

De gemiddelde correctietijd per corrector per versie is weergegeven in Tabel 40. Corrector 1 doet bijvoorbeeld over het nakijken van geannoteerde examenwerken gemiddeld 31 minuten per werk en voor de blanco werken is het gemiddeld 32 minuten. De vijf correctoren hadden voor de geannoteerde examenwerken gemiddeld 15.69 minuten nodig en voor de blanco examenwerken 17.82 minuten. Met behulp van multivariate variantie-analyse (MVA) is nagegaan of dit verschil in correctietijd statistisch significant is en dat bleek het geval ($p < .001$). Een geannoteerd werk vergt dus twee minuten en acht seconden minder correctietijd dan een in alle andere opzichten vergelijkbaar blanco examenwerk.

Tabel 40

Correctietijd per corrector voor geannoteerde en blanco examenwerken biologie

Corrector	Geannoteerde examenwerken		Blanco examenwerken	
	Gemiddelde	Standaard deviatie	Gemiddelde	Standaard deviatie
1	30.82	4.85	32.00	.00
2	9.06	2.44	11.41	3.48
3	15.06	3.07	18.53	4.93
4	17.53	2.94	19.06	2.33
5	6.00	1.46	8.12	2.32
Gemiddelde	15.69	9.19	17.82	8.80

Corrector 1 blijkt bij de blanco werken alleen de gemiddelde tijdsbesteding te hebben geregistreerd in plaats van de tijdsbesteding gedetailleerd per werk te noteren. De reden is dat deze corrector de set examenwerken horizontaal in plaats van verticaal heeft nagekeken (dit wil zeggen voor alle kandidaten eerst opgave 1, dan opgave 2 et cetera in plaats van eerst alle opgaven van kandidaat 1, dan alle opgaven van kandidaat 2 enzovoorts). Tabel 40 doet vermoeden dat de correctoren verschillen in de gemiddelde correctietijd en dat het verschil tussen de correctietijd van geannoteerd en blanco werk voor de ene corrector groter is dan voor de andere. Zowel het hoofdeffect van corrector als het interactie-effect tussen tijd en corrector blijkt statistisch significant ($p < .001$). Kennelijk verschillen correctoren in de hoeveelheid geïnvesteerde correctietijd en is het verschil tussen de correctietijd van geannoteerde en blanco examenwerken voor de ene corrector groter dan voor de andere. Corrector 1 besteedt bij voorbeeld meer tijd aan het nakijken dan de correctoren 2 en 5 en bij corrector 3 gaat het nakijken van geannoteerde werken in vergelijking met de andere correctoren naar verhouding wat sneller dan blanco werken.

Vanwege de grote verschillen in tijdsbesteding tussen correctoren en de interactie met de versie van het examenwerk mag aan het eerder gerapporteerde verschil in gemiddelde correctietijd tussen geannoteerde en blanco derde correctie geen grote betekenis worden gehecht.

Cijfer op het schoolexamen biologie

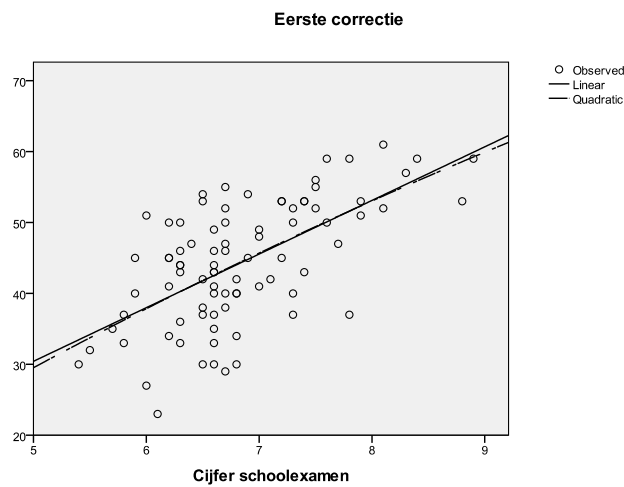
Nagegaan is of hoeverre de correcties samenhangen met het cijfer voor het schoolexamen en zo ja in hoeverre deze relatie verschilt per correctieprocedure. Tabel 41 geeft informatie over de verdeling van de drie correctieprocedures en het schoolexamencijfer.

Tabel 41

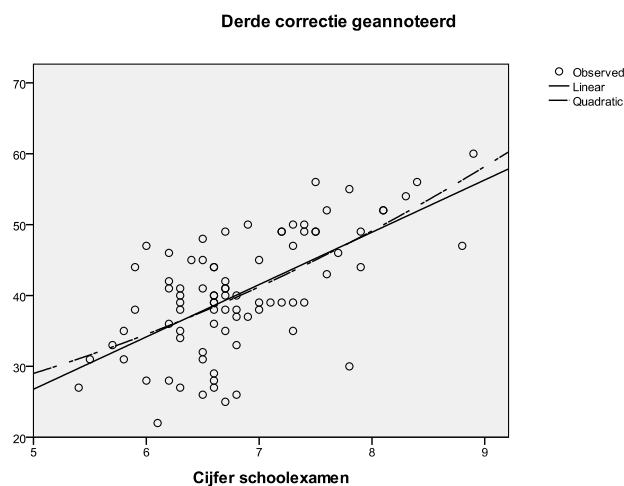
Karakterisering van de scoreverdeling voor drie correctieprocedures en het schoolexamencijfer biologie

	Aantal examenwerken	Min	Max	Gemiddelde	Standaard deviatie	Scheefheid	
						Statistic	SE
Eerste correctie	85	23	61	44.32	8.64	-.20	.26
Derde correctie geannoteerd	85	22	60	40.33	8.30	-.01	.26
Derde correctie blanco	85	22	54	40.13	8.18	-.21	.26
Cijfer schoolexamen	85	5.4	8.9	6.84	.71	.68	.26

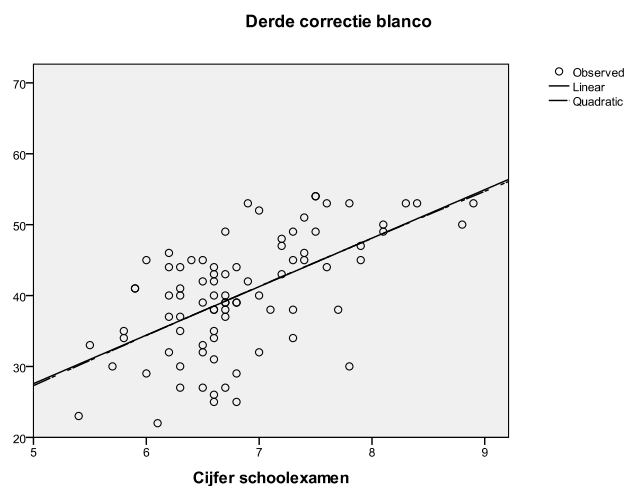
De scores voor de drie correctieprocedures en de schoolexamencijfers lijken redelijk normaal verdeeld. Cijfers lager dan vijf komen in de responsegroep niet voor. De spreidingsdiagrammen in Figuur 4a, 4b en 4c bevestigen dit beeld.



Figuur 4a Spreidingsdiagram eerste correctie versus cijfer schoolexamen biologie



Figuur 4b Spreidingsdiagram geannoteerde derde correctie versus cijfer schoolexamen biologie



Figuur 4c Spreidingsdiagram geannoteerde blanco correctie versus cijfer schoolexamen biologie

De resultaten van de multiple regressieanalyse zijn weergegeven in Tabel 42. Weergegeven zijn de regressiegewichten met bijbehorende standaardfouten van model 1 met alleen een lineaire term voor het schoolexamencijfer en van model 2 met zowel een lineaire als een kwadratische term. De tabel geeft ook R, de multiple correlatie, die in het geval van model 1 gelijk is aan de product-moment correlatie. De lineaire samenhang tussen de prestaties op het centraal examen en het schoolexamen is in alle drie gevallen relatief sterk (een toename van het schoolexamencijfer met één punt gaat gepaard met een toename van de score op het centraal schriftelijk examen met respectievelijk 7.6 (eerste correctie), 7.4 (derde correctie geannoteerd) en 6.8 scorepunten (derde correctie blanco). Er zijn geen aanwijzingen voor het veronderstelde kromlijng verband tussen de correcties en het schoolexamencijfer. In model 2 is de kwadratische term geen enkele keer significant, terwijl de multiple correlatie telkens gelijk is aan die van model 1.

Tabel 42

Uitkomsten van de analyse van de regressie van de totaalscore op het schoolexamencijfer biologie

		Ongestandaardiseerde regressiegewichten		Gestandaardiseerde regressiegewichten	T	Sig.	R
		b	se	Beta			
Eerste correctie							
Model 1	Lineair	7.56	1.05	.62	7.21	<.001	.62
Model 2	Lineair	1.97	15.25	.90	.72	.474	.62
	Kwadratisch	-.24	1.08	-.28	-.22	.823	
Derde correctie geannoteerd							
Model 1	Lineair	7.38	1.00	.63	7.40	<.001	.63
Model 2	Lineair	-1.05	14.48	-.09	-.07	.942	.63
	Kwadratisch	.60	1.02	.72	.58	.561	
Derde correctie blanco							
Model 1	Lineair	6.84	1.02	.59	6.72	<.001	.59
Model 2	Lineair	7.88	14.82	.68	.53	.596	.59
	Kwadratisch	-.07	1.04	-.09	-.07	.944	

3.5 Geschiedenis HAVO

Het examen geschiedenis HAVO bevat dertig opgaven met een maximumscore van 76. Het examen bestaat uit één combinatievraag, twee ordeningsvragen, drie kort-antwoordvragen, vijf kort-langantwoordvragen en negentien lang-antwoordvragen. Van de 95 examenwerken waren er twintig niet van annotaties voorzien, zodat voor de vergelijking van de correctieprocedures 75 werken beschikbaar zijn.

Tabel 43 toont enkele beschrijvende kenmerken van vier correctieprocedures, te weten eerste correctie, gezamenlijke eerste en tweede correctie, derde correctie van geannoteerde examenwerk en derde correctie van blanco examenwerken. De vier correctieprocedures zijn vergeleken in termen van de totaalscore op het examen, het gemiddeld verschil tussen de totaalscores, de gemiddelde absolute afwijking (d.w.z. het aantal scorepunten dat de correctieprocedures gemiddeld van elkaar verschillen waarbij het er niet toe doet of het verschil positief of negatief is), het cijfer voor het eerste tijdvak en het percentage onvoldoendes waartoe de correctieprocedure geleid zou hebben. De spreidingsdiagrammen zijn weergegeven in Bijlage 5.

Tabel 43

Enkele beschrijvende gegevens voor de vier correctieprocedures geschiedenis

	Aantal examenwerken	Minimum	Maximum	Gemiddelde/ Percentage	Standaard deviatie
Totaalscore					
Eerste correctie	75	22	66	44.31	10.33
Eerste en tweede correctie gezamenlijk	70	23	62	43.81	10.07
Derde correctie geannoteerd	75	16	58	35.83	10.16
Derde correctie blanco	75	9	56	34.89	9.87
Verskil tussen de totaalscores					
Eerste correctie minus gezamenlijke score	70	-6	4	-.13	1.35
Eerste correctie minus derde correctie geannoteerd	75	-2	21	8.48	5.37
Eerste correctie minus derde correctie blanco	75	-12	22	9.41	7.21
Derde correctie geannoteerd minus blanco	75	-23	13	.93	6.77
Gemiddelde absolute afwijking					
Eerste correctie versus gezamenlijke score	70	0	6	.67	1.18
Eerste correctie versus derde correctie geannoteerd	75	0	21	8.59	5.19
Eerste correctie minus derde correctie blanco	75	0	22	1.13	6.14
Derde correctie geannoteerd versus blanco	75	0	23	5.47	4.05
Cijfer eerste tijdvak					
Eerste correctie	75	3.90	9.10	6.54	1.22
Gezamenlijke correctie	70	4.00	8.60	6.49	1.19
Derde correctie geannoteerd	75	3.20	8.20	5.55	1.20
Derde correctie blanco	75	2.40	7.90	5.43	1.16
Percentage onvoldoendes					
Eerste correctie	75			23%	
Gezamenlijke correctie	70			23%	
Derde correctie geannoteerd	75			51%	
Derde correctie blanco	75			55%	

Verschillen tussen de vier correctieprocedures in gemiddelde geschiedenis

Kijken we eerst naar de totaalscore op het examen geschiedenis HAVO, dan zien we dat de examenkandidaten volgens de eerste correctoren gemiddeld 44 punten hebben behaald (van maximaal 76 punten). De verschillen tussen de gemiddelden van de vier correctieprocedures zijn getoetst met behulp van een multivariate variantieanalyse. De belangrijkste resultaten kunnen als volgt worden samengevat:

- Het verschil tussen de gemiddelden van geannoteerde en blanco derde correctie is niet significant ($p = .274$). Het verschil is minder dan één scorepunt (oftewel .12 cijferpunt).
- Het gemiddelde van eerste correctie verschilt significant van dat van geannoteerde derde correctie ($p < .001$) en blanco derde correctie ($p < .001$). Eerste correctie kent in vergelijking met

geannoteerde en blanco derde correctie een respectievelijk 8.5 en 9.4 punten hogere score (oftewel .99 en 1.11 cijferpunt).

- Het verschil tussen de gemiddelden van de eerste corrector en de gezamenlijk overeengekomen score is niet significant ($p = .428$). De gezamenlijke scores verschillen gemiddeld slechts een tiende punt van de scores van de eerste corrector (oftewel .05 cijferpunt).

Verschillen tussen vier correctieprocedures in spreiding geschiedenis

De gemiddelde absolute afwijkingen voor de vier correctieprocedures zijn weergegeven in het middelste gedeelte van Tabel 43. Een maat die laat zien hoe sterk correctoren of correctieprocedures van elkaar afwijken is de gemiddelde absolute afwijking. Deze spreidingsmaat is in ons geval het gemiddelde van de absolute waarde van het verschil tussen de scores van de beide correctieprocedures (dat wil zeggen dat indien het verschil negatief is het minteken is weggelaten). De belangrijkste resultaten kunnen als volgt worden beschreven:

- Het gemiddelde absolute verschil tussen geannoteerde en blanco derde correctie bedraagt 5.5 scorepunten. Dit komt overeen met 7% van de maximumscore van 76 (oftewel .65 cijferpunt).
- De scores van de eerste correctie wijken gemiddeld 8.6 punten af van geannoteerde derde correctie en gemiddeld 10.1 punten van blanco derde correctie. Dit komt overeen met respectievelijk 11% en 13% van de maximumscore (oftewel 1.02 en 1.20 cijferpunt).
- De scores van de eerste corrector en de gezamenlijke tussen eerste en tweede corrector overeengekomen scores liggen dicht bij elkaar. De gemiddelde absolute afwijking bedraagt slechts twee derde punt (oftewel .08 cijferpunt).

Tabel 44 laat zien hoe de GAA's verdeeld zijn (waarbij waarden van vijf en hoger zijn samengenomen). Kijken we bijvoorbeeld naar de kolom voor absolute afwijkingen van nul, dan blijkt dat de totaalscores van de eerste en de gezamenlijke correctie bij 64% van de nagekeken examenwerken volledig identiek zijn. Daarentegen bedraagt het percentage exacte overeenstemming tussen eerste en derde corrector slechts 4% (geannoteerd) en 8% (blanco). De kolom met de GAA's van vijf of hoger vertoont het omgekeerde beeld: voor eerste en gezamenlijke correctie heeft 1% van de GAA's een waarde van vijf of hoger, voor eerste versus derde geannoteerd is het 79% en voor eerste versus blanco derde correctie gaat het eveneens om 79%. Het patroon van uitkomsten maakt aannemelijk dat de eerste corrector het doorgaans sterker eens is met de tweede dan met de derde corrector.

Tabel 44

Verdeling van de gemiddelde absolute afwijking voor vier correctieprocedures geschiedenis

		0	1	2	3	4	5 of hoger	Totaal
Eerste correctie minus gezamenlijke score	Aantal	45	13	7	2	2	1	70
	Percentage	64%	19%	10%	3%	3%	1%	100%
Eerste correctie minus derde correctie geannoteerd	Aantal	3	4	3	1	5	59	75
	Percentage	4%	5%	4%	1%	7%	79%	100%
Eerste correctie minus derde correctie blanco	Aantal	6	2	3	2	3	59	75
	Percentage	8%	3%	4%	3%	4%	79%	100%
Derde correctie geannoteerd minus blanco	Aantal	4	6	4	12	11	38	75
	Percentage	5%	8%	5%	16%	15%	51%	100%

Cijfers en percentage onvoldoendes centraal examen geschiedenis

In het voorgaande is vastgesteld dat de vier correctieprocedures niet alle tot dezelfde totaalscores op het centraal examen leiden. In het onderste gedeelte van Tabel 43 is te zien tot welke cijfers en percentages onvoldoendes de vier correctieprocedures geleid zouden hebben. In de omzetting van scores naar cijfers is de normeringsterm 1,3 van het eerste tijdvak gebruikt. De inzet van de tweede corrector leidt niet tot een verhoging van het percentage onvoldoendes: het blijft 23%. Bij geannoteerde en blanco correctie zou respectievelijk 51% en 55% van de kandidaten een onvoldoende hebben behaald.

Overeenstemming tussen de vier correctieprocedures: totaalscore geschiedenis

De mate waarin de vier correctieprocedures tot vergelijkbare totaalscores op het examen leiden, is ook bepaald aan de hand van de intraclasscorrelatie ICC (zie Tabel 45):

- De samenhang tussen geannoteerde en blanco derde correctie is hoog (voor één correctieprocedure is de ICC .77 en voor beide samen .87). Omdat de hier niet gerapporteerde 95%-betrouwbaarheidsintervallen elkaar overlappen, moet dit verschil echter als betekenisloos worden beschouwd.
- Eerste correctie hangt sterker samen met geannoteerde dan met blanco derde correctie (.64 versus .52).
- De samenhang tussen eerste en gezamenlijke correctie (.99) is aanzienlijk hoger dan die tussen eerste en geannoteerde derde correctie (.64) en die tussen eerste en blanco derde correctie (.52).
- Voor eerste en gezamenlijk overeengekomen correcties is de ICC zeer hoog (voor een correctieprocedure .99 en voor beide 1.00). Kennelijk zijn deze twee correctieprocedures in statistisch opzicht niet of nauwelijks van elkaar te onderscheiden en voegt de gezamenlijk overeengekomen score weinig toe aan de totaalscore van de eerste corrector.

Tabel 45

Overeenstemming over de totaalscores (ICC) geschiedenis

Overeenstemming totaalscores	Aantal examenwerken	Correlatie	Intraclasscorrelatie	
			Individueel	Gezamenlijk
Eerste correctie en gezamenlijke correctie	75	.99	.99	1.00
Eerste correctie en derde correctie geannoteerd	70	.86	.64	.78
Eerste correctie en derde correctie blanco	75	.75	.52	.69
Derde correctie geannoteerd en blanco	75	.77	.77	.87

Overeenstemming tussen de vier correctieprocedures: scores per opgave geschiedenis

De mate waarin de correctieprocedures tot vergelijkbare scores voor de individuele opgaven leiden, is bepaald aan de hand van *Cohen's Kappa* (zie Tabel 46). De scores voor de schrijfopdracht zijn minimaal van intervalniveau (maximumscore 13) en zijn om deze reden afzonderlijk geanalyseerd. Omdat vestigingen de scores van de tweede corrector en de gezamenlijk overeengekomen score vrijwel nooit per opgave registreren, kunnen de derde correcties niet met deze beide correcties in verband worden gebracht en ontbreken deze gegevens in de tabel. Te zien is dat de overeenstemming in vergelijking met de vakken Nederlands, Engels, wiskunde en biologie relatief laag is. Gezien de grootte van de standaardfouten hangt eerste correctie sterker samen met geannoteerde dan met blanco derde correctie (.59 versus .47).

Tabel 46

Overeenstemming over de scores per opgave geschiedenis

	Aantal scores	Percentage exacte overeenstemming	Cohen's Kappa	Standaardfout
Eerste correctie en derde correctie geannoteerd	2250	70	.59	.013
Eerste correctie en derde correctie blanco	2250	61	.47	.014
Derde correctie geannoteerd en blanco	2250	68	.55	.013

Overeenstemming tussen drie correctieprocedures per type opgave geschiedenis

In de hiervoor besproken analyse is geen onderscheid gemaakt naar type opgave. Het is goed denkbaar dat de samenhang tussen de correctieprocedures voor het ene opgaventype anders is dan voor het andere opgaventype. Zo zou de samenhang tussen de scores van de drie correctieprocedures bij objectief scorebare meerkeuzeopgaven hoger kunnen zijn dan bij open opgaven. Vandaar dat de mate waarin de drie correcties tot dezelfde itemscores leiden ook per opgaventype is gerapporteerd (zie Tabel 47). Het examen bestaat uit één combinatievraag, twee ordeningsvragen, drie kort-antwoordvragen, vijf kort-lang-antwoordvragen en negentien lang-antwoordvragen. De resultaten vatten we als volgt samen:

- De percentages exacte overeenstemming en de voor toevallige samenhang gecorrigeerde overeenstemming (Kappa) zijn het hoogst voor de ordeningsvraag. Van de overige opgaventypen is de overeenstemming duidelijk lager waarbij het ene opgaventype niet veel voor het andere onder doet.
- Bij alle vraagtypen is het percentage exacte overeenstemming tussen eerste en geannoteerde derde correctie hoger dan tussen eerste en blanco derde correctie.
- Alleen bij de lang-antwoordvragen is de voor toevallige overeenstemming gecorrigeerde samenhang tussen eerste en derde correctie hoger voor geannoteerde dan voor blanco derde correcties (.56 versus .43). Bij de overige vraagtypen is het aantal vragen en correcties aanzienlijk kleiner en overlappen de 95%-betrouwbaarheidsintervallen van *Kappa* elkaar.

Tabel 47

Samenhang tussen drie correctieprocedures per opgaventype geschiedenis

	Aantal Scores	Percentage exacte overeenstemming	Cohen's Kappa	Standaardfout
Combinatievraag				
Eerste correctie en derde correctie geannoteerd	75	79	.64	.073
Eerste correctie en derde correctie blanco	75	64	.43	.081
Derde correctie geannoteerd en blanco	75	71	.54	.079
Orderingsvragen				
Eerste correctie en derde correctie geannoteerd	150	93	.89	.031
Eerste correctie en derde correctie blanco	150	88	.82	.040
Derde correctie geannoteerd en blanco	150	91	.86	.036
Kort-antwoordvragen				
Eerste correctie en derde correctie geannoteerd	225	72	.64	.038
Eerste correctie en derde correctie blanco	225	63	.51	.040
Derde correctie geannoteerd en blanco	225	69	.57	.041
Kort-lang-antwoordvragen				
Eerste correctie en derde correctie geannoteerd	375	66	.54	.032
Eerste correctie en derde correctie blanco	375	58	.43	.033
Derde correctie geannoteerd en blanco	375	65	.51	.033
Lang-antwoordvragen				
Eerste correctie en derde correctie geannoteerd	1425	68	.56	.017
Eerste correctie en derde correctie blanco	1425	59	.43	.017
Derde correctie geannoteerd en blanco	1425	66	.51	.017

Verskil in correctietijd geschiedenis

De gemiddelde correctietijd per corrector per versie is weergegeven in Tabel 48. Onder meer is te zien dat corrector 1 voor het nakijken van een geannoteerd examenwerk gemiddeld 12 minuten nodig had en voor een blanco werk gemiddeld 15 minuten. De vijf correctoren hadden per geannoteerde examenwerk gemiddeld 18.49 minuten nodig en per blanco werken gemiddeld 21.54 minuten. Met behulp van multivariate variantie-analyse (MVA) is nagegaan of dit verschil in correctietijd tussen geannoteerd en blanco examenwerk statistisch significant is en dat bleek het geval ($p < .001$). Een geannoteerd werk kost gemiddeld dus drie minuten en drie seconden minder correctietijd dan een in alle andere opzichten vergelijkbaar blanco examenwerk.

Tabel 48

Correctietijd per corrector voor geannoteerde en blanco examenwerken geschiedenis

Corrector	Geannoteerde examenwerken		Blanco examenwerken	
	Gemiddelde	Standaard deviatie	Gemiddelde	Standaard deviatie
1	12.33	3.09	15.47	4.55
2	21.00	4.31	20.67	4.95
3	19.00	5.73	26.33	8.34
4	21.87	3.44	27.53	5.00
5	18.27	8.40	17.67	5.35
Gemiddelde	18.49	6.20	21.53	7.38

Tabel 48 doet vermoeden dat de correctoren verschillen in de gemiddelde correctietijd en dat het verschil tussen de correctietijd van geannoteerd en blanco werk voor de ene corrector groter is dan voor de andere. Met behulp van een tweeweg variantie-analyse is nagegaan of het verschil in correctietijd tussen correctoren significant is en of het verschil in de tijdsbesteding aan geannoteerd en blanco werk voor de ene corrector groter of kleiner is dan voor de andere. Het hoofdeffect van corrector en de interactie tussen corrector en versie blijken significant (beide $p < .001$). Kennelijk zijn er niet alleen systematische verschillen tussen versies, maar ook tussen correctoren. Tegelijkertijd verschilt het verschil in de correctietijd van geannoteerde en blanco examenwerken van corrector tot corrector. Zo blijken correctoren 2 en 5 ongeveer evenveel tijd aan geannoteerde en blanco werken te besteden, terwijl met name corrector 3 en 4 geannoteerde werken naar verhouding sneller corrigeren dan blanco werken.

Vanwege de grote verschillen tussen correctoren en de interactie tussen corrector en versie kan aan het geconstateerde verschil in de tijdsbesteding aan geannoteerde en blanco werken geen grote betekenis worden gehecht.

Cijfer op het schoolexamen geschiedenis

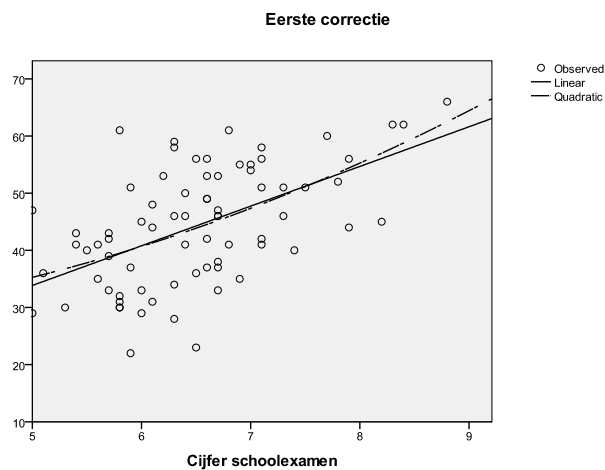
Nagegaan is of de correcties samenhangen met het cijfer voor het schoolexamen en zo ja in hoeverre deze relatie verschilt per correctieprocedure. Tabel 49 geeft informatie over de verdeling van de drie correctieprocedures en het schoolexamencijfer.

Tabel 49

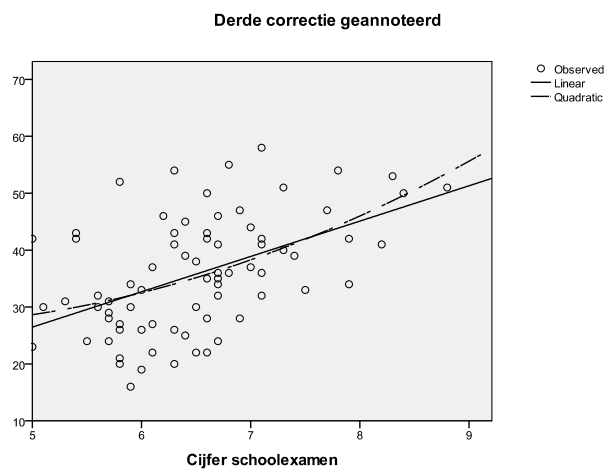
Karakterisering van de scoreverdeling voor drie correctieprocedures en het schoolexamencijfer geschiedenis

	Aantal examenwerken	Min	Max	Gemiddelde	Standaard deviatie	Scheefheid	
						Statistic	SE
Eerste correctie	75	22	66	44.31	10.33	-.01	.277
Derde correctie geannoteerd	75	16	58	35.83	10.16	.18	.277
Derde correctie blanco	75	9	56	34.89	9.87	-.04	.277
Cijfer schoolexamen	75	5.0	8.8	6.51	.81	.55	.277

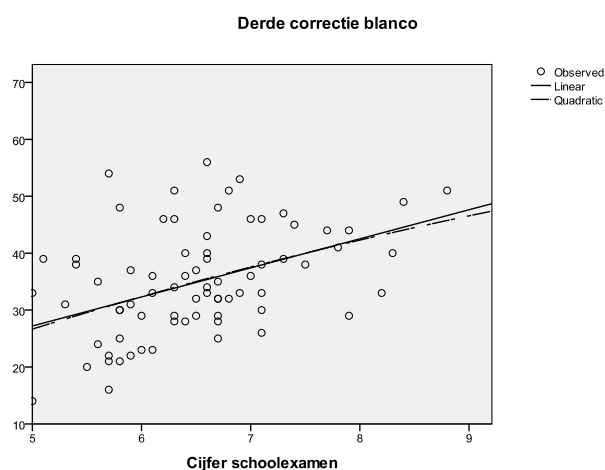
De scores voor de drie correctieprocedures en het schoolexamencijfer lijken redelijk normaal verdeeld. SE-cijfers lager dan vijf blijken niet gegeven te zijn. De spreidingsdiagrammen in Figuur 5a, 5b en 5c bevestigen dit beeld.



Figuur 5a Spreidingsdiagram eerste correctie versus cijfer schoolexamen geschiedenis



Figuur 5b Spreidingsdiagram geannoteerde derde correctie versus cijfer schoolexamen geschiedenis



Figuur 5c Spreidingsdiagram geannoteerde blanco correctie versus cijfer schoolexamen geschiedenis

De resultaten van de multiple regressieanalyse zijn weergegeven in Tabel 50. Weergegeven zijn de regressiegewichten met bijbehorende standaardfouten van model 1 met alleen een lineaire term voor het schoolexamencijfer en van model 2 met zowel een lineaire als een kwadratische term. De tabel

geeft ook R, de multiple correlatie, die in het geval van model 1 gelijk is aan de product-moment correlatie. De lineaire samenhang tussen de prestaties op het centraal examen en het schoolexamen is in alle drie gevallen relatief zwak (een toename van het schoolexamencijfer met één punt gaat gepaard met een toename van de score op het centraal schriftelijk examen met respectievelijk 6.9 (eerste correctie), 6.2 (derde correctie geannoteerd) en 5.1 scorepunten (derde correctie blanco). Er zijn geen aanwijzingen voor het veronderstelde kromlijinig verband tussen de correcties en het schoolexamencijfer. In model 2 zijn zowel de lineaire als de kwadratische term niet significant, terwijl de multiple correlatie telkens gelijk is aan die van model 1.

Tabel 50

Uitkomsten van de analyse van de regressie van de totaalscore op het schoolexamencijfer geschiedenis

		Ongestandaardiseerde regressiegewichten		Gestandaardiseerde regressiegewichten	T	Sig.	R
		b	se	Beta			
Eerste correctie							
Model 1	Lineair	6.94	1.25	.55	5.57	<.001	.55
Model 2	Lineair	-1.48	15.37	-.12	-.10	.924	.55
	Kwadratisch	.63	1.14	.66	.55	.584	
Derde correctie geannoteerd							
Model 1	Lineair	6.21	1.27	.50	4.89	<.001	.50
Model 2	Lineair	-6.73	15.63	-.54	-.43	.668	.50
	Kwadratisch	.96	1.16	1.04	.83	.409	
Derde correctie blanco							
Model 1	Lineair	5.11	1.29	.42	3.96	<.001	.42
Model 2	Lineair	8.32	15.94	.69	.52	.603	.42
	Kwadratisch	-.24	1.18	-.27	-.20	.840	

3.6 Tehatex HAVO

Het examen tehatex HAVO bevat 36 opgaven met een maximumscore van 75. Het examen bestaat uit 14 kort-antwoordvragen en 22 lang-antwoordvragen. Van de honderd examenwerken waren er vijf niet van annotaties voorzien. Deze zijn niet gebruikt voor de vergelijking van de correctieprocedures. Van de 95 resterende examenwerken was er een drie onbruikbaar voor analyse omdat het desbetreffende scoreformulier onvolledig was ingevuld.

Tabel 51 toont enkele beschrijvende kenmerken van vier correctieprocedures, te weten eerste correctie, gezamenlijke eerste en tweede correctie, derde correctie van geannoteerde examenwerk en derde correctie van blanco examenwerken. De vier correctieprocedures zijn vergeleken in termen van de totaalscore op het examen, het gemiddeld verschil tussen de totaalscores, de gemiddelde absolute afwijking (d.w.z. het aantal scorepunten dat de correctieprocedures gemiddeld van elkaar verschillen waarbij het er niet toe doet of het verschil positief of negatief is), het cijfer voor het eerste tijdvak en het percentage onvoldoendes waartoe de correctieprocedure geleid zou hebben. De spreidingsdiagrammen zijn weergegeven in Bijlage 6.

Tabel 51

Enkele beschrijvende gegevens voor de vier correctieprocedures tehatex

	Aantal examenwerken	Minimum	Maximum	Gemiddelde/ percentage	Standaard deviatie
Totaalscore					
Eerste correctie	89	25	61	41.94	8.55
Eerste en tweede correctie gezamenlijk	84	25	61	42.26	8.12
Derde correctie geannoteerd	94	19	56	36.16	7.69
Derde correctie blanco	94	19	55	34.50	7.70
Verskil tussen de totaalscores					
Eerste correctie minus gezamenlijke score	84	-9	4	-.25	2.28
Eerste correctie minus derde correctie geannoteerd	89	-6	26	6.06	6.21
Eerste correctie minus derde correctie blanco	89	-9	26	7.80	6.98
Derde correctie geannoteerd minus blanco	94	-15	17	1.66	7.17
Gemiddelde absolute afwijking					
Eerste correctie versus gezamenlijke score	84	0	9	1.37	1.83
Eerste correctie versus derde correctie geannoteerd	89	0	26	6.80	5.37
Eerste correctie minus derde correctie blanco	89	0	26	8.67	5.84
Derde correctie geannoteerd versus blanco	94	0	17	6.04	4.17
Cijfer eerste tijdvak					
Eerste correctie	89	4.30	8.60	6.33	1.02
Gezamenlijke correctie	84	4.30	8.60	6.37	.98
Derde correctie geannoteerd	94	3.60	8.00	5.64	.92
Derde correctie blanco	94	3.60	7.90	5.44	.92
Percentage onvoldoendes					
Eerste correctie	89			22%	
Gezamenlijke correctie	84			21%	
Derde correctie geannoteerd	94			47%	
Derde correctie blanco	94			61%	

Verschillen tussen de vier correctieprocedures in gemiddelde tehatex

Kijken we eerst naar de totaalscore op het examen tehatex, dan zien we dat de examenkandidaten volgens de eerste correctoren gemiddeld bijna 42 punten hebben behaald (van maximaal 75 punten). De verschillen tussen de gemiddelden van de vier correctieprocedures zijn getoetst met behulp van een multivariate variantieanalyse. De belangrijkste resultaten vatten we als volgt samen:

- Het verschil tussen de gemiddelden van geannoteerde en blanco derde correctie is significant ($p = .045$). Geannoteerde derde correctie leidt gemiddeld tot een 1.7 punten hogere score dan blanco derde correctie (oftewel .20 cijferpunt).
- Het gemiddelde van eerste correctie verschilt significant van dat van geannoteerde derde correctie ($p < .001$) en blanco derde correctie ($p < .001$). Eerste correctie leidt in vergelijking met

geannoteerde en blanco derde correctie tot een respectievelijk 6.1 en 7.8 punten hogere score (oftewel .69 en .89 cijferpunt).

- Het verschil tussen de gemiddelden van de eerste correctie en de gezamenlijk tussen eerste en tweede corrector overeengekomen score is niet significant ($p = .317$). De gezamenlijk door de eerste en tweede corrector toegekende score is nagenoeg identiek aan de score van de eerste corrector: de gezamenlijke score is gemiddeld slechts een kwart punt lager dan die van de eerste correctie (oftewel .04 cijferpunt).

Verschillen tussen vier correctieprocedures in spreiding tehatex

De gemiddelde absolute afwijking (GAA) voor de vier correctieprocedures zijn weergegeven in het middelste gedeelte van Tabel 51. Een maat die laat zien hoe sterk correctoren of correctieprocedures van elkaar afwijken is de gemiddelde absolute afwijking. Deze spreidingsmaat is in ons geval het gemiddelde van de absolute waarde van het verschil tussen de scores van de beide correctieprocedures (dat wil zeggen dat indien het verschil negatief is het minteken is weggelaten). De resultaten vatten we als volgt samen:

- De scores van geannoteerde en blanco derde correctie liggen gemiddeld zes punten uit elkaar (negatief dan wel positief). Dit komt overeen met 8% van de maximumscore van 75 (oftewel .72 cijferpunt).
- De scores van de eerste correctie wijken gemiddeld 6.8 punten af van geannoteerde derde correctie en gemiddeld 8.7 punten van blanco derde correctie. Dit komt overeen met met respectievelijk 9% en 11% van de maximumscore (oftewel .82 en 1.04 cijferpunt).
- Voor eerste correctie en gezamenlijke eerste en tweede correctie bedraagt de gemiddelde absolute afwijking slechts 1.4 scorepunten (oftewel .16 cijferpunt).

Tabel 52 laat zie hoe de GAA's verdeeld zijn (waarbij waarden van vijf en hoger zijn samengenomen). Kijken we bijvoorbeeld naar de kolom voor absolute afwijkingen van nul, dan blijkt dat de totaalscores van de eerste en de gezamenlijke correctie bij 50% van de nagekeken examenwerken volledig identiek zijn. Daarentegen bedraagt het percentage exacte overeenstemming tussen eerste en derde corrector slechts 4% (geannoteerd) en 2% (blanco). De kolom met de GAA's van vijf of hoger vertoont het omgekeerd beeld: voor eerste en gezamenlijke correctie heeft 6% van de GAA's een waarde van vijf of hoger, voor eerste versus derde geannoteerd is het 60% en voor eerste versus blanco derde correctie gaat het om 72%. Het patroon van uitkomsten doet vermoeden dat de eerste corrector het doorgaans sterker eens is met de tweede dan met de derde corrector en dat het verschil tussen eerste correctie en geannoteerde derde correctie kleiner is dan dat tussen eerste correctie en blanco derde correctie.

Tabel 52

Verdeling van de gemiddelde absolute afwijking voor vier correctieprocedures tehatex

		0	1	2	3	4	5 of hoger	Totaal
Eerste correctie minus gezamenlijke score	Aantal	42	12	8	12	5	5	84
	Percentage	50%	14%	10%	14%	6%	6%	100%
Eerste correctie minus derde correctie geannoteerd	Aantal	4	9	11	8	4	53	89
	Percentage	4%	10%	12%	9%	4%	60%	100%
Eerste correctie minus derde correctie blanco	Aantal	2	10	4	4	5	64	89
	Percentage	2%	11%	4%	4%	6%	72%	100%
Derde correctie geannoteerd minus blanco	Aantal	6	9	7	9	7	56	94
	Percentage	6%	10%	7%	10%	7%	60%	100%

Cijfers en percentage onvoldoendes centraal examen tehatex

In het voorgaande is vastgesteld dat de vier correctieprocedures niet alle tot dezelfde totaalscores op het centraal examen leiden. Het onderste gedeelte van Tabel 51 laat zien tot welke cijfers en percentages onvoldoende de vier correctieprocedures geleid zouden hebben. In de omzetting van scores naar cijfers is de normeringsterm 1.3 van het eerste tijdvak gebruikt. Een invloed van de tweede corrector op de examenscores lijkt klein: het percentage onvoldoendes daalt van 22% naar 21%. Bij geannoteerde en blanco correctie zou respectievelijk 47% en 61% van de kandidaten een onvoldoende hebben behaald.

Overeenstemming tussen de vier correctieprocedures: totaalscore tehatex

De mate waarin de vier correctieprocedures tot vergelijkbare totaalscores op het examen leiden, is ook bepaald aan de hand van de intraclasscorrelatie ICC (zie Tabel 53):

- Geannoteerde en blanco derde correctie lijken tot verschillende examenscores te leiden (voor één correctieprocedure is de ICC .56 en voor beide samen .71). Aangezien de hier niet gerapporteerde 95%-betrouwbaarheidsintervallen elkaar overlappen, moet het verschil in statistisch opzicht als verwaarloosbaar worden beschouwd.
- Eerst correctie hangt niet sterker samen met geannoteerde derde correctie dan met blanco derde correctie (.54 versus .42).
- De samenhang tussen eerste en gezamenlijke correctie lijkt hoger dan die tussen eerste en geannoteerde derde correctie (.96 versus .54) en die tussen eerste en blanco derde correctie (.96 versus .42). Gezien de hier niet gerapporteerde standaardfouten is het verschil echter verwaarloosbaar.
- De eerste en de gezamenlijk overeengekomen correcties hangen nagenoeg perfect samen (voor één correctieprocedure .96 en voor beide .98). Kennelijk zijn deze twee correctieprocedures in statistisch opzicht niet of nauwelijks van elkaar te onderscheiden en voegt de een zeer weinig toe aan de ander.

Tabel 53

Overeenstemming over de totaalscores (ICC) tehatex

Overeenstemming totaalscores	Aantal examenwerken	Correlatie	Intraclasscorrelatie	
			Individueel	Gezamenlijk
Eerste correctie en gezamenlijke correctie	84	.97	.96	.98
Eerste correctie en derde correctie geannoteerd	89	.70	.54	.70
Eerste correctie en derde correctie blanco	89	.62	.42	.59
Derde correctie geannoteerd en blanco	94	.67	.56	.71

Overeenstemming tussen de vier correctieprocedures: scores per opgave tehatex

De mate waarin de onderscheiden correctieprocedures tot vergelijkbare itemscores leiden, is bepaald aan de hand van *Cohen's Kappa* (zie Tabel 54). Omdat de scholen de scores van de tweede corrector en de gezamenlijk overeengekomen score vrijwel nooit per opgave registreren, kan de eerste correctie niet met deze beide correcties in verband worden gebracht en ontbreken deze gegevens in de tabel. Te zien is dat de overeenstemming lager is dan het geval is bij de vakken Nederlands, Engels, wiskunde en biologie. Gezien de grootte van de standaardfouten hangt eerste correctie hoger samen met geannoteerde dan met blanco derde correctie (.62 versus .43).

Tabel 54
Overeenstemming over de scores per opgave

	Aantal scores	Percentage exacte overeenstemming	Cohen's Kappa	Standaardfout
Eerste correctie en derde correctie geannoteerd	3202	74	.62	.011
Eerste correctie en derde correctie blanco	3204	61	.43	.013
Derde correctie geannoteerd en blanco	3382	65	.48	.012

Overeenstemming tussen drie correctieprocedures per type opgave tehatex

In de hiervoor besproken analyse is geen onderscheid gemaakt naar type opgave. Het is goed denkbaar dat de samenhang tussen de correctieprocedures voor het ene opgavetype anders is dan voor het andere opgavetype. Zo zou de samenhang tussen de drie correctieprocedures bij objectief scoorbare gesloten opgaven hoger kunnen zijn dan bij open opgaven. Vandaar dat de mate waarin de drie correcties tot dezelfde itemscores leiden ook per opgavetype is gerapporteerd (zie Tabel 55). De belangrijkste bevindingen zijn als volgt:

- De percentages exacte overeenstemming en de voor toevallige samenhang gecorrigeerde overeenstemming (*Kappa*) zijn voor de kort-antwoordvragen niet duidelijk anders dan voor de lang-antwoordvragen.
- Bij de kort-antwoordvragen is de overeenstemming tussen eerste en derde correcties hoger voor geannoteerde dan voor blanco derde correcties (.60 versus .41).
- Bij de lang-antwoordvragen is de overeenstemming tussen eerste en derde correcties hoger voor geannoteerde dan voor blanco derde correcties (.62 versus .45).

Tabel 55
Samenhang tussen drie correctieprocedures per opgavetype tehatex

	Aantal scores	Percentage exacte overeenstemming	Cohen's Kappa	Standaardfout
Kort-antwoordvragen				
Eerste correctie en derde correctie geannoteerd	1245	73	.60	.018
Eerste correctie en derde correctie blanco	1246	59	.41	.020
Derde correctie geannoteerd en blanco	1315	65	.48	.019
Lang-antwoordvragen				
Eerste correctie en derde correctie geannoteerd	1957	75	.62	.015
Eerste correctie en derde correctie blanco	1958	63	.45	.016
Derde correctie geannoteerd en blanco	2067	66	.48	.016

Verschil in correctietijd tehatex

De gemiddelde correctietijd per corrector per versie is weergegeven in Tabel 56. De standaarddeviatie van de tijdsbesteding van corrector 1 is nul. De verklaring is dat deze corrector alleen de gemiddelde tijdsbesteding heeft geregistreerd (in plaats van per examenwerk). Te zien is onder meer dat corrector

4 voor het nakijken van een geannoteerd examenwerk gemiddeld bijna 11 minuten nodig had en voor een blanco werk gemiddeld ruim 15 minuten. De vijf correctoren hadden voor de geannoteerde werken gemiddeld 29.76 minuten nodig en voor de blanco examenwerken 34.41 minuten. Met behulp van multivariate variantie-analyse (MVA) is nagegaan of dit tijdsverschil significant is en dat bleek het geval ($p < .001$). Het verschil blijkt omgerekend vier minuten en 39 seconden te zijn.

Tabel 56

Correctietijd per corrector voor geannoteerde en blanco examenwerken tehatex

Corrector	Geannoteerde examenwerken		Blanco examenwerken	
	Gemiddelde	Standaard deviatie	Gemiddelde	Standaard deviatie
1	60.00	.00	60.00	.00
2	31.63	5.27	40.00	6.27
3	13.33	3.43	19.21	5.73
4	10.95	1.90	15.22	3.61
5	32.00	9.12	36.63	10.48
Gemiddelde	29.76	18.37	34.41	17.24

Corrector 1 blijkt alleen de gemiddelde tijdsbesteding te hebben geregistreerd in plaats van de tijdsbesteding gedetailleerd per werk te noteren. De reden is dat deze corrector de set examenwerken horizontaal in plaats van verticaal heeft nagekeken (dit wil zeggen voor alle kandidaten eerst opgave 1, dan opgave 2 et cetera in plaats van eerst alle opgaven van kandidaat 1, dan alle opgaven van kandidaat 2 enzovoorts). Met behulp van een tweeweg variantie-analyse is nagegaan of het verschil in correctietijd tussen geannoteerd en blanco examenwerk significant is en of dit verschil voor de ene corrector groter of kleiner is dan voor de andere. Het hoofdeffect van corrector blijkt significant ($p < .001$), evenals de interactie tussen corrector en versie ($p < .001$). Kennelijk zijn er systematische verschillen in correctietijd tussen de beide versies en tussen correctoren. Daarnaast is het verschil in correctietijd tussen geannoteerde en blanco examenwerken voor de ene corrector groter dan voor de andere.

Vanwege de grote verschillen tussen correctoren en de interactie tussen corrector en versie is de zeggingskracht van het geconstateerde verschil in tijdsbesteding aan geannoteerde en blanco werken gering. Gezien het grote verschil met de overige correctoren heeft corrector 1 de tijdsbesteding waarschijnlijk ernstig overschat. De gerapporteerde gemiddelde correctietijd per examenwerk tehatex moet daarom met extra terughoudendheid geïnterpreteerd worden.

Cijfer op het schoolexamen tehatex

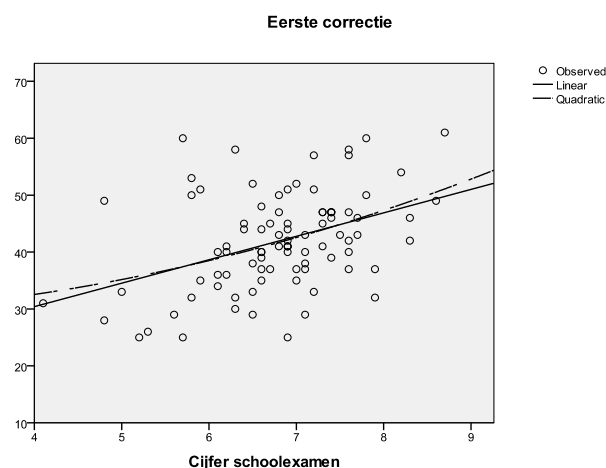
Nagegaan is of de correcties samenhangen met het cijfer voor het schoolexamen en zo ja in hoeverre deze relatie verschilt per correctieprocedure. Tabel 57 geeft informatie over de verdeling van de drie correctieprocedures en het schoolexamencijfer.

Tabel 57

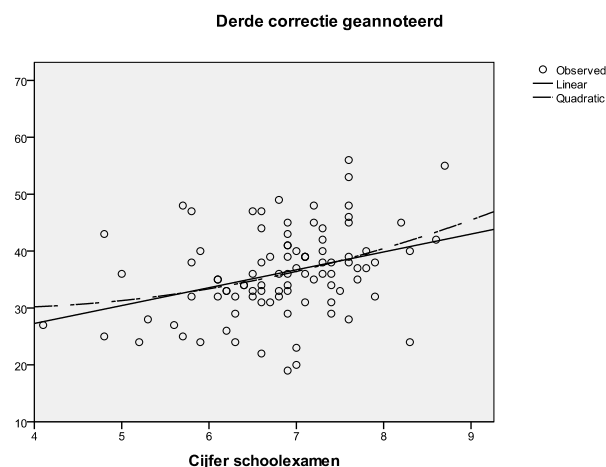
Karakterisering van de scoreverdeling voor drie correctieprocedures en het schoolexamencijfer tehatex

	Aantal examen- werken	Min	Max	Gemiddelde	Standaard deviatie	Scheefheid	
						Statistic	SE
Eerste correctie	89	25	61	41.94	8.55	.11	.26
Derde correctie geannoteerd	94	19	56	36.16	7.69	.15	.25
Derde correctie blanco	94	19	55	34.50	7.70	.47	.25
Cijfer schoolexamen	94	4.1	8.7	6.83	.84	-.57	.25

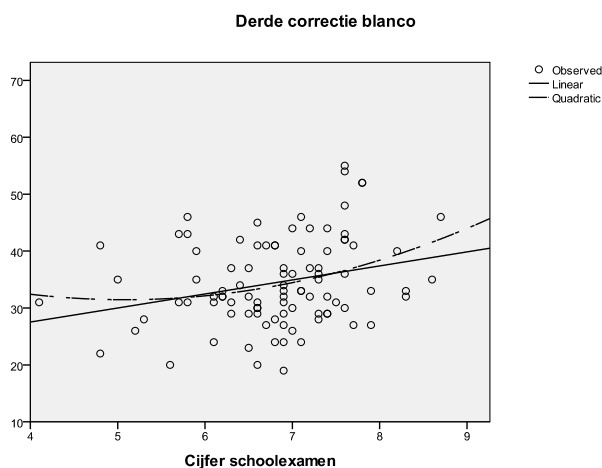
De scores voor de drie correctieprocedures lijken redelijk normaal verdeeld. De verdeling van de schoolexamencijfers is negatief scheef: lage cijfers en cijfers rond de grens tussen voldoende en onvoldoende worden minder gegeven dan hogere cijfers. De spreidingsdiagrammen in Figuur 6a, 6b en 6c bevestigen dit beeld.



Figuur 6a Spreidingsdiagram eerste correctie versus cijfer schoolexamen tehatex



Figuur 6b Spreidingsdiagram geannoteerde derde correctie versus cijfer schoolexamen tehatex



Figuur 6c Spreidingsdiagram geannoteerde blanco correctie versus cijfer schoolexamen tehatex

De resultaten van de multiple regressieanalyse zijn weergegeven in Tabel 58. Weergegeven zijn de regressiegewichten met bijbehorende standaardfouten van model 1 met alleen een lineaire term voor het schoolexamencijfer en van model 2 met zowel een lineaire als een kwadratische term. De tabel geeft ook R, de multiple correlatie, die in het geval van model 1 gelijk is aan de product-moment correlatie. De lineaire samenhang tussen de prestaties op het centraal examen en het schoolexamen is in alle drie gevallen zwak (een toename van het schoolexamencijfer met één punt gaat gepaard met een toename van de score op het centraal schriftelijk examen met respectievelijk 4.1 (eerste correctie), 3.1 (derde correctie geannoteerd) en 2.5 scorepunten (derde correctie blanco). Er zijn geen aanwijzingen voor het veronderstelde kromlijinig verband tussen de correcties en het schoolexamencijfer. In model 2 zijn zowel de lineaire als de kwadratische term niet significant, terwijl de multiple correlatie gelijk of niet veel hoger is dan die van model 1.

Tabel 58

Uitkomsten van de analyse van de regressie van de totaalscore op het schoolexamencijfer tehatex

		Ongestandaardiseerde regressiegewichten		Gestandaardiseerde regressiegewichten	T	Sig.	R
		b	se	Beta			
Eerste correctie							
Model 1	Lineair	4.12	.98	.41	4.19	<.001	.41
Model 2	Lineair	-.62	1.01	-.06	-.06	.951	.41
	Kwadratisch	.36	.76	.47	.48	.636	
Derde correctie geannoteerd							
Model 1	Lineair	3.14	.90	.34	3.49	.001	.34
Model 2	Lineair	-3.31	9.21	-.36	-.36	.720	.35
	Kwadratisch	.49	.69	.71	.70	.483	
Derde correctie blanco							
Model 1	Lineair	2.46	.92	.27	2.67	.009	.27
Model 2	Lineair	-8.30	9.42	-.90	-.88	.380	.29
	Kwadratisch	.82	.71	1.18	1.15	.254	

4 Discussie

In de centrale schriftelijke examens wordt het examenwerk eerst door de eigen docent gecorrigeerd en vervolgens door een docent van een andere school. Eerste correctoren krijgen daarbij het advies om het examenwerk van annotaties te voorzien. Volgens de richtlijnen van de VO-raad (2009) dient de eerste corrector duidelijk aan te geven waar in het werk de onvolkomenheden zitten en de toegekende punten per vraag voor de linker kantlijn van het examenwerk te noteren. In dit onderzoek is nagegaan in hoeverre het verwijderen van de annotaties van de eerste corrector van invloed is op de scores van de tweede corrector. Daartoe heeft een steekproef van correctoren een steekproef van examenwerken opnieuw gecorrigeerd. Deze hernieuwde correctie is uitgevoerd door zogeheten derde correctoren. Daarbij is van elk examenwerk zowel een geannoteerde als een blanco versie nagekeken. Afgezien van het ontbreken van de annotaties van de eerste correctie was de blanco versie volledig identiek aan de geannoteerde versie. Daarnaast zijn de scores van de derde correctoren vergeleken met de scores van de eerste corrector en met de gezamenlijk tussen eerste en tweede corrector overeengekomen scores (eerste tijdvak). Voor de vakken Nederlands, Engels, wiskunde, biologie, geschiedenis en tehatex hebben dertig derde correctoren in totaal 616 examenwerken opnieuw gecorrigeerd (waarvan 562 annotaties van de eerste corrector bevatten). Per vak waren vijf correctoren in het onderzoek betrokken. Een check op de kwaliteit van de correcties toonde aan dat de scores van de vijf correctoren voor onderzoeksdoeleinden in voldoende mate overeenstemmen. In dit hoofdstuk relateren we de onderzoeksresultaten aan de onderzoeksvragen en doen we enkele aanbevelingen voor vervolgonderzoek.

Geannoteerde versus blanco derde correctie

Doel van het onderzoek was na te gaan in hoeverre het verwijderen van de annotaties van de eerste corrector van invloed is op de scores van een tweede corrector. De eerste onderzoeksvraag luidde: 'In hoeverre is er een verschil tussen de scores van geannoteerde en blanco derde correctie?' Zoals verwacht hangen de scores van geannoteerde en blanco derde correctie sterk samen. De hoogte van de samenhang is echter niet voor alle examenvakken gelijk. Bij de examenvakken Engels, wiskunde en biologie maakt het voor de rangordening van de kandidaten naar vaardigheidsniveau niet veel uit of het werk nog de annotaties van de eerste corrector bevat (de correlaties zijn respectievelijk .92, .94 en .93). Bij de vakken Nederlands, geschiedenis en tehatex zijn de verschillen in de rangordening van de kandidaten aanmerkelijk groter (de correlaties zijn respectievelijk .73, .77 en .67).

De verschillen tussen de gemiddelde examenscores van geannoteerde en blanco derde correctie zijn relatief klein. Ze variëren afhankelijk van het examenvak van een kwart tot ruim anderhalve scorepunt. Klaarblijkelijk zijn de gemiddelde scores van geannoteerde en blanco derde correctie in hoge mate vergelijkbaar. Alleen bij wiskunde en tehatex is het verschil tussen de twee gemiddelden statistisch significant. In beide gevallen is het gemiddelde van geannoteerde derde correctie hoger dan dat van blanco derde correctie. Bij wiskunde gaat het om ruim een half scorepunt en bij tehatex betreft het ruim anderhalve punt.

De hiervoor gerapporteerde gemiddelden geven aan welke invloed het verwijderen van de annotaties zou kunnen hebben op de gemiddelde prestaties van de totale groep kandidaten in het onderzoek. Voor het doel van ons onderzoek zijn deze gemiddelden minder informatief. Centrale examens hebben immers tot doel de prestaties van individuele kandidaten te meten en niet die van groepen kandidaten. De gemiddelde absolute afwijking is een meer geschikte maat omdat deze weergeeft hoe ver de scores van individuele kandidaten gemiddeld uit elkaar liggen (positief dan wel negatief). Deze spreidingsmaat is in ons geval het gemiddelde van de absolute waarde van het verschil tussen de scores van de beide correctieprocedures (dat wil zeggen dat indien het verschil negatief is het minteken is weggelaten). Gemiddeld liggen de scores van geannoteerde en blanco derde correctie afhankelijk van het examenvak tussen de twee en zes punten uit elkaar (positief dan wel negatief). Dit komt overeen met respectievelijk 4% en 8% van de maximumscore. De drie vakken met de grootste

absolute scoreverschillen zijn Nederlands, geschiedenis en tehatex met respectievelijk 7%, 7% en 8% van de maximumscore (oftewel .64, .65 en .72 cijferpunt).

De verschillen tussen geannoteerde en blanco derde correctie worden ook zichtbaar in de percentages onvoldoendes waartoe beide correctieprocedures geleid zouden hebben. Bij Nederlands, Engels en wiskunde liggen ze hooguit twee percentagepunten uit elkaar. Bij biologie, geschiedenis en tehatex is het percentage onvoldoendes voor blanco derde correctie echter respectievelijk vier, vier en veertien percentagepunten hoger dan voor geannoteerde derde correctie. Bij deze vier vakken zou het verwijderen van de punten en aantekeningen van de eerste corrector derhalve tot meer onvoldoendes kunnen leiden (bij gelijkblijvende normering).

Eerste correctie versus geannoteerde en blanco derde correctie

Hiervoor constateerden we verschillen tussen geannoteerde en blanco derde correctie in gemiddelden, spreiding, samenhang en percentage onvoldoendes. Omdat de geannoteerde en blanco werken aselekt aan derde correctoren zijn toegewezen, lijkt het verwijderen van de annotaties van de eerste corrector een aannemelijke verklaring. Helemaal zeker is dit echter niet. Elke versie van een examenwerk is immers slechts door één corrector nagekeken; het oordeel van de ene corrector over de ene versie legt niet meer gewicht in de schaal dan het oordeel van de andere corrector over de andere versie van hetzelfde examenwerk. Er ontbreekt met andere woorden een criterium waartegen de scores van de derde correctoren kan worden afgezet. Vandaar dat we de scores van geannoteerde en blanco derde correctie ook hebben vergeleken met die van de eerste correctoren. De desbetreffende onderzoeksvraag luidde: 'In hoeverre is de samenhang tussen eerste en derde correctie voor geannoteerde werken anders dan voor blanco werken?' Ervan uitgaande dat derde correctoren zich door de annotaties van de eerste corrector laten beïnvloeden verwachtten we dat eerste correctie sterker zou samenhangen met geannoteerde dan met blanco derde correctie.

Inderdaad blijkt de samenhang tussen de scores van de eerste en de derde correctoren bij alle zes examenvakken voor geannoteerde werken hoger dan voor blanco werken. Het verschil is echter niet voor alle examenvakken even groot. Wederom kunnen we de zes examenvakken in twee groepen van drie indelen. Bij Engels, wiskunde en biologie, waar de samenhang tussen eerste en derde correctie relatief hoog is, hangt eerste correctie maar weinig sterker samen met geannoteerde dan met blanco derde correctie. Bij Nederlands, geschiedenis en tehatex, waar de samenhang tussen eerste en derde correctie naar verhouding laag is, hangt eerste correctie duidelijk sterker samen met geannoteerde dan met blanco derde correctie. Kennelijk is het voor de rangordening van de kandidaten van belang of het examenwerk nog de punten en aantekeningen van de eerste corrector bevat.

Een verschil tussen geannoteerde en blanco derde correctie zien we ook terug in de spreiding van de scores (de gemiddelde absolute afwijking). Bij Engels, wiskunde en biologie, waar de scores van eerste en derde correctie relatief dicht bij elkaar liggen, wijkt eerste correctie ongeveer even sterk af van geannoteerde als van blanco derde correctie (positief dan wel negatief). Zo is bij Engels het gemiddelde absolute verschil tussen eerste correctie en geannoteerde derde correctie gemiddeld 2.2 punten en tussen eerste correctie en blanco derde correctie gemiddeld 2.4 punten, hetgeen overeenkomt met 5% en 5% van de maximumscore (ofwel .41 en .45 cijferpunt). Bij Nederlands, geschiedenis en tehatex, waar de scores van eerste en derde correctie naar verhouding sterker verschillen, wijken de eerste correcties minder sterk af van geannoteerde dan van blanco derde correcties (positief dan wel negatief). Zo liggen bij geschiedenis de scores van eerste en geannoteerde derde gemiddeld 8.6 punten uit elkaar en het overeenkomstige verschil tussen eerste correctie en blanco derde correctie bedraagt bij dit vak 10.1 punten, hetgeen overeenkomt met respectievelijk 9% en 13% van de maximumscore (ofwel 1.02 en 1.20 cijferpunt).

De bevinding dat de eerste correctie over het algemeen sterker samenhangt met geannoteerde dan met blanco derde correctie wijst erop dat derde correctoren zich laten beïnvloeden door de annotaties van de eerste corrector. Kennelijk is onafhankelijke tweede correctie het beste gegarandeerd als het examenwerk geen annotaties van de eerste corrector bevat. De derde correctoren in het onderzoek hadden de opdracht om het examenwerk zo volledig, objectief en zorgvuldig mogelijk na te kijken. Bovendien kregen zij ruim de tijd om het correctiewerk uit te voeren en ontvingen zij een vergoeding die vergelijkbaar is met hetgeen correctoren van de staatsexamens ontvangen. Hiermee kan de vraag

of het verwijderen van de annotaties van de eerste corrector nog nodig is als tweede correctoren voldoende tijd krijgen bevestigend beantwoord worden.

Het geconstateerde verschil tussen geannoteerde en blanco derde correctie lijkt op gespannen voet te staan met het uitgangspunt dat iedere leerling recht heeft op een zo onafhankelijke mogelijke correctie van zijn of haar examenwerk. Vanuit het perspectief van een rechtvaardige beoordeling valt te overwegen de huidige richtlijn voor het annoteren van examenwerk opnieuw te bezien (VO-raad, 2009). Een eventueel besluit ten faveure van blanco correctie dient te worden genomen op basis van een zorgvuldige afweging van de voor- en nadelen van blanco versus geannoteerde correctie. Daarbij dienen ook de omstandigheden waaronder de correctie plaatsvindt in ogenschouw genomen te worden.

Het onderhavige onderzoek is niet onder reële examenomstandigheden uitgevoerd (onder meer omdat de derde correctoren de examenwerken buiten de examenperiode nakijken en zoals gezegd ruim de tijd kregen en een redelijke vergoeding ontvingen). Uit onderzoek van het Platform VVVO (2008a) blijkt dat integrale tweede correctie voor docenten een hoge taakbelasting met zich meebrengt; docenten blijken de integrale tweede correctie alleen ten koste van hun eigen vrije tijd en nachtrust te kunnen uitvoeren. Als de overheid aan een integrale tweede correctie wil vasthouden, aldus het Platform VVVO, zal de voor de correctie benodigde tijd 'zichtbaar en geormerkt in de taakbelasting van de betrokken docenten opgenomen moeten worden (p. 1)'. Vandaar dat wij pleiten voor een proef met 'onafhankelijke' tweede correctie van blanco examenwerken tijdens de examencampagne. Voor de uitvoerbaarheid is het van belang dat eerste en tweede correctoren voldoende tijd krijgen om het correctiewerk integraal uit te voeren. Een mogelijke uitwerking is dat de examenwerken na gescand te zijn digitaal naar de correctoren verzonden worden. Zo krijgen alle correctoren een blanco examen onder ogen en hoeven tweede correctoren niet meer te wachten tot de 'eigen' docent klaar is. Voor de onafhankelijkheid van de correctie is het van belang dat de tweede corrector geen inzage krijgt in de schoolexamencijfers van de kandidaten. Enkele van de onderzoeksvragen die de proef met het scannen zou kunnen beantwoorden zijn:

- In hoeverre is de gewijzigde correctieprocedure uitvoerbaar (o.a. scannen)?
- Treedt de veronderstelde tijdswinst daadwerkelijk op nu de tweede corrector niet meer op de eerste corrector hoeft te wachten?
- In hoeverre lukt het om de correctieomstandigheden overeenkomstig de voorstellen van het Platform VVVO (2008) te optimaliseren (ondanks de voor velen zeer drukke examenperiode)?
- In hoeverre gaat de eerste corrector anders corrigeren in de wetenschap dat de tweede correctie integraal wordt uitgevoerd?
- In hoeverre gaan tweede corrector daadwerkelijk anders corrigeren als zij voldoende tijd en een redelijke vergoeding krijgen? Gaan tweede correctoren strenger corrigeren (bijvoorbeeld even streng als de derde correctoren in het onderhavige onderzoek)? Wordt de tweede correctie vaker integraal uitgevoerd in plaats van steekproefsgewijs?
- Hoe verloopt het overleg tussen de eerste en tweede corrector? Vindt overleg vaker dan nu plaats? Zijn er meer conflicten? Is er meer sprake van intervisie?
- In hoeverre percipiëren de direct betrokken de wijzigingen als een verbetering?

Eerste correctie versus derde correctie

Op grond van eerder onderzoek in andere examensituaties (o.a. Inspectie van het Onderwijs, 2009a,b) verwachtten we dat eerste correctoren over het algemeen hogere scores toekennen dan derde correctoren (de tweede verwachting bij de tweede onderzoeksvraag). Dit blijkt inderdaad het geval. Het verschil in de gemiddelde examenscore van eerste en derde correctie varieert afhankelijk van het examenvak en de versie van één tot ruim negen scorepunten. Telkens waren de scores van de eerste correctie gemiddeld hoger dan die van de geannoteerde en blanco derde correctie. Uitgedrukt als een percentage van de maximumscore gaat het bij Engels en wiskunde om 2% à 3%, bij Nederlands om 4% à 5%, bij biologie om 5% à 6%, bij tehatex om 8% à 10% en bij geschiedenis om 11% à 12%. Uitgedrukt in termen van cijferpunten gaat het bij Engels om .18 à .24, bij wiskunde

om .23 à .30, bij Nederlands om .38 à .48, bij biologie om .49 à .52, bij tehatex om .69 à .89 en bij geschiedenis om .99 à 1.11.

De gemiddelde absolute afwijkingen vertonen een soortgelijk beeld. Bij wiskunde liggen de scores van eerste en derde correctoren gemiddeld 4% à 5% uit elkaar, bij Engels gaat het om 5%, bij biologie om 6% à 7%, bij Nederlands om 7% à 9%, bij tehatex om 9% à 12% en bij geschiedenis om 11% à 13% (positief dan wel negatief). Uitgedrukt in termen van cijferpunten gaat het bij Engels om .41 à .45, bij wiskunde om .35 à .43, bij Nederlands om .61 à .82, bij biologie om .52 à .59, bij tehatex om .82 à 1.04 en bij geschiedenis om 1.02 à 1.20.

Het verschil in soepelheid tussen de eerste en derde correctoren wordt ook zichtbaar in de percentages onvoldoendes. Als we de blanco derde correctie uit het onderzoek als maatstaf nemen en deze vergelijken met de eerste correctie, stijgt het percentage onvoldoendes bij wiskunde van 23% tot 29%, bij Engels van 27% tot 34%, bij biologie van 15% tot 26%, bij Nederlands van 20% tot 30%, bij geschiedenis van 23% tot 55% en bij tehatex van 22% tot 61%. De geconstateerde verschillen tussen eerste en derde correctie staan op gespannen voet met het uitgangspunt dat iedere leerling recht heeft op een rechtvaardige beoordeling. Voor het cijfer op het examen zou het niet uit mogen maken wie het examenwerk nakijkt. De opzet van het onderzoek laat geen honderd procent zekere uitspraken over mogelijke oorzaken toe. Daartoe is een relatief kleinschalig vervolgonderzoek nodig waarbij een panel van intensief getrainde correctoren een beperkte steekproef uit de examenwerken onder gecontroleerde omstandigheden nogmaals beoordeelt. Hiermee worden zogeheten 'ware' scores verkregen waartegen de scores van de eerste, tweede en derde correctoren uit het onderzoek kunnen worden afgezet. Verder lijkt onderzoek geboden naar de wijze waarop docenten de eerste correctie uitvoeren. In allerlei beoordelingscontexten is namelijk aangetoond dat beoordelaars te soepel beoordelen en te weinig spreiding aanbrengen als zij degenen die zij moeten beoordelen zelf hebben opgeleid (Albanese, 2000; Guilford, 1936; Inspectie van het Onderwijs, 2009a,b; Landy & Farr, 1980; Murphy & Cleveland, 1995; Symonds, 1931; Thorndike, 1920). Vooruitlopend op de resultaten van het voorgestelde vervolgonderzoek lijkt het daarom verleidelijk het geconstateerde verschil tussen eerste en derde correctie mede toe te schrijven aan het gegeven dat de derde corrector de kandidaat niet persoonlijk kent en geen persoonlijk belang heeft bij te soepele (of te strenge) correcties.

Verschillen tussen opgaventypen

De zes examenvakken maken gebruik van verschillende typen opgaven. Zo bestaat het examen Engels vooral uit meerkeuzevragen en de examens geschiedenis en tehatex vooral uit lang-antwoordvragen. De derde onderzoeksvraag luidde: 'In hoeverre is de overeenstemming tussen de scores van verschillende correctieprocedures afhankelijk van het opgaventype? De drie correctieprocedures waarvan we de scores per opgave konden vergelijken waren eerste correctie, geannoteerde derde correctie en blanco derde correctie. De verwachting was dat de verschillen tussen de scores van verschillende correctieprocedures voor gesloten vragen kleiner zijn dan voor open vragen. Daarnaast verwachtten we dat dit verschil in overeenstemming tussen eerste en derde correctoren in het nadeel van open vragen voor geannoteerde examenwerken minder groot is dan voor blanco examenwerken. Als vergelijkingsmaatstaf hanteerden we het percentage exacte overeenstemming en *Cohen's Kappa*, een overeenstemmingsmaat die corrigeert voor toevallige samenhang. Uit het onderzoek komt naar voren dat het percentage exacte overeenstemming en de voor toevallige samenhang gecorrigeerde overeenstemming voor gesloten vragen hoger is dan voor open vragen. In overeenstemming met de verwachting zijn de verschillen tussen de scores van de drie correctieprocedures groter naarmate het opgaventype meer open is. Bij de meerkeuzevragen leiden de drie correctieprocedures tot nagenoeg dezelfde scores. Hier maakt het dus weinig uit of de opgaven door eerste of derde correctoren worden nagekeken of dat de derde correctie plaatsvindt op basis van geannoteerd dan wel blanco examenmateriaal (afgezien van sporadische slordigheidsfoutjes). Bij de lang-antwoordvragen blijken de drie correctieprocedures tot duidelijk verschillende scores te leiden. Bij dit opgaventype zijn de scoreverschillen tussen eerste en derde correctie voor geannoteerde werken minder groot dan voor blanco werken. Zo hangt de eerste correctie bij de lang-antwoordvragen Nederlands, wiskunde, biologie en tehatex duidelijk sterker

samen met geannoteerde dan met blanco derde correctie (en bij geschiedenis is dit verschil net niet significant). Bij de opgaventypen die qua openheid het midden houden tussen meerkeuzevragen en lang-antwoordvragen is het beeld minder eenduidig in de zin dat er soms wel verschillen zijn aangetoond en soms niet. Zo hangt de eerste correctie bij het vak tehatex ook voor kort-antwoordvragen sterker samen met geannoteerde dan met blanco derde correctie, maar bij de kort-antwoordvragen Nederlands is dit verschil niet significant. Kennelijk zijn het vooral de annotaties bij de lang-antwoordvragen die de scores van de derde correctoren hebben beïnvloed. Een voor de hand liggende verklaring is dat de antwoordmodellen bij open vragen de correctoren minder steun bij het nakijken bieden dan de correctievoorschriften bij gesloten vragen die vaak objectief scorebaar zijn. De ruimere mogelijkheden voor eigen interpretaties verhogen mogelijk de onzekerheid van de corrector waardoor deze eerder geneigd zal zijn zich door externe informatie te laten leiden, zoals de annotaties van de eerste corrector.

Correctietijd van geannoteerde en blanco derde correctie

Het onderzoek had ook tot doel om inzicht te verschaffen in de tijdsinvestering die gemoeid is met integrale tweede correctie en in de vraag of geannoteerde werken meer, evenveel of minder correctietijd kosten dan blanco werken. De vierde onderzoeksvraag luidde: 'In hoeverre is er een verschil in de gemiddelde correctietijd van geannoteerde en blanco examenwerken?' Uit het onderzoek komt naar voren dat de correctie de ene docent veel meer tijd kost dan de andere. Daarnaast blijkt de gemiddelde correctietijd sterk van vak tot vak te verschillen. Het nakijken van een geannoteerd werk Engels kost gemiddeld ruim 11 minuten, wiskunde ruim 12 minuten, biologie bijna 16 minuten, geschiedenis ruim 18 minuten, Nederlands ruim 23 minuten en tehatex bijna 30 minuten. Bij Nederlands en Engels besteden correctoren evenveel tijd aan geannoteerde dan aan blanco werken, maar bij de overige vakken vergen geannoteerde werken minder correctietijd dan blanco werken. Bij wiskunde, biologie en geschiedenis gaat het om iets meer dan twee minuten per werk en bij tehatex om bijna vijf minuten per werk. Deze onderzoeksresultaten lijken op gespannen voet te staan met het gegeven dat het lezen en interpreteren van de annotaties van de eerste corrector tijd kost. Deze extra tijdsinvestering zou zich echter weer terug kunnen verdienen als de annotaties de derde corrector in staat zouden stellen om sneller tot een gefundeerde beslissing over de score te komen. Een alternatieve verklaring stelt dat de annotaties de derde corrector ertoe verleiden te vroeg te kiezen voor de score of interpretatie van de eerste corrector. Annotaties zouden dan bijdragen aan het ontstaan van een tunnelvisie waarbij de eerste corrector ertoe neigt om belangrijke aanwijzingen die niet bij de score of interpretatie van de eerste corrector passen over het hoofd te zien. Nader onderzoek lijkt hier geboden, bijvoorbeeld in de vorm van cognitieve interviews volgens de hardopdenkmethode met correctoren tijdens het uitvoeren van de correctietaak. De in ons onderzoek verkregen registraties van de correctietijd per werk zijn aanmerkelijk lager dan de ramingen van het Platform VVVO (2008b): bij wiskunde gaat het om 12 versus 20 minuten, bij biologie om 16 versus 62 minuten, bij geschiedenis om 18 versus 45 minuten en bij tehatex om 30 versus 55 minuten. Naar mogelijke oorzaken kan men slechts gissen. Een mogelijke verklaring is de ruimere ervaring van de derde correctoren met het toepassen van het correctievoorschrift. Op een uitzondering na hadden allen het correctievoorschrift al eerder gebruikt tijdens de eerste en meestal ook tweede correctie van het examen schooljaar 2008-2009.

De invloed van de tweede correctie op de score voor het centraal schriftelijk examen

Menings- en interpretatieverschillen tussen eerste en tweede correctoren horen onvermijdelijk bij het correctiewerk en hebben een belangrijke functie: ze leiden tot 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, maar zou er wel voor moeten zorgen dat de eigen docent bij de correctie niet zomaar zijn gang kan gaan (Algra, 2004). De vijfde onderzoeksvraag luidde: 'In hoeverre verschilt de gezamenlijk tussen de eerste en tweede corrector overeengekomen score van die van de eerste corrector?' De verwachting was dat de gezamenlijk overeengekomen score lager zou zijn dan de score van de eerste corrector. De redenering was dat eerste correctoren eerder geneigd zijn tot het geven van te soepele dan te strenge

correcties en dat de huidige tweede correctie de neiging tot het geven van overwegend te soepele correcties bestrijdt. Anders dan verwacht blijken de gezamenlijke scores niet noemenswaard af te wijken van die van de eerste correctoren. Afhankelijk van het examenvak varieert het gemiddelde verschil in de totale responsgroep van een tiende tot een derde scorepunt. Kijken we naar de gemiddelde absolute afwijking dan liggen de scores gemiddeld een half tot bijna anderhalf scorepunt uit elkaar (hetgeen overeenkomt met .06 tot .16 cijferpunt). De Pearson product moment correlatie tussen beide scores bedraagt bij Nederlands, Engels, wiskunde, biologie, geschiedenis en tehatec respectievelijk .97, .99, 1.00, .99, .99 en .96. Kennelijk resulteren eerste en gezamenlijke correctie in een nagenoeg identieke rangordening van de examenkandidaten. De intraclasscorrelatie (ICC), die rekening houdt met verschillen in gemiddelde scores, is normaliter lager dan de 'gewone' correlatie, maar is dat in dit geval niet. 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 vakken is het verschil nul of hooguit een percentagepunt. Al met al kunnen we niet anders dan concluderen dat eerste en gezamenlijke correctie in statistisch opzicht niet of nauwelijks onderscheidbaar zijn.

Voor de ogenschijnlijk geringe invloed van de tweede correctie zijn ten minste twee verklaringen te geven. Een eerste verklaring stelt dat er geen aanleiding is om de scores van de eerste corrector in twijfel te trekken (VO-raad, 2008) en dat eerste en tweede corrector het daarom vrijwel altijd volledig met elkaar eens zijn. Een tweede verklaring is dat de eerste en/of tweede correctie tekort schieten (Inspectie van het Onderwijs, 2006, 2009a,b). De opzet van het onderzoek laat het niet toe om een onderscheid tussen beide verklaringen te maken. Daartoe is nader onderzoek nodig, bijvoorbeeld inhoudsanalyse van de documenten waarop tweede correctoren hun wijzigingssuggesties noteren of veldonderzoek naar de wijze waarop de eerste en tweede correctie tijdens de examencampagne worden uitgevoerd. Mochten vervolgonderzoek de veronderstelde tekortkomingen in de eerste en/of tweede correctie bevestigen, kan dat aanleiding zijn om de huidige procedures van eerste en tweede correctie nog eens kritisch te bezien. Iedere leerling heeft immers recht op een op een zo onafhankelijk en zorgvuldig mogelijke correctie van zijn of haar examenwerk.

De gegevens over de eerste correctie hebben we bij de scholen opgevraagd. Het gebruik van secundaire gegevens brengt met zich mee dat geen informatie is verkregen over de eventuele invloed van een gewijzigde tweede correctie op de eerste correctie. We weten bijvoorbeeld niet in hoeverre de eerste corrector zorgvuldiger of juist minder zorgvuldig gaat corrigeren in de wetenschap dat de tweede corrector integraal nakijkt en voldoende tijd heeft. Een gerandomiseerd en gecontroleerd veldonderzoek tijdens de examencampagne zou hierover meer uitsluitstel kunnen geven.

De invloed van het schoolexamencijfer op de correctie

Uit het onderzoek is naar voren gekomen dat de eerste correctoren bij alle zes vakken gemiddeld aanzienlijk hogere scores toekenden dan de 'onafhankelijke' derde correctoren uit het onderzoek. Tijdens de eerste correctie beschikken docenten over de cijfers op het schoolexamen. Gezien de ervaringen bij de vmbo-praktijkexamens (Inspectie van het Onderwijs, 2009a,b) is het niet ondenkbaar dat eerste correctoren van de centraal schriftelijke examens bewust dan wel onbewust met deze SE-cijfers rekening houden. De verwachting was dat kandidaten met onvoldoendes op het schoolexamen meer van de soepelheid van de eerste corrector zouden profiteren dan hun klasgenoten met hogere cijfers. De vijfde onderzoeksvraag luidde dan ook 'In hoeverre is de samenhang tussen de correcties en het schoolexamencijfer voor kandidaten met cijfers onder of rond de grens tussen voldoende en onvoldoende anders dan voor degenen met hogere SE-cijfers?' Deze verwachting kon niet worden bevestigd. Van het veronderstelde kromlijnjige verband tussen cijfers en correcties bleek geen sprake. Eerste correctoren blijken leerlingen onder of rond de grens tussen voldoende en voldoende niet soepeler te beoordelen dan hun klasgenoten met hogere cijfers voor het schoolexamen. Hierbij zij aangetekend dat deze resultaten alleen betrekking hebben op het eerste tijdvak (waaraan nagenoeg alle kandidaten deelnemen). Bij het tweede tijdvak, dat alleen bestemd is voor inhalers en herkansers, staat er naar verhouding meer op het spel en zou het schoolexamencijfer wellicht wel een rol kunnen spelen.

Implicaties voor de examenmakers

De evaluatie van de kwaliteit van de derde correcties (zie paragraaf 2.10) maakt aannemelijk dat de betrouwbaarheid van de scores Nederlands, Engels, wiskunde en biologie hoger is dan bij geschiedenis en tehatex het geval is. Geschiedenis en tehatex zijn tevens de vakken waarbij de scores voor de vier correctieprocedures het meest uiteenlopen. Beide vakken bevatten veel open vragen waarbij het correctievoorschrift de corrector mogelijk minder steun biedt (omdat de openheid van de voorschriften aanleiding kan geven tot verschillende interpretaties en scores). Het verdient aanbeveling de correctievoorschriften geschiedenis en tehatex op dit aspect nog eens nader te bezien. Daarbij dient te worden nagegaan of en zo ja, op welke wijze de interpretatieruimte voor de corrector beperkt kan en moet worden zonder afbreuk te doen aan de validiteit van de examinering.

Literatuur

- Albanese, M.A. (2000). Challenges in using rater judgements in medical education. *Journal of Evaluation in Clinical Practice*, 6, 3, 305-319.
- Alberts, R.V.H., & Erens, B.J.M. (2009). *Verslag van de examencampagne 2009 voortgezet onderwijs*. Arnhem: Cito.
- Algra, A. (2004). Eerste en tweede correctie examens: problemen en regels. *Schoolmanagers_VO* #6, 8-10.
- Bramley, T. & Pollitt, A. (1996). *Key Stage 3 English: Annotations Study. A report by the University of Cambridge Local Examinations Syndicate for the Qualifications and Curriculum Authority*. London: QCA.
- Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies*, 52, 1, 29-46.
- Crisp, V., & Johnson, M. (2007). The use of annotations in examination marking: Opening a window into markers' minds. *British Educational Research Journal*, 33, 6, 943-961.
- Douma, L. (2008). De tweede correctie: rekkelijken en preciezen bijna met elkaar op de vuist. *AOB-het Onderwijsblad*.
- Fearnley, A. (2005). *An investigation of targeted double marking for GCSE and GCE*. London: Qualifications and Curriculum Authority.
- Guilford, J.P. (1936). *Psychometric methods*. New York: McGraw-Hill.
- Inspectie van het Onderwijs (2006). *Onderwijsverslag 2004/2005*. Utrecht: Inspectie van het Onderwijs.
- Inspectie van het onderwijs (2009a). *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*. Utrecht: Inspectie van het onderwijs.
- Inspectie van het onderwijs (2009b). *Onderwijsverslag 2007/2008*. Utrecht: Inspectie van het Onderwijs.
- Kuhlemeier, H. (2007). *Internationaal gangbare moderatieprocedures voor het vaststellen en verbeteren van de kwaliteit van schoolexamens*. Paper gepresenteerd op de 34e Onderwijs Research Dagen, 6-8 juni 2007, Groningen.
- Kuhlemeier, H., Weeren, J. van, & Werf, M.P.C. van der (2006). Scheiding van opleiden en examineren in Nederland en omringende landen. Studie verricht in opdracht van de Onderwijsraad. Cito: Arnhem.
- Landy, F.J., & Farr, J.L. (1980). Performance rating. *Psychological Bulletin*, 87, 1, 72-107.
- Luijten, T. (1977). *Procedure bij het construeren van examenopgaven in open vragen en het beoordelen van examenwerk. Voorstel aan de CVO van de Cito-projectgroep 924 (CITO-memo 248)*. Arnhem: Cito.

Massey, A. & Foulkes, J. (1994). Audit of the 1993 KS3 Science National Test Pilot and the concept of quasi-reconciliation. *Evaluation & Research in Education*, 8, 3, 119-132.

Murphy, R.J.L. (1987). Reliability of marking in eight GCE-examinations. *British Journal of Educational Psychology*, 48, 196-200.

Murphy, R.J.L. (1979). Removing the marks from examination scripts before re-marking them: does it make any difference? *British Journal of Educational Psychology*, 49, 73-78.

McVey, P.J. (1975). The errors in marking examination scripts in electronic engineering. *International Journal of Electronic Engineering Education*, 12, 203-216.

Meadows, M.L., & Baird, J. (2005). *What is the right mark? Respecting other examiners' views in a community of practice*. Poster presented at the AEA Europe conference in Dublin, November 2005.

Newton, P. E. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, 22, 4, 405-420

Pilliner, A. E. G. (1965). Review of the marking of scripts in A level History. *British Journal of Educational Psychology*, 35, 110-111.

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>

Shaw, S. (2008). Essay marking on-screen: Implications for assessment validity. *E-Learning*, 5, 3, 256-274.

Shaw, S., & Imam, H. (2008). *On-screen essay marking reliability: towards an understanding of marker behaviour*. Paper presented at the IAEA Conference Cambridge.

Suto, W.M.I, & Greateorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, 34, 2, 213-233.

Symonds, P.M. (1931). *Diagnosing personality and conduct*. New York: Appleton-Century.

Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 1, 25-29.

Vidal Rodeiro (2007). Agreement between outcomes from different double marking models. *Research Matters*, 4, 28-34.

VO-raad (2008a). *Onderzoek naar steekproefsgewijze tweede correctie* (persbericht van 13 mei 2008). Online beschikbaar via
<http://www.vosabb.nl/werkgevers-in-onderwijs/item/artikel/13491/42/>

VO-raad (2008b). *Rapportage ledenpanel examinering*. Online beschikbaar via:
<http://www.vo-raad.nl/zoekresultaten?term=vergoedingenbeleid>

VO-raad (2009). *Protocol eerste en tweede correctie centrale examens vmbo, havo en vwo*. Online beschikbaar via www.vo-raad.nl

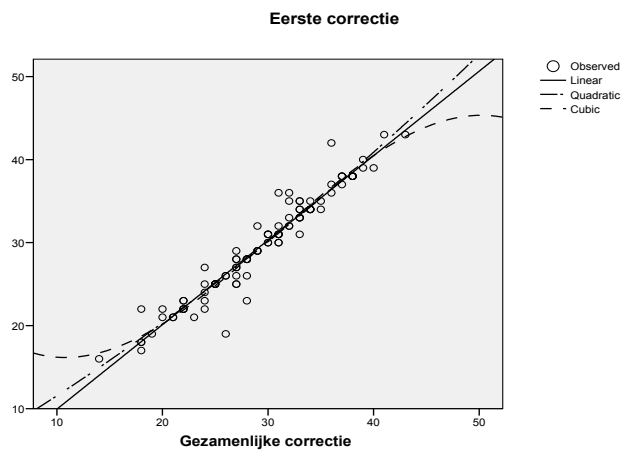
Wilmot, J. (1984). A pilot study of the effects of complete or partial removal of marks and comments from scripts before re-marking them. *AEB Research Report RAC 315*.

Wolfe, J.L. (2000). Effects of annotation on student readers and writers. *Journal of the International Conference on Digital Libraries*, San Antonio, Texas, 19-26.

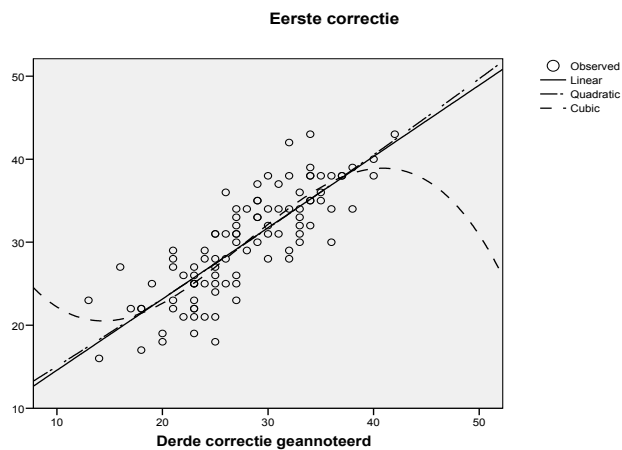
Wolfe, J.L., & Neuwirth, C.M. (2001). From the margins to the center: The future of annotation. *Journal of Business and Technical Communication*, 15, 3, 333-371.

Bijlagen

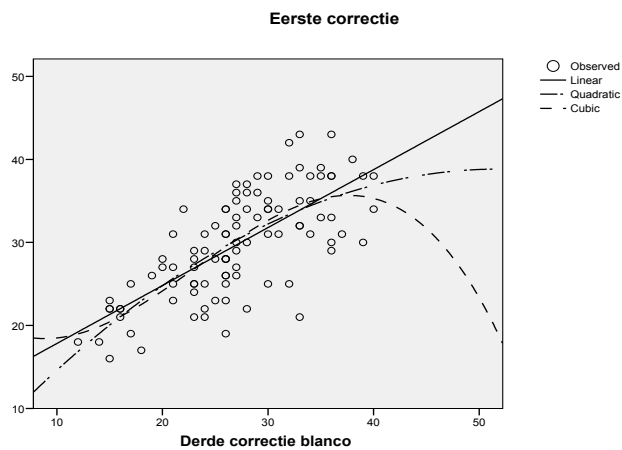
Bijlage 1 Spreidingsdiagrammen Nederlands VWO



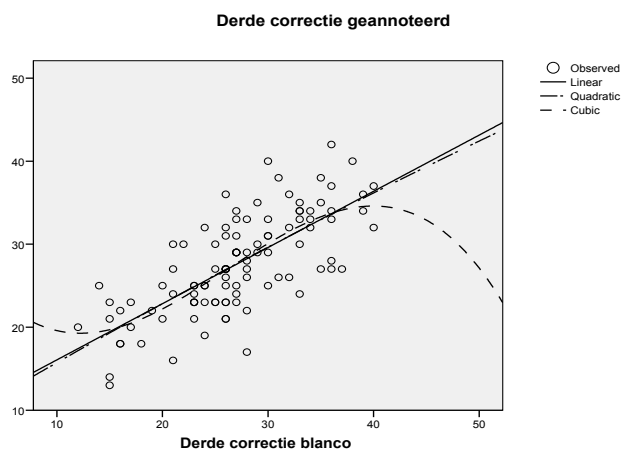
Figuur A Spreidingsdiagram eerste correctie versus eerste en tweede correctie gezamenlijk Nederlands



Figuur B Spreidingsdiagram eerste correctie versus geannoteerde derde correctie Nederlands

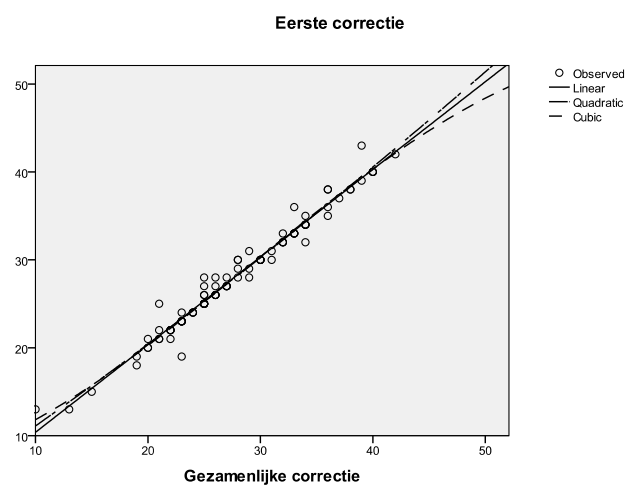


Figuur C Spreidingsdiagram eerste correctie versus blanco derde correctie Nederlands

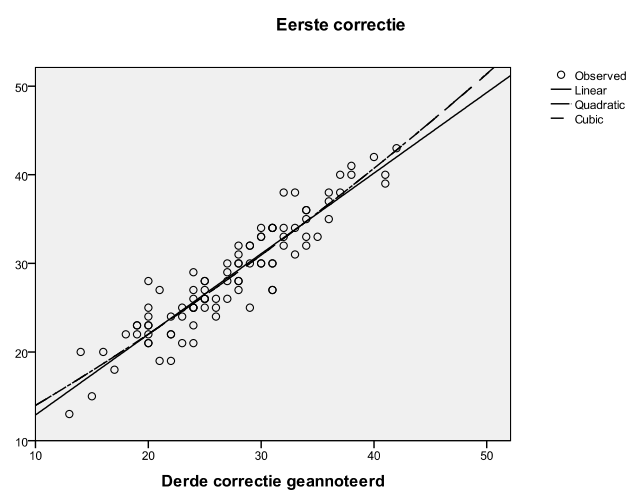


Figuur D Spreidingsdiagram geannoteerde versus blanco derde correctie Nederlands

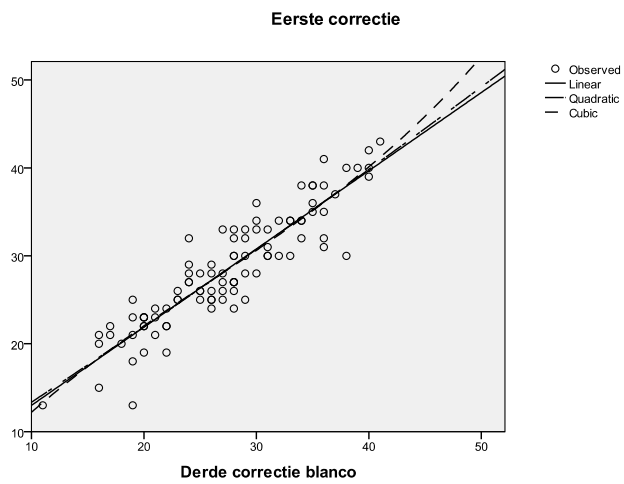
Bijlage 2 Spreidingsdiagrammen Engels GLT



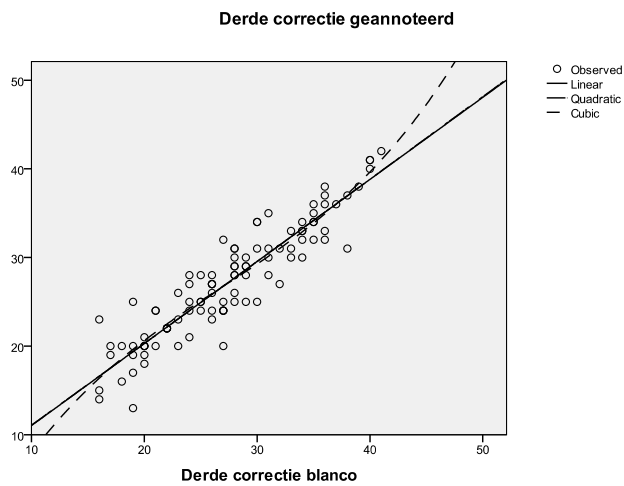
Figuur A Spreidingsdiagram eerste correctie versus eerste en tweede correctie gezamenlijk Engels



Figuur B Spreidingsdiagram eerste correctie versus geannoteerde derde correctie Engels

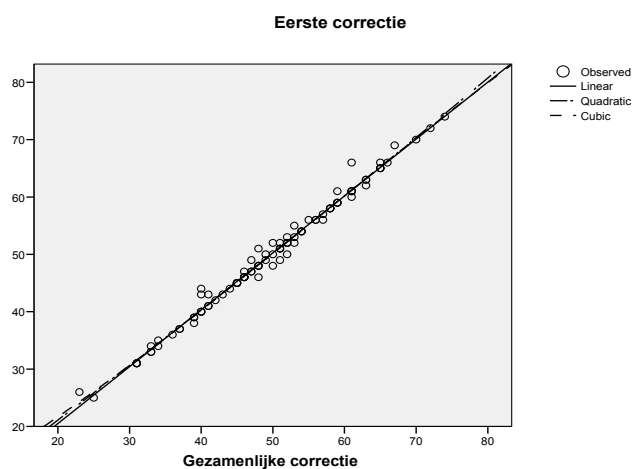


Figuur C Spreidingsdiagram eerste correctie versus blanco derde correctie Engels

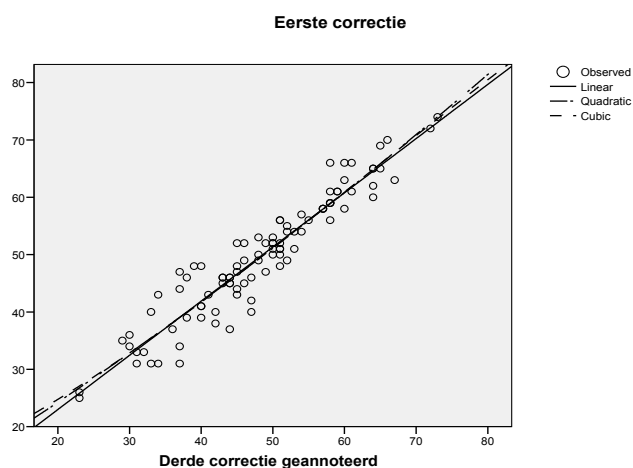


Figuur D Spreidingsdiagram geannoteerde versus blanco derde correctie Engels

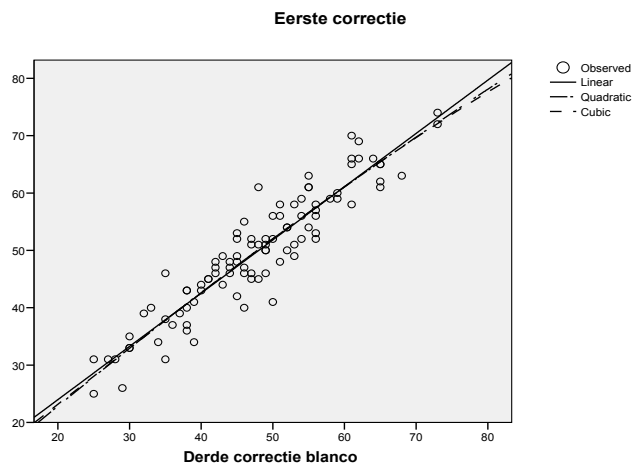
Bijlage 3 Spreidingsdiagrammen wiskunde GLTL



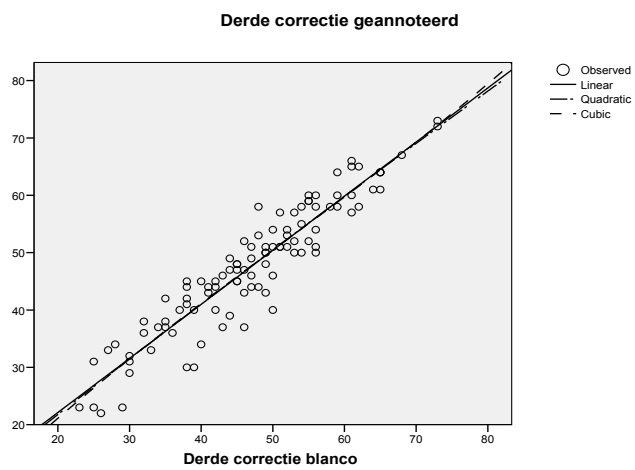
Figuur A Spreidingsdiagram eerste correctie versus eerste en tweede correctie gezamenlijk wiskunde



Figuur B Spreidingsdiagram eerste correctie versus geannoteerde derde correctie wiskunde

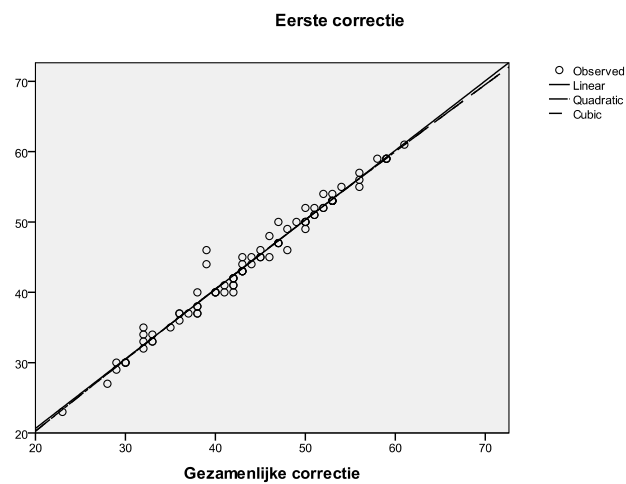


Figuur C Spreidingsdiagram eerste correctie versus blanco derde correctie wiskunde

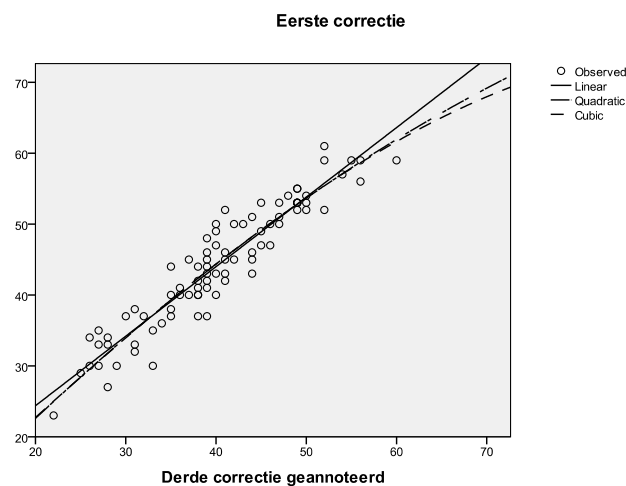


Figuur D Spreidingsdiagram geannoteerde versus blanco derde correctie wiskunde

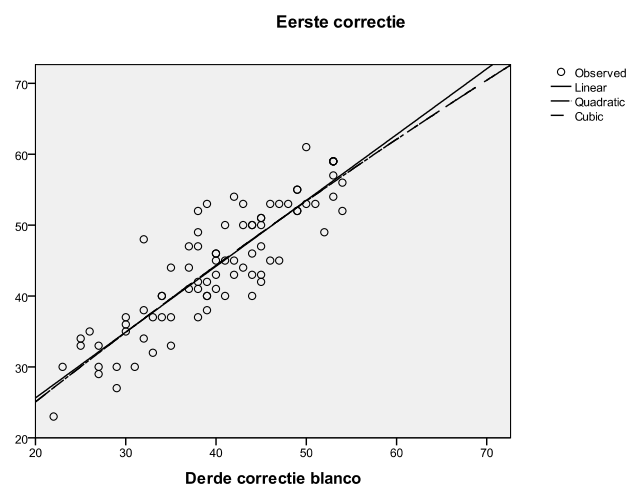
Bijlage 4 Spreidingsdiagrammen biologie VWO



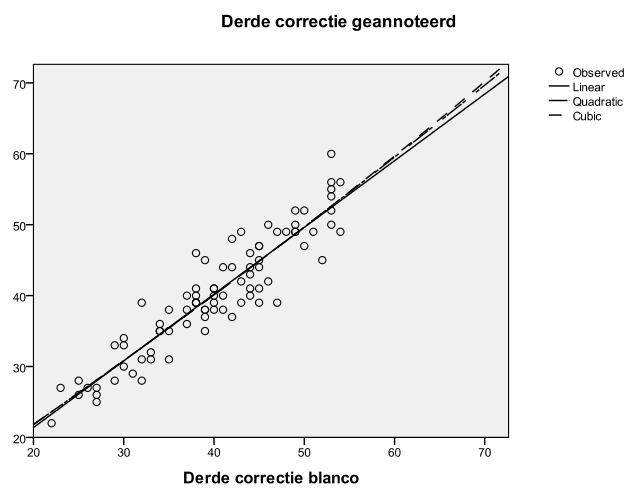
Figuur A Spreidingsdiagram eerste correctie versus eerste en tweede correctie gezamenlijk biologie



Figuur B Spreidingsdiagram eerste correctie versus geannoteerde derde correctie biologie

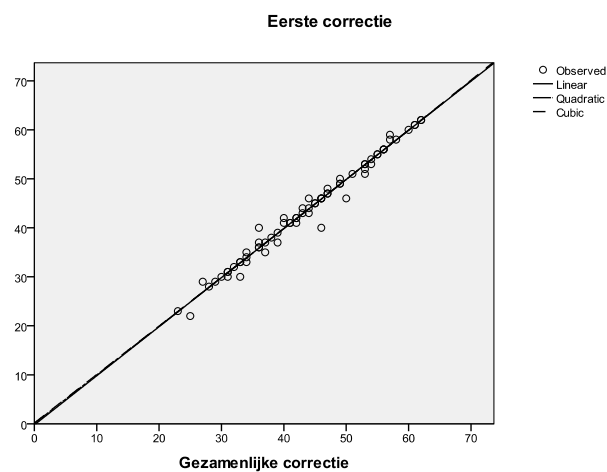


Figuur C Spreidingsdiagram eerste correctie versus blanco derde correctie biologie

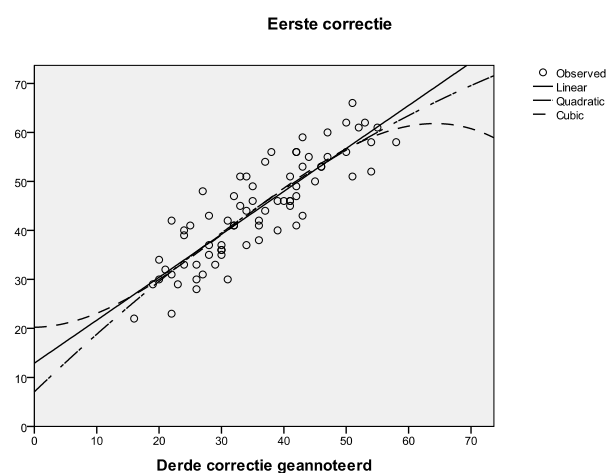


Figuur D Spreidingsdiagram geannoteerde versus blanco derde correctie biologie

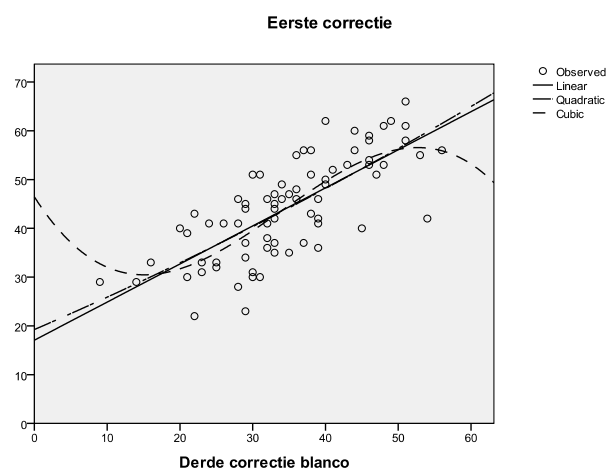
Bijlage 5 Spreidingsdiagrammen geschiedenis HAVO



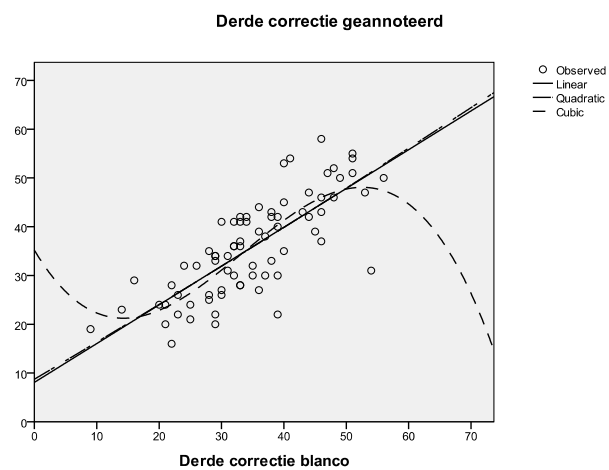
Figuur A Spreidingsdiagram eerste correctie versus eerste en tweede correctie gezamenlijk geschiedenis



Figuur B Spreidingsdiagram eerste correctie versus geannoteerde derde correctie geschiedenis

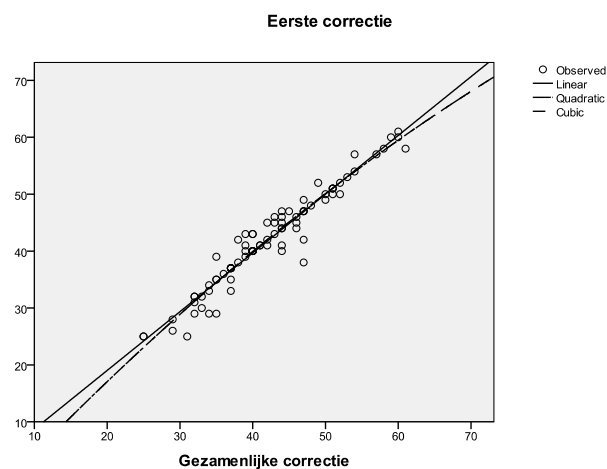


Figuur C Spreidingsdiagram eerste correctie versus blanco derde correctie geschiedenis

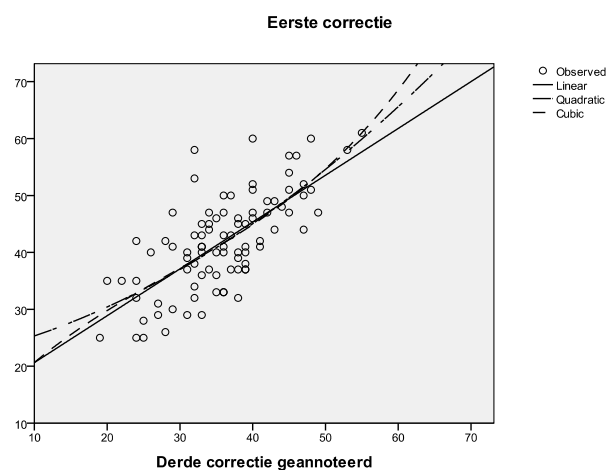


Figuur D Spreidingsdiagram geannoteerde versus blanco derde correctie geschiedenis

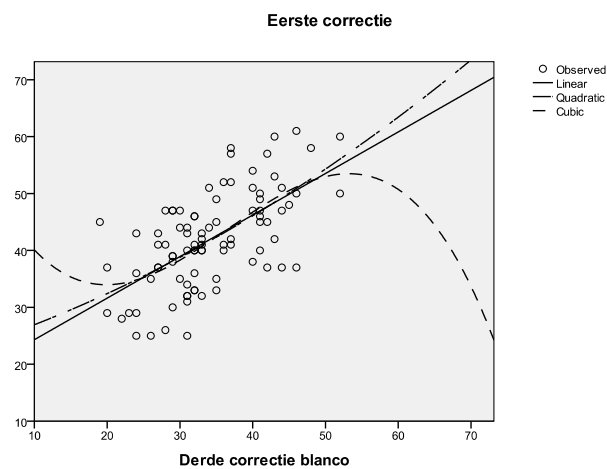
Bijlage 6 Spreidingsdiagrammen tehatex HAVO



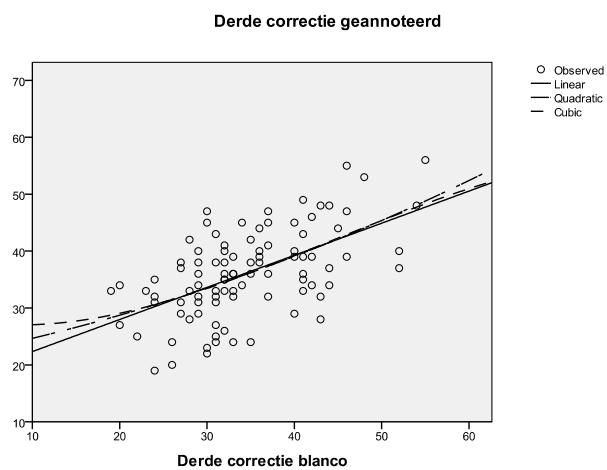
Figuur A Spreidingsdiagram eerste correctie versus eerste en tweede correctie gezamenlijk tehatex



Figuur B Spreidingsdiagram eerste correctie versus geannoteerde derde correctie tehatex



Figuur C Spreidingsdiagram eerste correctie versus blanco derde correctie tehatex



Figuur D Spreidingsdiagram geannoteerde versus blanco derde correctie tehatex

