Ineke Verheul is onderwijspsycholoog. Met haar adviesbureau Game Onderwijs Onderzoek richt zij zich op de inzet van games in het onderwijs door middel van advies, onderzoek, presentaties en workshops.
E-mail: game.ondd@gmail.com

De invloed van het LT-examenverslag op de scores

Uriël Schuurs, Hans Kuhlemeier & Hugo Gitsels

Jaarlijks publiceert het sectiebestuur Nederlands van Levende Talen een verslag van een docentenbespreking van het vwo-examen Nederlands. Dat bevat, zo staat te lezen in het meest recente verslag over het examen vwo, een synthese van de verschillende verslagen van de voorbespreking en van de bijeenkomst in Utrecht, waarbij doorgaans 50 docenten aanwezig zijn. Een overgrote meerderheid van de docenten Nederlands in de examenklassen gebruikt dit verslag tijdens de correctie van de examens. Tot dusverre ontbrak echter onderzoek naar het effect van het gebruik ervan. Heeft dit gebruik bijvoorbeeld invloed op de hoogte van de examenscores? Of wordt de betrouwbaarheid van de correctie er positief – of negatief – door beïnvloed? Dit artikel beschrijft de opzet en de resultaten van een onderzoek uit 2016 waarin we de effecten van het LT-verslag in beeld brengen op de hoogte van de scores en de betrouwbaarheid van de scores. Daarnaast is er in 2017 een test Correctievoorschrift geweest, waarbij docenten na correctie van leerlingwerk feedback kunnen leveren op het correctievoorschrift. Aan het eind van dit artikel besteden we aandacht aan deze test.

Jaarlijks wordt het centraal examen Nederlands vastgesteld in opdracht van het College voor Toetsen en Examens (CvTE). Die vaststelling is inclusief de bijbehorende correctievoorschriften. De bedoeling is dat de eerste en de tweede correctie plaatsvinden met inachtneming van de gepubliceerde correctievoorschriften: bij de eerste correctie beoordeelt een docent het werk van de eigen leerlingen, de tweede correctie is een check waarbij een docent van een andere school het leerlingwerk beoordeelt dat eerst door de eigen docent is beoordeeld. Jaarlijks houdt Levende Talen enkele dagen na afname van het examen Nederlands een vergadering, waarbij docenten de antwoorden in het gepubliceerde correctievoorschrift bespreken en daar, mede op basis van nagekeken leerlingwerk, aanvullingen en concretiseringen bij geven. Het verslag dat Levende Talen van deze jaarlijkse docentbespreking op internet publiceert (vanaf hier aangeduid als het LT-verslag), wordt door maar liefst 94% van de docenten Nederlands gebruikt bij de correctie van de examens (Cito/CvTE, 2015; Schuurs & Van den Bergh, 2015). Het merendeel van de opmerkingen betreft de voorschriften voor de correctie van de open vragen in het examen. Veel opmerkingen in het LT-verslag

betreffen antwoorden die, volgens docenten en het sectiebestuur Nederlands van Levende Talen, ook goed gerekend moeten worden naast dat wat er in het officieel geldende correctievoorschrift staat. Zo is het LT-verslag bij vwo 2015-1 bij vier open vragen milder dan het officiële correctievoorschrift, bij drie open vragen iets milder of gelijk, bij zes open vragen identiek en bij één open vraag strenger (dat wil zeggen preciezer geformuleerd) dan het officiële correctievoorschrift. Dit roept de vraag op wat de invloed is van het gebruik van het LT-verslag op de hoogte van de examenscores. Bovendien is het de vraag of het gebruik ervan bij de correctie tot betrouwbaarder scores leidt. Voor zover we weten is naar deze vragen tot dusverre geen onderzoek gedaan. Om die reden hebben we najaar 2015 een experimenteel onderzoek uitgevoerd naar de invloed van het gebruik van het LT-verslag op genoemde variabelen. Er waren dus twee onderzoeksvragen:

- Welke invloed heeft het gebruik van het LT-verslag op de hoogte van de examenscores.
- Welke invloed heeft het gebruik van het LT-verslag op de betrouwbaarheid van de examenscores?

Het examen Nederlands is erg correctie-intensief (Kuhlemeier & Van der Molen, 2016) en kent veel open vragen met een principieel onvolledig beoordelingsmodel. Bij de open vragen Nederlands is het, net als bijvoorbeeld bij geschiedenis, onmogelijk om van tevoren een volledige lijst met goed te rekenen antwoorden op te stellen (vgl. ook Kuhlemeier et al., 2012).[1]

Vooraf hebben we enkele verwachtingen geformuleerd over de mogelijke invloed van het LT-verslag op de hoogte van de beoordelingen en op de betrouwbaarheid van de scores.

Waar het de hoogte van de scores betreft, waren er twee uiteenlopende verwachtingen. Veel opmerkingen van docenten betreffen andere verwoordingen van het officieel gepubliceerde correctiemodel en aanvullende antwoorden die goed gerekend zouden moeten worden. Daarom is het, aan de ene kant, redelijk om te verwachten dat de in het LT-verslag opgenomen toevoegingen op het officiële correctievoorschrift leiden tot een versoepeling van de beoordelingsnormen en dus tot hogere scores. Een examenvraag waarbij dat het geval kan zijn, is vraag 22 die in figuur 1 onverkort is weergegeven.

In het LT-verslag voor vwo 2015 zijn toevoegingen aanwijsbaar die opgevat mogen worden als een versoepeling van het correctievoorschrift bij vraag 22:

> ‘In de vraag gaat het over de houding die de gebruikers aannemen ten opzichte van de nadelen van internet en sociale media. In de kolomtitels van het schema keert het woord ‘nadelen’ niet terug.
>
> Leerlingen kunnen dus twee sporen volgen: “de houding ten opzichte van internet en social media” en daarnaast “de houding ten opzichte van de nadelen van internet en social media”. Van belang is dat er een logisch verband is tussen kolom 1 en 2, per doelgroep.’ (LT-verslag 2015)

In het vervolg van het LT-verslag zijn deze beide sporen verder uitgewerkt en van voorbeelden voorzien.

Aan de andere kant is het ook niet onaannemelijk dat de toevoegingen in het LT-verslag leiden tot een strenger beoordelingsvoorschrift, met lagere scores als gevolg. Het LT-verslag bevat dan ook opmerkingen die in die richting wijzen, zoals bij vraag 25 (figuur 2).

Die opgave vraagt aan leerlingen om het tekstdoel en de beoogde doelgroep van verschillende teksten te benoemen. In het officiële correctievoorschrift wordt daarbij

opgemerkt: ‘alles wat niet beperkt is tot een specifieke doelgroep kan worden goed gerekend.’ Die opmerking ging in de ogen van de docentenvergadering van Levende Talen duidelijk te ver en in het LT-verslag wordt daaraan dan ook toegevoegd:

> ‘De opmerking dat alles wat niet beperkt is tot een specifieke doelgroep goed kan worden gerekend, betekent niet dat alles juist is. Mogelijke alternatieve juiste antwoorden voor de doelgroep van de hoofdtekst zijn “burgers”, “lezers van *De Groene Amsterdammer*”, “hoogopgeleiden”, etc.
>
> Niet juist zijn antwoorden zoals: “alle internet- en sociale mediagebruikers”, “internetgebruikers”, “het volk”.’ (LT-verslag 2015)

Overigens bevat het LT-verslag vaak bij een en dezelfde vraag zowel toevoegingen die het correctievoorschrift soepeler maken als toevoegingen die tot een strengere beoordeling

6p **22** In de tekst wordt de houding besproken van internetgebruikers, het bedrijfsleven en de overheid ten opzichte van de nadelen die internet en de sociale media met zich meebrengen.
Neem onderstaand schema over en vat daarin samen welke houding elk van deze groepen volgens de tekst tegen deze nadelen aanneemt en waar die houding uit voortkomt.

GROEP	houding ten opzichte van internet en sociale media	houding komt voort uit
internetgebruikers		
bedrijfsleven		
overheid		

Figuur 1

2p **25** De hoofdtekst en tekstfragment 3 verschillen fundamenteel van elkaar doordat ze gericht zijn op andere doelgroepen en op andere tekstdoelen.
Karakteriseer het verschil tussen beide teksten door van beide het tekstdoel en de doelgroep te benoemen.
Kies bij tekstdoel uit de volgende termen: betogend, beschouwend, expressief of informerend.

Figuur 2

3p	**29**	Blijkens alinea 5 heeft de bemoeienis van de National Science Foundation een voordeel, maar ook een nadeel. Benoem dit voordeel en dit nadeel. Gebruik voor je antwoord niet meer dan 20 woorden.

Figuur 3

leiden, zodat de tegengestelde effecten elkaar wellicht uitmiddelen. Zoiets is bijvoorbeeld het geval bij vraag 29 (figuur 3).

Bij deze vraag bevat de toevoeging in het LT-verslag zowel een uitbreiding van de goed te rekenen antwoorden als een inperking:

> 'Alternatieve formuleringen van het nadeel zoals 'het beperkt de onderzoeker' zijn ook juist. Het gaat immers om de beperking van de mogelijkheden tot ontdekken/onderzoeken.
>
> Niet juist zijn formuleringen zoals "het levert beperkingen op", "het vertraagt, daardoor gaat het niet snel" of "er is geen plek meer voor vooruitgang".' (LT-verslag 2015)

We verwachtten – anders gezegd – dat de invloed van het LT-verslag op de hoogte van de scores uiteenlopend is, afhankelijk van de aard van de in het verslag opgenomen toevoegingen per vraag. Daarbij maken we onderscheid in drie categorieën:

type A: de toevoegingen maken het correctievoorschrift soepeler, wat naar verwachting tot een hogere score leidt;

type B: de toevoegingen bestaan uit opmerkingen die het correctievoorschrift soepeler maken, maar evenzeer uit opmerkingen die het correctievoorschrift strenger maken, zodat het verwachte effect per saldo nul is;

type C: de toevoegingen maken het correctievoorschrift strenger, wat zoals verwacht een lagere score tot gevolg heeft.

Waar het de betrouwbaarheid van de scores betreft, mag worden verwacht dat het LT-verslag een positief effect heeft. De voorbeelden en preciseringen die ter vergadering genoemd worden zijn immers meestal gebaseerd op antwoorden die leerlingen hebben gegeven. Men creëert er als het ware consensus over welke antwoorden nog wel en welke niet meer goed gerekend moeten worden. Bovendien vermindert dit waarschijnlijk het aantal keren dat een individuele docent een beroep doet op vakspecifieke regel 3.3. Deze regel biedt ruimte om scorepunten toe te kennen aan een antwoord op een open vraag dat niet in het beoordelingsmodel voorkomt, terwijl dit antwoord op grond van aantoonbare, vakinhoudelijke argumenten wel als (gedeeltelijk) juist aangemerkt kan worden. De gezamenlijke bespreking en de voorbeeldantwoorden zorgen er samen voor dat er meer op dezelfde manier wordt beoordeeld, wat de betrouwbaarheid ten goede komt. Zelfs al wordt er ter vergadering niets aan het officiële correctievoorschrift toegevoegd, dan nog fungeert de bespreking van een open vraag en de goed te rekenen antwoorden daarop in elk geval voor de aanwezigen als een oefening in het corrigeren van het leerlingwerk.

Methode van onderzoek

Het hier gerapporteerde onderzoek had betrekking op het vwo-examen Nederlands van 2015. Dit examen bestaat uit 36 vragen met maximumscores van 1, 2, 3, 4 of 6. Er waren 12 open vragen en 24 meerkeuzevragen; van de 36 vragen waren er 14 die in het LT-verslag van een toevoeging werden voorzien. Bij dit examen waren maximaal 63 punten te behalen. De gemiddelde score was 32,25.

Ten behoeve van het onderzoek is bij een aantal scholen geanonimiseerd examenwerk opgevraagd. Daaruit zijn twintig examenwerken geselecteerd, afkomstig van vijf zwakke, tien gemiddelde en vijf goede kandidaten. Van de examenwerken is een digitale 'blanco' kopie gemaakt (nadat deze van een identificatiecode en een paginering waren voorzien). Van deze digitale kopie werden de gegevens van de vestiging, de scorepunten en aantekeningen van de eerste corrector digitaal verwijderd evenals – waar nog nodig – de voor- en achternaam van de kandidaat.

Ten behoeve van dit correctie-experiment hebben we gewerkt met twee versies van het correctievoorschrift: naast de officiële versie zoals door het CvTE gepubliceerd (hierna aangeduid als CV-min) hebben we een tweede versie gemaakt waarin per vraag de toevoegingen van de vakvereniging onverkort en ongewijzigd zijn opgenomen (hierna CV-plus genoemd).

De twintig examenwerken zijn nagekeken door achttien correctoren. Alle correctoren hadden minstens drie jaar ervaring als eerste en/of tweede corrector en allen gaven aan in eerdere jaren de toevoegingen van de vakvereniging gebruikt te hebben. We hebben geprobeerd om ook docenten Nederlands te werven die in 2015 geen examenklas vwo hadden en daardoor met dit specifieke examen geen eerdere correctie-ervaring hadden. Dat is maar ten dele gelukt: vijf docenten hadden in 2015 geen vwo-examenklas onder hun hoede, de anderen wel. Om de invloed van het cv-min zuiver te kunnen vergelijken met die van het cv-plus is het van belang dat de beide groepen docenten die ermee nakijken vergelijkbaar zijn. Als de ene groep bijvoorbeeld meer strenge correctoren telt dan de andere groep, kan dat een eerlijke vergelijking in de weg staan. Daarom heeft de toewijzing van correctoren aan de twee condities cv-min en cv-plus plaatsgevonden op basis van de resultaten van een korte vragenlijst. Die bevatte zeven frequentievragen van het type nooit=1, soms=2 en altijd =3. Eén vraag ging over het gebruik van de toevoegingen van de vakvereniging, zes vragen over de strengheid van de docent als corrector en als deelnemer aan het gezamenlijk overleg, en één vraag over het verschil tussen de gemiddelde cijfers op het centraal examen en het schoolexamen. Na rangordening van de correctoren op basis van hun totaalscore over de zeven vragen zijn de correctoren om-en-om aan de condities CV-min en CV-plus toegewezen.

Uitvoering van het onderzoek

Elke corrector heeft alle twintig geselecteerde examenwerken nagekeken. Van de achttien correctoren keken er negen na aan de hand van het CV-min en negen met het CV-plus. De correctoren hebben de examenwerken op een zelf gekozen tijdstip en locatie gecorrigeerd. Zij hadden de opdracht het werk zo objectief mogelijk na te kijken aan de hand van het originele correctievoorschrift (in de CV-min conditie) of het correctievoorschrift met de toevoegingen van de vakvereniging (in de CV-plus conditie). De docenten in de CV-min-conditie is gevraagd de eventuele kennis van het verslag van de vakvereniging niet te gebruiken. Zij zijn er uitdrukkelijk op gewezen dat de gegevens niet meer bruikbaar zouden zijn voor het onderzoek als zij dat wel zouden doen.

CORRECTIE-VOORSCHRIFT	AANTAL CORRECTOREN	AANTAL EXAMEN-SCORES	MIN	MAX	GEM	STD.DEV
CV-min	9	180	22	51	36,85	5,97
CV-plus	9	180	24	50	37,43	6,31

Tabel 1. Gemiddelde examenscores per correctievoorschrift

Resultaten

De invloed op de hoogte van de scores

Tabel 1 geeft de gemiddelde totaalscore per type correctievoorschrift. Het verschil tussen de gemiddelde scores van degenen die met het CV-min en het CV-plus beoordeelden blijkt -0,58. Dit lijkt erop te wijzen dat het CV-plus tot iets hogere scores heeft geleid dan het CV-min. Het verschil in het voordeel van het CV-plus is echter in statistisch opzicht niet significant (op 5%-niveau). Hiermee kan niet worden aangetoond dat de toevoegingen uit het LT-verslag een betekenisvolle invloed hebben op de hoogte van de examenscores.

In tabel 1 is er geen rekening gehouden met het gegeven dat slechts 14 van de 36 vragen van een toevoeging zijn voorzien. Baseren we ons op de 22 vragen zonder toevoeging, dan is het verschil tussen de scores van de beide groepen die met het CV-min en het CV-plus werkten met -0,01 nagenoeg gelijk aan nul (waarbij een negatieve waarde aangeeft dat de correctie met CV-plus heeft geleid tot een hogere score). Dit wijst erop dat de twee groepen beoordelaars hun werk goed gedaan hebben.

Baseren we ons alleen op de vragen met een toevoeging, dan bedraagt het scoreverschil tussen beide correctievoorschriften -0,57 in het voordeel van het CV-plus. Het eerder gerapporteerde verschil in het voordeel van het CV-plus op het examen als geheel komt dus vrijwel volledig voor rekening van de vragen met een toevoeging. Een significantietoetsing brengt echter aan het licht dat het verschil tussen beide correctievoorschriften in statistisch opzicht niet van betekenis is (op 5%-niveau). Kijken we dus alleen naar de vragen met een toevoeging, dan nog maakt het niet uit of de correctoren met of zonder de toevoegingen beoordeeld hebben.

Zoals eerder uiteengezet, verwachten we van toevoegingen van type A een verhoging van de examenscores en van type C een verlaging, terwijl toevoegingen van type B neutraal van aard zijn. Omdat bij sommige opgaven de toevoegingen alle in één richting gingen, hebben we daar het te verwachten effect aangeduid met AA of met CC. In tabel 2 zijn de gemiddelde scores op de vragen met een toevoeging uitgesplitst per vraag per type toevoeging. Behalve de gemiddelden toont de tabel ook de standaarddeviaties en de effectgrootte van het verschil tussen het gemiddelde voor CV-plus minus CV-min. In de effectgrootte is het verschil tussen beide gemiddelden uitgedrukt als een proportie van de standaarddeviatie.

Tabel 2 laat zien dat vier examenvragen uit het beoordelingsmodel Nederlands een toevoeging van type A kennen, negen vragen een toevoeging van type B en één vraag een toevoeging van type C. Voor de type A-toevoegingen bedraagt het totaal van de verschilscores tussen de gemiddelden van CV-plus minus CV-min -0,95 en bij de type C-toevoegingen gaat het om 0,91 (met de kanttekening dat dit verschil door één van de toevoegingen bij deze opgave is veroorzaakt). Bij de opgaven met neutrale toevoegingen bedraagt het totaal van de verschilscores 0,12. De effecten van beide typen toevoegingen middelen elkaar vrijwel volledig uit: het 'netto' effect kan dus als volstrekt verwaarloosbaar worden beschouwd.

Kijken we naar de effectgroottes per vraag, dan zien we dat die bij vijf opgaven buiten de bandbreedte van -0,20 tot +0,20 ligt. Dat zijn de opgaven waarbij de toevoegingen vanuit het LT-verslag er mogelijk toe doen. De opgaven 10, 13, 21 en 22 zijn soepeler beoordeeld met het LT-verslag. Dit is in overeenstemming met onze verwachtingen. Bij opgave 25 is er juist strenger beoordeeld. Ook dit is in overeenstemming met onze verwachtingen: de toevoeging bij vraag 25 betreft immers een type C-toevoeging die naar soepelheid neigende docenten ervan moet weerhouden aan een fout antwoord punten toe te kennen (zie de eerder geciteerde opmerking uit het LT-verslag bij vraag 25).

Examenvraag	Type vraag*	CVmin		CVplus		Verschilscore	Effectgrootte
		Gem	Std.dev.	Gem	Std.dev		
7	B	1,79	0,90	1,74	0,90	0,04	0,05
10	A	2,38	1,13	2,59	1,04	-0,22	**-0,20**
13	A	1,11	0,99	1,38	1,07	**-0,27**	**-0,26**
15	B	1,19	0,98	1,25	0,97	-0,06	-0,06
18	B	1,35	0,91	1,26	0,92	0,09	0,10
19	B	0,68	0,98	0,77	0,84	-0,09	-0,10
20	B	0,64	0,48	0,63	0,48	0,02	0,03
21	A	1,27	0,67	1,46	0,56	-0,19	**-0,31**
22	AA	3,73	1,26	4,00	1,26	**-0,27**	**-0,22**
25	CC	0,91	0,74	0,64	0,67	**0,26**	**0,37**
26	B	1,56	1,19	1,61	1,29	-0,04	-0,04
29	B	1,76	1,07	1,59	0,99	0,17	0,16
31	B	0,61	0,92	0,61	0,89	0,00	0,00
33	B	0,62	0,58	0,62	0,58	-0,01	-0,01

*A = soepeler, B = neutraal, C = strenger

Tabel 2. Gemiddelden, standaarddeviaties en effectgrootte van het verschil tussen scores met het CV-min en CV-plus voor de vragen met een toevoeging in het CV-plus Nederlands

Correctie-voorschrift	Aantal correctoren	Percentage exacte overeen-stemming	Cohens multi-rater Kappa	Standaardfout	95%-betrouw-baarheids-interval
ALLE VRAGEN					
CV-min	9	89	0,82	0,01	0,80 - 0,83
CV-plus	9	90	0,85	0,01	0,83 - 0,86
ALLEEN OPEN VRAGEN					
CV-min	9	69	0,59	0,01	0,57 - 0,60
CV-plus	9	73	0,65	0,01	0,63 - 0,67

Tabel 3. Percentage exacte overeenstemming en Cohens multi-rater Kappa per correctievoorschrift

De invloed op de beoordelaarsovereenstemming Nederlands

Een eenvoudige maat voor de mate waarin beoordelaars het met elkaar eens zijn, is het percentage exacte overeenstemming. Deze maat houdt er echter geen rekening mee dat correctoren alleen al op basis van toeval tot op zekere hoogte met elkaar zullen overeenstemmen. Denk bijvoorbeeld aan een vierkeuze-opgave, 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). 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: 0,00 tot 0,20 is *slight*, 0,21 tot 0,40 is *fair*, 0,41 tot 0,60 is *moderate*, 0,61 tot 0,80 is *substantial* en 0,81 tot 1,00 is *almost perfect*. De overeenstemming tussen de correctoren is bepaald met *Cohen's multi-rater Kappa*. De mate van overeenstemming over de totaalscores op het examen als geheel is bepaald met behulp van de 'gewone' product-moment correlatie.

Overeenstemming per examenvraag

Tabel 3 toont het gemiddeld percentage exact gelijke scores en Cohens multi-rater Kappa, afzonderlijk voor de negen correctoren die zonder en de negen correctoren die met de toevoegingen nakeken.

Voor het CV-min bedraagt het gemiddelde percentage exact gelijke scores 89% en voor het CV-plus 90%. Cohens Kappa bedraagt 0,82 versus 0,85. Aangezien de betrouwbaarheidsintervallen van de beide Kappa's elkaar overlappen, is hier geen sprake van een significant verschil op 5%-niveau. Kijken we alleen naar de open vragen, dan blijken de percentages overeenstemming en de Kappa's beduidend lager. Nu is het betrouwbaarheidsverschil tussen CV-min en CV-plus wel significant op 5%-niveau, aangezien de betrouwbaarheidsintervallen elkaar niet overlappen.

Overeenstemming over de examenscores

Ter beantwoording van de vraag of de toepassing van het CV-plus inderdaad tot meer

overeenstemming leidt, zijn de correlaties berekend tussen de examenscores van de correctoren die met het CV-min en het CV-plus werkten. Het verschil blijkt klein: voor het CV-min bedraagt de mediaan over de 36 correlaties tussen de examenscores van de negen correctoren 0,89 en voor het CV-plus is de mediaan over hetzelfde aantal correlaties 0,91. De laagste correlatie in de groep CV-min bedraagt 0,81 en de hoogste correlatie is 0,97. Voor het CV-plus bedraagt de laagste correlatie 0,77 en de hoogste correlatie 0,97. Deze correlaties wijzen op een grote overeenstemming tussen de correctoren.

Samenvatting en discussie

Er is experimenteel onderzocht welk effect het gebruik van het LT-verslag heeft op de hoogte van de examenscores en de betrouwbaarheid van de correctie. Het onderzoek is uitgevoerd aan de hand van examen Nederlands vwo 2015-1. Achttien correctoren hebben ieder het examenwerk van twintig kandidaten gecorrigeerd, waarvan negen met het LT-verslag (cv-plus) en negen zonder dat verslag (cv-min). Voor de hoogte van de scores op beide examens blijkt het niet uit te maken of de correctoren met of zonder de toevoegingen van de vakvereniging nakeken. De overeenstemming tussen de correctoren bleek wel beïnvloed door het gebruik van het LT-verslag: bij de open vragen vonden we een betrouwbaarheidsverschil in het voordeel van conditie CV-plus, zowel op itemniveau als op basis van de totale examenscore. De conclusie is dat open vragen op basis van de toevoegingen van het LT-verslag iets betrouwbaarder beoordeeld worden; dit effect van de aanvullingen bij de open vragen is echter klein en komt daardoor niet tot uiting in een significant hogere betrouwbaarheid van het examen als geheel.

Generalisatie naar de onderwijspraktijk

De onderzoeksresultaten zijn verzameld in de artificiële context van een onderzoek en daarmee niet zonder meer generaliseerbaar naar de 'echte' examenpraktijk. Hieronder bespreken we vier verschillen tussen het onderzoek en de correctiepraktijk die de generaliseerbaarheid kunnen beperken:

Ten behoeve van dit onderzoek zijn de opmerkingen uit het LT-verslag gecombineerd met het officiële cv: per vraag zijn eerst de officiële richtlijnen vermeld en direct daarna de opmerkingen vanuit het LT-verslag. Dit wijkt af van de praktijk waarin de corrector te midden van alle leerlingwerk, het vragenboekje, het tekstboekje, het officiële correctievoorschrift én het extra document met LT-toevoegingen zit. Mogelijk bestaat in die situatie een grotere kans dat af en toe een van genoemde documenten niet wordt geraadpleegd.

Anders dan in de echte examensituatie beoordeelden de docenten uit het onderzoek alleen andermans leerlingen. Er stonden dus geen persoonlijke belangen op het spel (denk bijvoorbeeld aan de neiging van eerste correctoren om voor hun eigen kandidaten het onderste uit de kan te halen). Uit onderzoek blijkt dat objectieve derde correctie van andermans leerlingen in een onderzoekssituatie gemiddeld tot lagere scores leidt dan eerste correctie van 'eigen' leerlingen in de 'echte' examenpraktijk (Kuhlemeier & Kremers, 2013). Mogelijk hebben de correctoren uit ons onderzoek strenger nagekeken dan zij gewend zijn, wat zou leiden tot een onderschatting van de examenprestaties in het huidige onderzoek. Nader onderzoek zou hierover meer uitsluitsel kunnen geven. Te denken valt daarbij aan een non-invasief veldonderzoek tijdens de correctieperiode waarbij de examenprestaties bij CV-plus gebruikers worden vergeleken met die bij CV-min gebruikers.

In het onderzoek corrigeerde iedere docent

twintig examenwerken (met het CV-min of met het CV-plus). In de echte examensituatie corrigeert elke docent echter gemiddeld minimaal twee keer zoveel leerlingen als in ons onderzoek. Vrijwel elke docent is in werkelijkheid immers zowel eerste als tweede corrector.

Een ander 'onnatuurlijk' element is dat vrijwel alle correctoren het examen al eerder als eerste en/of tweede corrector hadden nagekeken. Onduidelijk is in hoeverre het de docenten in de CV-min conditie gelukt is de eventuele kennis van de toevoegingen daadwerkelijk te 'onderdrukken'. Mogelijk is de invloed van de toevoegingen van de vakverenigingen in de 'echte' examenpraktijk groter dan uit ons onderzoek blijkt.

Aanbevelingen voor verder onderzoek

Voor eventueel vervolgonderzoek naar de invloed van toevoegingen op het correctievoorschrift op de hoogte van de scores, de betrouwbaarheid en de tijdbesteding aan de correctie en het gezamenlijk overleg verdient het aanbeveling meer examens en meer correctoren in het onderzoek te betrekken.

In vervolgonderzoek verdient veldonderzoek tijdens de correctieperiode de voorkeur. Natuurlijk is het bij echte examencorrectie niet mogelijk om correctoren random aan een CV-plus en een CV-min conditie toe te wijzen. Wel zou de onderzoeker gebruik kunnen maken van de natuurlijke variatie in het gebruik van de toevoegingen door correctoren: een deel van de correctoren gebruikt het LT-verslag immers niet en binnen de groep die dat wel doet, kan een onderscheid worden gemaakt in lichte, middelmatige en integrale gebruikers. Een indicatie van het effect van de toevoegingen kan worden verkregen door de mate van het gebruik (getrouwheid en eventueel tijdbesteding) te relateren aan de beschikbare examenprestaties (o.a. hoogte van de scores; interne consistentie; overeenstemming tussen eerste en tweede corrector). Wel dient het gebruik van de toevoegingen dan nauwgezet geregistreerd te worden, bijvoorbeeld aan de hand van het ontwikkelen van profielen van onvoldoende, minimale, middelmatige (acceptabele) en volledige (ideale) implementatie (Leithwood & Montgomery, 1987).

Besluit

Zoals door het CvTE was aangekondigd in onder andere de zogenoemde Maartmededeling 2017 (zie www.examenblad.nl), heeft er in 2017 onder regie van het CvTE een test Correctievoorschrift plaatsgevonden voor het vwo-examen Nederlands: op woensdagmiddag 10 mei 2017 kregen 20 vwo-docenten Nederlands het correctievoorschrift toegestuurd van het examen dat die dag was afgenomen. Aan de hand van dat correctievoorschrift hebben zij elk een aantal leerlingwerken gecorrigeerd. Hun bevindingen zijn een dag later onder begeleiding van het CvTE besproken. Nog diezelfde avond en de daarop volgende ochtend heeft de Vaststellingscommissie Nederlands van het CvTE de daaruit voortvloeiende suggesties voor wijziging besproken en voor zover mogelijk verwerkt in het correctievoorschrift. Daarna is het correctievoorschrift definitief vastgesteld en gepubliceerd.

Deze test is een interessante vorm van co-creatie die, wanneer de evaluatie ervan positief uitvalt, docenten Nederlands in de toekomst extra steun kan bieden bij de correctie. Het CvTE zal verslag doen van een evaluatie van deze test. Wij noemen de test omdat de gedachte post zou kunnen vatten dat ze de LT-examenbesprekingen en de verslaglegging daarvan overbodig zou maken. De LT-bespreking en verslaglegging wijkt echter op minstens twee manieren af van wat er in de test van het CvTE gebeurde: ze biedt de mogelijkheid tot bespreking van het examen in een ruimere context, en bovendien genereert de LT-bijeenkomst een veelheid aan andersoortige verwoordingen. Zo biedt het LT-verslag

2015 – onder het kopje 'Voorbeelden' – bij acht open vragen in totaal maar liefst 48 alternatieve antwoorden die ontleend zijn aan leerlingwerk, elk voorzien van een indicatie van het aantal toe te kennen punten. Dit moge duidelijk maken dat het LT-verslag naast een gecontinueerde test Correctievoorschrift altijd nog bestaansrecht heeft.

Noot

1. Een parallel onderzoek is uitgevoerd voor het havo-examen Geschiedenis. Daarover is verslag gedaan door Kuhlemeier, Boom, Gitsels & Schuurs (in druk).

Literatuur

Cito/CvTE (2015). Terugblik CE Nederlands vwo 2015. www.examenblad.nl/evaluatie/terugblik-ce-nederlands-vwo-2015/2015/vwo/f=/terugblik_CE_Nederlands_vwo_2015_def.pdf

Cohen, Jacob (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Kuhlemeier, H., Boom, S., Gitsels, H. & Schuurs, U. (in druk). Maken de toevoegingen van de vakverenigingen op de correctievoorschriften het CSE VO betrouwbaarder? Te verschijnen in *Examens, Tijdschrift voor de toetspraktijk*, 2017–4.

Kuhlemeier, H., Gitsels, H., Boom, S., Kerkhof, A. van de, & Sinkeldam, R. (2012). *Examenkenmerken en verschillen tussen correctoren: verslag van een panelonderzoek bij het CSE geschiedenis, tehatex en Nederlands.* Arnhem: Cito.

Kuhlemeier, H,. & Kremers, E. (2013). Eerste, tweede, derde en vierde correctie: Wat is het verschil? *Examens, Tijdschrift voor de toetspraktijk*, 3(10), 6–10.

Kuhlemeier, H., & Molen, P. van der (2016). *De praktijk van de eerste en tweede correctie van het CSE. Technische rapportage van de landelijke enquête van 2016 in vergelijking met 2011.* Arnhem: Cito.

Landis, J.R., & Koch, G.G. (1997) The measurement of observer agreement for categorical data. *Biometrics*, 1, 159–174.

Leithwood, K.A., & Montgomery, D.J. (1987). *Improving classroom practice: Using innovation profiles*. Toronto: The Ontario Institute for Studies in Evaluation.

LT-verslag (2015). Verslag eindexamenbespreking Nederlands 13 mei 2015. https://nederlandslevendetalen.files.wordpress.com/2016/03/verslagen-examenbesprekingen-lt-2006-2015.zip

Schuurs, U., & Bergh, H. van den (2015): Veranderingen in het centrale examen Nederlands en de mening van docenten. In Mottart, A. & S. Vanhooren (red.), 29e *Conferentie Onderwijs Nederlands* (pp. 111–115). Gent: Academia Press.

Uriël Schuurs studeerde Nederlands en Toegepaste Taalwetenschap, promoveerde in 1991 op onderzoek naar schrijfonderwijs en werkte achtereenvolgens als onderzoeker en docent aan Universiteit Utrecht en aan de Radboud Universiteit Nijmegen. Momenteel werkt hij onder andere bij Cito, waar hij betrokken is bij de ontwikkeling van een veelheid aan toetsen.
E-mail: uriel.schuurs@cito.nl

Hans Kuhlemeier studeerde onderwijskunde aan de Katholieke Universiteit te Nijmegen, promoveerde in 1996 op een onderzoek naar de structuur van taalvaardigheid en werkt als onderzoeker bij Cito.
E-mail: hans.kuhlemeier@cito.nl

Hugo Gitsels studeerde aan de Amsterdamse Hogeschool voor de Kunsten en kunstgeschiedenis aan de Vrije Universiteit Amsterdam. Sinds 2008 werkt hij als toetsdeskundige bij Cito waar hij zich bezig houdt met ontwikkeling van centrale examens en met onderzoek gericht op beoordeling- en correctiepraktijk.
E-mail: hugo.gitsels@cito.nl