

Examenkenmerken en verschillen tussen correctoren

□ Hans Kuhlemeier, Esther Kuijs, Stefan Boom, Hugo Gitsels en Ed Kremers

In de twee eerdere nummers van *EXAMENS* lieten Kuhlemeier en Kremers zien dat onafhankelijke correctoren gemiddeld beduidend lagere scores toekennen dan de 'eigen' docent. Eerste correctoren bleken sterk te verschillen in de soepelheid van de beoordeling. In dit derde artikel wordt nagegaan in hoeverre verschillen tussen correctoren te maken hebben met kenmerken van de examenvraag en het correctievoorschrift. Recent hebben Cito en CvE maatregelen genomen om de examenvragen en de antwoordmodellen te verduidelijken.

Beoordelen van examenwerk geen sinecure

Het correctievoorschrift

Correctoren van het centraal schriftelijk examen moeten het werk van hun kandidaten nakijken volgens het correctievoorschrift. Het beoordelen van antwoorden op open vragen is niet altijd eenvoudig. Behalve de algemene en vakspecifieke regels voor de beoordeling moet de corrector het antwoordmodel grondig bestuderen. Daarnaast moet de corrector over voldoende vakinhoudelijke kennis beschikken. Bij beoordeling van antwoorden op open vragen zijn beoordelingsverschillen tussen correctoren niet volledig te vermijden. Bij sommige vakken is het vaak niet mogelijk om alle geheel of gedeeltelijk juiste

antwoorden die kandidaten kunnen geven in het antwoordmodel op te nemen. Het antwoordmodel volstaat dan noodgedwongen met een beperkt aantal voorbeelden van juiste antwoorden of geeft alleen aan wat de strekking van een juist antwoord is. Naarmate het antwoordmodel de verzameling van mogelijke antwoorden minder goed dekt, heeft de corrector minder houvast bij het beoordelen. Er ontstaat dan meer ruimte voor eigen interpretaties en verschillen in soepelheid.

Algemene beoordelingsregels

Het correctievoorschrift bevat een aantal algemene beoordelingsregels (zie kader). Die moet de corrector toepassen als het antwoordmodel onverhoopt te weinig sturing biedt. Naast de algemene beoordelingsregels zijn er bij sommige vakken vakspecifieke regels. Zo kennen de examens geschiedenis, aanvullend op de hiervoor genoemde algemene regel, de eis dat 'Vakinhoudelijke argumenten afkomstig moeten zijn uit gezaghebbende, wetenschappelijke publicaties'. Als een antwoord op een open vraag niet in het antwoordmodel voorkomt, mag de corrector dus alleen scorepunten toekennen naar analogie of in de geest van het antwoordmodel als hij of zij de volledige of gedeeltelijke juistheid van het antwoord aan de hand van gezaghebbende, wetenschappelijke publicaties kan aantonen.

Al met al trekt de correctie van examenwerk een zware wissel op de (vakinhoudelijke) deskundigheid van de docent. De beoordelingspraktijk leert dat correctoren in situaties waarin het antwoordmodel niet of onvoldoende sturing biedt vaak verschillende scores toekennen, zonder dat altijd onmiddellijk duidelijk is wie het bij het rechte eind heeft.



De beoordeling van open vragen trekt een zware wissel op de vakinhoudelijke deskundigheid van de docent

Artikel 3. Algemene regels

3. Scorepunten worden toegekend met inachtneming van de volgende regels:

- 3.1 indien een vraag volledig juist is beantwoord, wordt het maximaal te behalen aantal scorepunten toegekend;
- 3.2 indien een vraag gedeeltelijk juist is beantwoord, wordt een deel van de te behalen scorepunten toegekend, in overeenstemming met het beoordelingsmodel;
- 3.3 indien een antwoord op een open vraag niet in het beoordelingsmodel voorkomt en dit antwoord op grond van aantoonbare, vakinhoudelijke argumenten als juist of gedeeltelijk juist aangemerkt kan worden, moeten scorepunten worden toegekend naar analogie of in de geest van het beoordelingsmodel;
- 3.4 indien slechts één voorbeeld, reden, uitwerking, citaat of andersoortig antwoord gevraagd wordt, wordt uitsluitend het eerstgegeven antwoord beoordeeld;
- 3.5 indien meer dan één voorbeeld, reden, uitwerking, citaat of andersoortig antwoord gevraagd wordt, worden uitsluitend de eerstgegeven antwoorden beoordeeld, tot maximaal het gevraagde aantal;
- 3.6 indien in een antwoord een gevraagde verklaring of uitleg of afleiding of berekening ontbreekt dan wel foutief is, worden 0 scorepunten toegekend tenzij in het beoordelingsmodel anders is aangegeven;
- 3.7 indien in het beoordelingsmodel verschillende mogelijkheden zijn opgenomen, gescheiden door het teken /, gelden deze mogelijkheden als verschillende formuleringen van hetzelfde antwoord of onderdeel van dat antwoord;
- 3.8 indien in het beoordelingsmodel een gedeelte van het antwoord tussen haakjes staat, behoeft dit gedeelte niet in het antwoord van de kandidaat voor te komen;
- 3.9 indien een kandidaat op grond van een algemeen geldende woordbetekenis, zoals bij voorbeeld vermeld in een woordenboek, een antwoord geeft dat vakinhoudelijk onjuist is, worden aan dat antwoord geen scorepunten toegekend, of tenminste niet de scorepunten die met de vakinhoudelijke onjuistheid gemoeid zijn.

Bron: http://wetten.overheid.nl/BWBR0027449/geldigheidsdatum_31-03-2014

Het panelonderzoek vierde correctie (P4C)

In het zogeheten panelonderzoek vierde correctie (P4C) is nagegaan in hoeverre verschillen tussen de beoordelingen van de 'eigen docent' en onafhankelijke vierde correctoren samenhangen met kenmerken van de vraagstelling en het correctievoorschrift (Kuhlemeier e.a., 2013). Panels van getrainde docenten keken ieder zestig

antwoorden op vijf vragen uit de examens geschiedenis, tehatex en Nederlands opnieuw na. Gekozen zijn de vragen en antwoorden uit het hoofdonderzoek (Kuhlemeier, Van Rijn & Kremers, 2013; Kuhlemeier & Kremers, 2013a) waarbij de scores van de eigen docent en de onafhankelijke correctoren het verst uit elkaar lagen. Het panelonderzoek bevestigde het bestaan van grote verschillen tussen eerste correctoren in de soepelheid van de beoordeling. Daarnaast bleken de verschillen tussen correctoren samen te hangen met kenmerken van de vraag en het antwoordmodel.

Welke examenkenmerken geven aanleiding tot verschillen tussen correctoren?

Verschillen tussen correctoren blijken vooral voor te komen in de volgende situaties:

1. *Het antwoordmodel is onvolledig.* Kandidaten geven veelvuldig antwoorden die niet in het antwoordmodel voorkomen. Bij examens zoals geschiedenis is de corrector dan aangewezen op algemene regel 3.3 en op de vakspecifieke regel. Het onderzoek maakt aannemelijk dat de scores van de correctoren verder uiteenlopen naarmate zij vaker een beroep moeten doen op algemene regel 3.3.
2. *Het antwoordmodel geeft alleen een voorbeeld van een volledig juist antwoord.* Het antwoordmodel biedt geen steun bij het herkennen van en het toekennen van punten aan minder juiste of foute antwoorden. Daardoor moet de corrector een beroep doen op algemene regel 3.2. Uit de panelbeoordelingen komt naar voren dat algemene regel 3.2 vaak aanleiding geeft tot verschillen tussen correctoren in soepelheid.
3. *Een inconsistentie tussen vraag en antwoordmodel.* Dit geeft aanleiding tot verschillen tussen correctoren in soepelheid. Er wordt bijvoorbeeld een kort en bondig antwoord gevraagd terwijl het antwoordmodel uit lange en complexe zinnen bestaat.
4. *De kandidaat geeft meer antwoorden dan volgens de vraagstelling is toegestaan.* De corrector moet dan algemene regel 3.5 toepassen. De ene corrector past deze algemene regel zoals het hoort wel toe

Verschillen tussen correctoren kunnen gedeeltelijk verklaard worden vanuit kenmerken van het examen



en de ander doet dat ten onrechte niet en geeft de kandidaat ook punten voor antwoorden boven het toegestane aantal.

5. *Het aantal te beoordelen inhoudselementen is groter dan de maximumscore.* Om één punt te verdienen moet de kandidaat bijvoorbeeld zowel een juiste uitleg geven als een juiste bron noemen. Als de kandidaat in deze situatie alleen een juiste uitleg geeft of alleen een juiste bron noemt, kent de ene corrector (zoals het hoort) 0 punten toe, terwijl de andere corrector dat één punt dan ten onrechte wel toekent.
6. *De kandidaat geeft een foute toelichting bij een juist antwoord of een goede toelichting bij een fout antwoord, zonder dat in de vraagstelling om verduidelijking of uitleg gevraagd wordt.* Verschillen tussen correctoren ontstaan als de ene corrector de toelichting wel in de beoordeling betreft en de ander dat niet doet.
7. *Het voorschrift van een als 0-1 te scoren vraag beschrijft meerdere kenmerken van een juist antwoord (element), zonder dat volledig duidelijk is hoeveel en welke kenmerken in het antwoord van de kandidaat aanwezig moeten zijn om dat ene punt te mogen toekennen.* De ene corrector is al blij met één van de kenmerken, terwijl de ander eist dat alle kenmerken in het antwoord aanwezig zijn.
8. *Het antwoordmodel is geformuleerd in academische vaktaal/jargon (bijvoorbeeld uit de katernen),* terwijl kandidaten hun antwoorden in alledaags Nederlands formuleren. Een voorbeeld is de omschrijving van een juist antwoord 'Het is een enorme 'blow-up' van een (klein) gebruiksvoorwerp' uit het antwoordmodel tehatex. De ene corrector eist een expliciete verwijzing naar een blow-up of een enorme uitvergroting van een gebruiksvoorwerp, terwijl de ander genoeg neemt met ieder antwoord waarin iets van grootte naar voren komt.
9. *Een dichotoom antwoordmodel dat alleen de scores 0 en 2 toestaat,* waarbij expliciet vermeld wordt dat gedeeltelijk juiste antwoorden de score 0 moeten

krijgen. Verschillen tussen correctoren ontstaan als de ene corrector voor een gedeeltelijk juist antwoord 0 punten toekent en de ander dat antwoord in strijd met het antwoordmodel met 1 scorepunt honoreert.

Interpretatie

Bij de interpretatie van de resultaten van het P4C-onderzoek moet men bedenken dat de onderzochte examens, kandidaten en examenvragen geen representatieve steekproef vormen uit de totale verzameling van examens, kandidaten en examenvragen. Er zijn slechts drie examens in het onderzoek betrokken die vooral gekozen zijn vanwege het grote aantal open vragen. Waren bijvoorbeeld examens zonder open vragen gekozen, dan zouden de resultaten met betrekking tot de kwaliteit van de vragen en de antwoordmodellen ongetwijfeld gunstiger zijn (maar had het onderzoek ook minder verbeteringssuggesties opgeleverd). Daarnaast beoordeelden de panels van het P4C-onderzoek alleen vragen en examenwerken waarbij het verschil tussen de scores van de 'eigen docent' en de onafhankelijke correctoren sterk positief was. De kans op het aantreffen van problematische vragen of te soepele correctoren is daardoor veel groter dan wanneer we de panels een representatieve steekproef van vragen en werken hadden voorgelegd. De resultaten van dit deel van het P4C-onderzoek zijn daarom niet geldig voor het gemiddelde examen, de gemiddelde examenvraag of de gemiddelde corrector.

Voorbeeld van een examenvraag met bijbehorend antwoordmodel

De Republiek in een tijd van vorsten

Toen Karel V in 1543 heer over alle Nederlanden werd, waren Vlaanderen en Brabant voor de uitvoering van zijn centralisatiebeleid de belangrijkste gewesten.

2) 1 Leg uit waarom dit zo was.

De Republiek in een tijd van vorsten

1 maximumscore 2

Kern van een juist antwoord is:
Vlaanderen en Brabant waren de rijkste gewesten / gewesten met de meeste inwoners van de Nederlanden. Daardoor konden zij het meeste bijdragen aan de schatkist (zodat Karel zijn centralisatiebeleid kon voortzetten/uitvoeren).

Opmerking

Als in het antwoord alleen verwezen wordt naar de keuze van zijn hoofdstad (Brussel/Brabant) wordt 1 scorepunt toegekend.



Een minder bewerkelijke en objectievere correctie is te realiseren

Verbeteringssuggesties

Kuhlemeier en anderen (2013) deden een groot aantal aanbevelingen die verschillen tussen correctoren en de verleiding tot te soepel beoordelen kunnen tegengaan. De belangrijkste aanbeveling voor de examenmakers was het verzamelen van antwoorden van kandidaten tijdens de constructiefase. Zo zou men de lang-antwoord vragen onder strikte geheimhoudingscondities aan een kleine groep kandidaten kunnen voorleggen. Door de antwoorden vervolgens na te kijken krijgt de examenmaker kwalitatieve informatie ter verdere verbetering van de vraagstelling en het antwoordmodel. Een aanvullend middel is het uitvoeren van interviews met kandidaten volgens de zogeheten 'cognitive labs' methode. De kandidaat wordt daarbij gevraagd zijn of haar gedachten bij het beantwoorden van de examenvraag onder woorden te brengen. Hiermee kan de examenmaker nagaan in hoeverre kandidaten de vraag interpreteren en het antwoord formuleren zoals hij of zij dat voor ogen had en de opgave zo nodig aanpassen. Een tweede aanbeveling gaat over het gebruik van principieel onvolledige antwoordmodellen zoals 'Voorbeeld van een juist antwoord is,' 'Een juist antwoord moet de volgende strekking hebben' en 'De kern van een juist antwoord is' (zie de bijgevoegde voorbeeldvraag). Uiteraard is het zelfstandig kunnen formuleren van een antwoord een essentieel onderdeel van veel vakken. Waar principieel onvolledige antwoordmodellen echter de boventoon voeren, zou men kunnen overwegen om een deel van de lang-antwoordvragen geleidelijk aan te vervangen, bijvoorbeeld door objectief scorebare voorgestructureerde vragen of door vraagtypen waarbij een korter antwoord volstaat.

Wat wil de Tweede Kamer?

Een samenvatting van het onderzoek (Kuhlemeier & Kremers, 2013c) is in 2013 aan de Tweede Kamer voorgelegd. De vaste commissie voor OCW vroeg de staatssecretaris wanneer hij aan de slag zou gaan met de aanbevelingen die snel en gemakkelijk zijn uit te voeren, zoals de door docenten gewenste verduidelijking van de correctievoorschriften. In zijn antwoord gaf de staatssecretaris aan dat CvE en Cito deze verduidelijking al ter hand hebben genomen. Hieronder geven we een indruk van de reeds genomen maatregelen. Daarnaast doen we verslag van enkele maatregelen die op dit moment nog

onderwerp van onderzoek zijn. We beperken ons hierbij tot de centraal schriftelijke examens voor de kunst- en maatschappijvakken.

Reeds genomen verbeteringsmaatregelen

Binnen het cluster kunst- en maatschappijvakken is een begin gemaakt met het implementeren van de aanbevelingen van het P4C-onderzoek. De resultaten ervan zijn niet alleen besproken binnen de constructieteams en met de CvE-vaststellingscommissies, maar zijn ook vakoverstijgend aan de orde gesteld via lezingen en workshops. Procedures en richtlijnen zijn aangescherpt en gedocumenteerd in handboeken voor examenconstructie. Een procedure voor het uitvoeren van kwalitatieve try-outs van examenvragen tijdens de constructiefase is in een vergevorderd stadium.

De geleidelijke veranderingen in de examens zelf worden zichtbaar als we de examens geschiedenis voor havo en vwo van 2012 en 2013 met elkaar vergelijken. In beide examenjaren ligt het zwaartepunt nog steeds op antwoordmodellen van het type 'Voorbeeld van een juist antwoord is,' maar het aandeel daarvan is in 2013 wel iets kleiner dan in 2012 (59% versus 62%) en zal de komende jaren nog verder afnemen. Tegelijkertijd is het aandeel van het type antwoordmodel 'Kern van een juist antwoord' gestegen van 10 procent naar 28 procent. Dit is een verbetering omdat de examenmakers de essentie van het juiste antwoord met dit type antwoordmodel op een duidelijke en gestroomlijnde manier voor het voetlicht kunnen brengen. Dat verdere verbetering mogelijk is, blijkt hieruit dat het examen nog steeds voor slechts 7 procent uit voorgestructureerde vragen bestaat (denk bijvoorbeeld aan de zogeheten chronologische volgordevraag).

Verdere acties de komende jaren

Cito doet ook de komende jaren onderzoek naar een minder belastende en objectievere correctie (zonder daarbij de validiteit uit het oog te verliezen). Voordat we een overzicht van recent uitgevoerd onderzoek geven, plaatsen we eerst een relativerende kanttekening. Bewezen effectieve maatregelen leiden namelijk niet direct en ook niet per definitie tot aanpassing van de examens. Zo kunnen de onderzoeksresultaten van 2013 pas op zijn vroegst invloed hebben op de examens van 2015. De productie van de examens van 2015 is namelijk al in volle

Door uitbreiding van het antwoordmodel worden interpretatieverschillen verkleind



gang. Daarnaast is het implementeren van onderzoeksresultaten in de examens de gezamenlijke verantwoordelijkheid van CvE en Cito. Tot slot noemen we de vaktraditie in de mens-, maatschappij- en kunstvakken. Van kandidaten in deze vakken wordt verwacht dat zij een eigen, op niveau geformuleerd antwoord op een open vraag kunnen geven. Objectief scorebare vragen zijn dan slechts beperkt toepasbaar.

Aanvullend onderzoek

Naar aanleiding van de aanbevelingen uit het P4C-onderzoek is binnen het cluster kunst- en maatschappijvakken voor drie examens aanvullende onderzoek uitgevoerd: kunst havo (Gitsels & Zwitser, 2013), filosofie h/v (Van Dorp & Wessel, 2013) en muziek, drama en dans g/t (Denissen & Schoenmakers, 2013). Deze vakken kennen relatief veel open vragen en mede daardoor een hoge correctielast per kandidaat. In het onderzoek naar het examen kunst havo hebben de onderzoekers een aantal open vragen omgewerkt tot gesloten vragen. De onderzoeksuitkomsten laten zien dat het mogelijk is om een deel van de open vragen te vervangen door gesloten vragen zonder dat het examen te gemakkelijk of minder betrouwbaar wordt. Een minder bewerkelijke en objectievere correctie lijkt bij kunst havo dus vrij eenvoudig te realiseren.

Ook in het onderzoek bij het examens filosofie h/v en muziek, dans en drama g/t lag het zwaartepunt op de beoordeling van lang-antwoord vragen. De resultaten doen vermoeden dat uitbreiding van het antwoordmodel een probaat middel is om interpretatieverschillen te verkleinen. De uitgebreidere antwoordmodellen 'nieuwe stijl' bevatten aanvullende aanwijzingen voor het aantal punten dat de corrector mag toekennen aan veel voorkomende antwoorden die geheel of gedeeltelijk fout zijn. Daarnaast bevatten de uitgebreidere antwoordmodellen voorbeelden van zowel volledig juiste als gedeeltelijk juiste en foute antwoorden. Ook zijn waar dat zinvol is per vraag 'opmerkingen' opgenomen die correctoren extra steun geven. Vaak verwijzen deze opmerkingen overigens naar algemene regels die nogal eens over het hoofd worden gezien.

Andere vaardigheden

De laatste decennia is er veel veranderd in de doelstellingen van het voortgezet onderwijs. Het gaat steeds meer

om het kunnen toepassen van kennis in nieuwe situaties, om het zelf ontwikkelen van weer andere kennis en om het kunnen reflecteren op de vergaarde kennis. De komende periode gaan we ook onderzoeken hoe deze vaardigheden op een efficiënte, objectieve, betrouwbare en valide manier in het CSE geëxamineerd kunnen worden.

Literatuur

- Denissen, H. & Schoenmakers, P. (2013). *De eerste en de onafhankelijke tweede correctie bij muziek, drama en dans G/T in 2011 en 2012* (ongepubliceerd rapport). Arnhem: Cito.
- Dorp, F. van & Wessels, H. (2013). *Onderzoek kwaliteit correctievoorschrift examen filosofie h/v* (ongepubliceerd rapport). Arnhem: Cito.
- Gitsels, H. & Zwitser, R. (2013). *Verslag van het onderzoek naar voorgestructureerde en gesloten vragen in het examen kunst*. Arnhem: Cito.
- Kuhlemeier, H., Gitsels, H., Boom, S., e.a. (2013). *Examenkenmerken en verschillen tussen correctoren. Verslag van een panelonderzoek bij het CSE geschiedenis, tehatex en Nederlands*. Arnhem: Cito.
- Kuhlemeier, H. & Kremers, E. (2013a). Eerste, tweede, derde en vierde correctie: Wat is het verschil? Correctie van examenwerk. *Examens, tijdschrift voor toetspraktijk*, 10 (3), 6-10.
- Kuhlemeier, H. & Kremers, E. (2013b). De uitvoering van de tweede correctie. Nakijken, nakijken en nog eens nakijken. *Examens, tijdschrift voor toetspraktijk*, 11 (1), 11-16.
- Kuhlemeier, H. & Kremers, E. (2013c). *De praktijk van de eerste en tweede correctie. Samenvatting van onderzoek naar het functioneren van het CSE*. Arnhem: Cito.
- Kuhlemeier, H., Rijn, P. van & Kremers, E. (2013). *Eerste, tweede en derde correctie van examenwerk: Wat is het verschil?* Arnhem: Cito.

□ Dr. H. Kuhlemeier is als onderzoeker werkzaam bij Cito; mw. drs. E. Kuijs-Pieters is clustermanager kunst- en maatschappijvakken; drs. S. Boom is toetsdeskundige geschiedenis voor havo en vwo; H. Gitsels, MA is toetsdeskundige voor de kunstvakken; drs. E.J.J. Kremers is sectormanager centrale examens havo/vwo en Coördinator toetsing Doorlopende Leerlijnen Taal en Rekenen. E-mail: hans.kuhlemeier@cito.nl.