Vraag- en antwoorddocument webinar over normering 9 maart 2017,
bijlage bij mail aan deelnemers 4-4-2017

College voor Toetsen en Examens

Inhoud

1 Referentie-examen en de prestatie-eis

Er zijn veel vragen binnengekomen over het **referentie-examen** en de referentienorm (**prestatie-eis**). Denk hierbij aan vragen als:
- *Hoe en door wie wordt een referentie-examen gemaakt?*
- *Zijn referentie-examens in te zien?*
- *Hoe groot is de groep leerlingen die het referentie-examen maakt?*
- *Wat is de reden dat het gemiddelde cijfer rond de 6 moet zijn?*
- *Wie bepaalt wat het 'juiste' percentage onvoldoendes is?*

In de hierna volgende beschrijving proberen we de vele vragen op deelaspecten te beantwoorden. Voor het begrip volgt hier eerst een vraag: Welk eerste tijdvak examen uit de laatste 5 jaar vond u het 'mooist'? Dan zou dat examen 'referentie-examen' genoemd kunnen worden. Een referentie-examen is dus een oud eerder afgenomen examen.

Eind vorige eeuw is men begonnen met het aanwijzen van referentie-examens. Dit gebeurde toen nog niet voor alle vakken. Een referentie-examen is een eerder afgenomen eerste tijdvak examen dat 'goed is ontvangen'. Dit wordt (mede) afgelezen uit de waardering die docenten geven in de zogenaamde quickscan, de vier vragen die zij over het examen invullen nadat zij de scores in Wolf hebben ingevuld. Een ander criterium waaraan een referentie-examen moet voldoen, is dat het examen geen fouten mag bevatten, het moet naar het oordeel van vakinhoudelijk experts een 'mooi' examen zijn waarin de verdeling over de stof goed is en de uitwerking van de vakinhoud, zoals beschreven in de syllabus, goed aansluit op het gegeven onderwijs. Last but not least moet de moeilijkheid van het gekozen referentie-examen passend zijn. Dat betekent dat het niet te moeilijk en niet te makkelijk mag zijn. De afnamegegevens van het betreffende examen (de N-term, het gemiddelde cijfer en het percentage onvoldoendes) zijn onlosmakelijk met het referentie-examen verbonden. Met het referentie-examen is dus ook de prestatie-eis voor dat vak vastgelegd in een % onvoldoendes en gemiddeld cijfer.

Rond de eeuwwisseling ontstond de notie dat bij alle vakken een referentie-examen gewenst was. In 2004 is dat voor het eerst in de volle breedte gelukt. Vanaf dat moment bestaat er dus voor elk vak een referentie-examen waarin de prestatie-eis is vastgelegd. Voor pilot-examens wordt geen pilot-referentie-examen aangewezen. Deze examens volgen de prestatie-eis van de reguliere examens. Bijvoorbeeld voor het pilot-examen maatschappijwetenschappen vwo worden de gegevens gebruikt van het referentie-examen maatschappijwetenschappen vwo regulier.

Er zijn veel redenen om het referentie-examen te willen actualiseren. Op het moment dat er een ingrijpende wijziging van het examenprogramma is, moet gezocht worden naar een referentie-examen binnen dat nieuwe programma. Het is gebruikelijk dat het eerste en het tweede jaar eerst wordt aangekeken hoe de examens zich ontwikkelen. In deze jaren worden de gegevens van het referentie-examen (oud programma) gehanteerd.[1] Pas na een aantal jaar na invoering van een nieuw programma wordt bekeken of er een examen geschikt wordt bevonden om als referentie-examen dienst te doen.

Een andere manier om voor een vak met een (gedeeltelijk) nieuw programma of voor een geheel nieuw vak de prestatie-eis te bepalen is door een standaardsetting uit te voeren. Er zijn verschillende manieren om dat te doen, maar wat de verschillende methoden gemeenschappelijk hebben, is dat een groep experts, meestal zijn dat docenten die examenklassen lesgeven in het betreffende vak, bepaalt waar de cesuur zou moeten liggen. Zo is in 2016 een standaardsetting uitgevoerd bij vwo natuurkunde, scheikunde en biologie omdat daar een nieuw examenprogramma werd ingevoerd. Er is toen gebruik

[1] Er wordt dan alleen gekeken naar het % onvoldoendes of het gemiddeld cijfer en niet naar de inhoud, aangezien de inhoud gewijzigd en dus niet meer representatief is. Wel is de populatie die een examen volgens het oude programma gemaakt heeft vergelijkbaar met de populatie die het examen volgens het nieuwe programma gemaakt heeft. Op deze manier wordt de prestatie-eis van het oude programma overgezet op het nieuwe programma.

gemaakt van een nieuw type standaardsetting, namelijk de 3DC standaardsetting. Bij deze vorm van standaardsetting wordt gebruik gemaakt van de afnamegegevens. De docenten die deze standaardsetting uitvoeren, worden geholpen bij het zetten van de standaard doordat voor hen inzichtelijk gemaakt wordt wat de consequenties zijn van de keuzes die zij maken. De uitslag van deze standaardsetting helpt dan bij het vaststellen van de meest passende N-term voor dat specifieke examen, maar daarmee ook met het stellen van de prestatie-eis (het juiste % onvoldoendes) voor toekomstige examens.

Een referentie-examen dat meer dan 7 jaar geleden is afgenomen, krijgt het stempel 'belegen' en het verdient aanbeveling om een nieuwer examen te kiezen. Het referentie-examen wordt om deze reden om de zoveel tijd geactualiseerd. Omdat de prestatie-eis gehandhaafd is, geldt voor dit nieuwe referentie-examen dezelfde eis als voor het oude referentie-examen. Normaal gesproken betekent dat dat het % onvoldoendes en het gemiddeld cijfer ongeveer gelijk zijn. Alleen als er in die '7 jaar' een ontwikkeling heeft plaatsgevonden (zoals bv de waargenomen vaardigheidsstijging na de invoering van de aangescherpte uitslagregels) kan het zijn dat het nieuwe referentie-examen een ander % onvoldoendes en gemiddeld cijfer kent dan het oude. Daarmee is de prestatie-eis niet veranderd, maar is de populatie examenleerlingen voor dat vak vaardiger geworden.

Een referentie-examen heeft dus meestal niet een N-term van 1,0. Maar idealiter wijkt het daar ook niet te veel van af. Welk examen op dit moment dienst doet als referentie-examen is niet openbaar.

Waarom maken jullie gebruik van leerlingen als referentiegroep en niet van docenten? Dit om de mate van normvervaging (leerlingen die door de jaren minder talig worden en daardoor lager gaan scoren op taalgevoelige onderdelen) zoveel mogelijk te beperken (je gaat immers uit van de best scorende groep).
Het blijkt onmogelijk dat docenten (of andere experts) voorafgaand aan een examen een goede en nauwkeurige inschatting maken van de moeilijkheid van het examen. Het kan wel zijn dat in de huidige systematiek normvervaging optreedt. Deze zou eens in de, zeg 10 jaar, gemeten moeten worden. Daarbij moet wel bedacht worden dat exameneisen over zulke lange perioden ook niet altijd constant zijn. Het CvTE heeft de opdracht om de norm te handhaven en doet dat, met deze problematiek op het netvlies, zo goed als mogelijk.

Gelden de voorbeelden uit het webinar ook voor vmbo?
Ook op het vmbo worden referentie-examens gebruikt op exact dezelfde manier als op het havo/vwo.

Als het enkele jaren duurt voordat er een referentie-examen is, hoe krijg je dan een referentie-examen voor een vak als Filosofie met een programma dat om de vier jaar vrijwel volledig gewijzigd wordt?
Bij een vak als filosofie komt het relatief vaak voor dat de inhoud van het referentie-examen en het actuele examen verschillend zijn. De getallen die bij het referentie-examen horen (percentage onvoldoendes) kunnen echter wel gebruikt worden.

Hoe weet je dat de populatie die het referentie-examen destijds maakte een gemiddelde populatie was en geen `sterke' of 'zwakke' lichting?
Tot 2011 veronderstellen we dat er geen wijzigingen zijn op nationaal niveau met betrekking tot het vaardigheidsniveau. Een sterke of zwakke lichting kan op schoolniveau wel voorkomen, op nationaal niveau middelt dit uit. Na 2011 hebben we de generieke vaardigheidsstijging in beeld. Zie de notitie op Examenblad.nl: https://www.examenblad.nl/document/de-normeringssystematiek-van-de-ce/2017/f=/De_normeringssystematiek_van_de_CEs_vo_30_november_2016.pdf

Is het met dit systeem ook mogelijk dat alle leerlingen een voldoende scoren?
In theorie is het met dit systeem wel mogelijk dat alle leerlingen in Nederland een voldoende scoren op een examen. Maar in de praktijk gebeurt dit niet. Leerlingen die voldoen aan de eisen verdienen een voldoende. Dat zou het geval zijn als alle leerlingen aan de exameneisen voldoen. In de praktijk komt dat echter nooit voor. Wel zien het % onvoldoendes dalen als de populatie vaardiger is geworden dan in voorgaande jaren. Dit hebben we bv. gezien bij Engels vwo in 2013 waar het % onvoldoendes zakte tot 12,6%. Voor wiskunde B vwo pilot was het % onvoldoendes in dat jaar zelfs 8,4%. Dit is overigens al uitzonderlijk laag. De kans dat dit 0% wordt is bijzonder klein.

2 Verschillen tussen vakken

Vele vragen voorafgaand aan het webinar, maar ook nog daarna, gingen over **verschillen tussen vakken**. Denk hierbij aan vragen als:
- *Wordt de N-term van verschillende vakken ook wel eens naast elkaar gelegd?*
- *Waarom wordt de N-term voor het vak Duits niet opgehoogd?*
- *Wat moeten vakken met een laag gemiddeld cijfer doen om toch een 6,5 gemiddeld te halen?*
- *Waarom krijgen niet alle vakken hetzelfde gemiddeld cijfer?*

Allereerst is het belangrijk om een onderscheid te maken tussen de N-term en het gemiddeld cijfer (of % onvoldoendes). Het vergelijken van N-termen tussen vakken heeft helemaal geen zin. Een N-term is een instrument om te compenseren voor verschillen in moeilijkheid van examens *van hetzelfde vak over jaren heen*. Een hoge N-term (>1,5) betekent dus dat een examen relatief moeilijk was en een lage N-term (<0,5) betekent dat een examen relatief makkelijk was. Als een vak vaak een erg lage N-term of juist een erg hoge N-term heeft, is er maar één manier om dat op te lossen en dat is door de examens voor dat vak moeilijker respectievelijk makkelijker te maken. Doordat we de lat gelijk moeten houden, zal de N-term dan automatisch omhoog respectievelijk omlaag gaan. Het gemiddeld cijfer (of % onvoldoendes) blijft dan gewoon hetzelfde.

Verschillen in gemiddeld cijfer (of % onvoldoendes) tussen vakken is een ander verhaal. Deze verschillen zijn ontstaan door de keuze van het referentie-examen en eventuele ontwikkelingen in het verleden.

Binnen de huidige context is er maar één manier waarop een vak een hoger gemiddeld cijfer (en lager % onvoldoendes) kan krijgen en dat is doordat de populatie examenkandidaten vaardiger wordt. Dat kan bijvoorbeeld doordat ze met z'n allen harder gaan studeren dan de leerlingen in eerdere jaren of dat zij op een andere manier nog beter worden voorbereid op hun centraal examen. Het kan ook door op de school strengere selectie toe te passen waardoor de zwakkere kandidaten het vak niet meer kiezen en alleen de sterkere kandidaten overblijven. De populatie als geheel wordt dan vaardiger. Dergelijke maatregelen kunnen niet zomaar genomen worden en zeker niet op landelijke schaal. Dit is dus niet zo realistisch, maar wel de enige mogelijkheid om een hoger gemiddeld cijfer te krijgen voor een vak zonder de prestatie-eis aan te passen.

Als we voor een bepaald vak het gemiddeld cijfer (of % onvoldoendes) willen aanpassen, betekent dat een aanpassing van de prestatie-eis. Het CvTE mag dat niet doen. Alleen de politiek kan daarover beslissen. Wilt u dus dat voor uw vak de prestatie-eis wordt bijgesteld, dan dient u daarvoor de politiek te benaderen (al dan niet via uw vakvereniging).

Als CvTE zijn wij ons overigens wel bewust van de verschillen tussen vakken en de mogelijke ongewenste effecten die dat kan hebben op bijvoorbeeld de vakkenkeuze van leerlingen. Op dit moment zijn wij ons aan het beraden op de wenselijkheid om de gemiddelde cijfers (of % onvoldoendes) over vakken heen (meer) gelijk te trekken. De vraag is dan natuurlijk wat wenselijke/redelijke gemiddelde cijfers zijn voor vakken. En zijn verschillen tot op zekere

hoogte wel gerechtvaardigd of juist niet? En welke vakken zouden dan een hoger gemiddeld cijfer moeten krijgen? U kunt zich voorstellen dat dit een discussie is die niet zomaar beslecht is. Hierbij zal het veld gehoord moeten worden en zal de politiek uiteindelijk een beslissing moeten nemen.

3 Wolf-scores

Er zijn diverse vragen gesteld over de Wolf-scores. Denk daarbij aan de volgende vragen:
- *Welke rol spelen de WOLF-scores bij het bepalen van de N-term?*
- *Door de 2e correctie verandert de score nog. Dit wordt niet meegenomen in WOLF. Hoe wordt hiermee omgegaan?*
- *Worden bij de WOLF-scores alleen de eerste 5 leerlingen meegenomen.*

De Wolf-scores spelen een belangrijke rol bij het bepalen van de N-term. Op basis van de Wolf-scores verkrijgen wij een beeld van de behaalde scores. Zo weten wij bij elke mogelijke N-term wat het gemiddelde cijfer en percentage onvoldoendes zal zijn. Op basis van de Wolf-scores wordt een technisch normeringsadvies bepaald. Dat technisch advies is gebaseerd op één of meer specifieke aannames. Bijvoorbeeld dat de nieuwe populatie even vaardig is als de referentiepopulatie.

De N-termen zijn dus gebaseerd op de door de examinator ingezonden scores. Cito onderzoekt op gezette tijden de representativiteit van deze gegevens door ze te vergelijken met de door de scholen aan DUO gerapporteerde cijfers. Er zullen ongetwijfeld aanzienlijke verschillen voor en na 2e correctie voorkomen, maar in algemene zin leveren de Wolf-scores een zeer representatief totaalbeeld op. Het is bij de controles nog nooit voorgekomen dat het verschil tussen de steekproef van Cito en de aan DUO gerapporteerde gegevens groter was dan 0,1 cijferpunt

Het aantal kandidaten dat wordt opgevraagd via Wolf, bedraagt voor de meeste examens 5 per groep/klas. Dat is het minimale aantal voor een verantwoorde normering. We ontvangen echter van meer kandidaten gegevens. Alle resultaten die beschikbaar zijn op het moment dat Cito aan de analyse van een examen begint, worden in de analyses opgenomen en dragen bij aan het totaalbeeld van hoe goed een examen gemaakt is.

4 De totstandkoming en de hoogte van de N-term

Voor de inleiding bij dit onderwerp verwijzen we graag naar het webinar, waar uitgebreid op dit onderwerp wordt ingegaan. Hieronder de verhelderende vragen die wij in het webinar niet konden beantwoorden.

Vraag	Antwoord
De systematiek van referentie-examens en referentievragen, die er voor zou moeten zorgen dat de prestatie-eis gelijk blijft. Die zou de N-term moeten bepalen, maar ik zie niet hoe dat werkt.	Uitgangspunt is dat het elk jaar even moeilijk moet zijn om een voldoende te halen. Als twee populaties even vaardig zijn, mag je ervan uitgaan dat in beide populaties procentueel evenveel leerlingen een onvoldoende halen. Dit gebeurt tijdens de eerste stap van het normeringsproces (bij vakken met voldoende leerlingen en zonder aanvullende dataverzameling). Er wordt gezorgd dat het percentage leerlingen dat op het referentie-examen een onvoldoendes haalde wordt 'overgebracht' naar het actuele examen. Met andere woorden: de hoogte van de N-term wordt zodanig bepaald dat het percentage van het actuele examen gelijk is aan dat van het referentie-examen.

Vraag	Antwoord
Welke vergelijkingen worden precies toegepast om tot een cijfer te komen met de N-term dicht bij het cijfer 1 en het cijfer 10?	De formule voor de hoofdrelatie is: C = 9 * (S/L) + N (met S de behaalde score en L de lengte van de scoreschaal). Bij een centraal examen met een maximum aantal scorepunten van 68 wordt dat: C = 9 * (S/68) + N Bij N > 1,0 geldt voor de laagste scores de formule: C = 1,0 + S* (9/68)*2 En voor de hoogste scores: C =10,0 – (68-S)* 9/68) * 0,5 Bij N < 1,0 geldt voor de laagste scores de formule: C = 1,0 + S* (9/68)* 0,5 en voor de hoogste scores C = 10,0 – (68-S)*(9/68)*2 Aan de formules is te zien dat de grafiek van een grensrelatie 2x zo stijl dan wel ½x zo stijl loopt als de grafiek van de hoofrelatie.
Hoe wordt er bepaald waar het cijfer wordt afgevlakt?	Dat gebeurt door van iedere grafiek van een grensrelatie het snijpunt te bepalen met de grafiek van de hoofdrelatie. De formules worden gegeven in het antwoord op de vorige vraag.
De N-term ligt dus in principe voor het examen al vast?	Bij een pretest en een posttest hebben we vooraf een schatting. Maar de feitelijke afnamegegevens (die Cito verkrijgt via Wolf), blijven nodig. Aan een pretest kun je bijvoorbeeld niet zien of sprake was van tijdnood. Op basis van de Wolf-gegevens kan dat wel.
Leidt het aanpassen van de N-term op den duur niet tot het verlagen van het eindniveau van opleidingen.	Bij ieder centraal examen wordt de N-term bepaald die de moeilijkheidsgraad van dat examen het best weerspiegelt. Het doel is om de prestatie-eisen aan leerlingen over de jaren heen gelijk te houden. Vroeger was het inderdaad zo dat we het niet konden zien als het niveau van leerlingen ongemerkt daalde. Doordat we in ons huidige systeem ook aanvullende gegevens hebben als pretest en posttest, is het mogelijk om ook zicht te krijgen op een eventuele verandering in vaardigheid van leerlingen. Daar kunnen we dan rekening mee houden om te voorkomen dat we mee zakken met een dalende vaardigheid. Het antwoord op de vraag is dus: juist niet.
Ik weet nog steeds niet hoe de N-term doorberekend wordt. De berekening in u Dia levert volgens mij een hoger cijfer op dan 1 bij 0 punten	Bij een score van 0 pas je altijd een van de volgende grensrelaties toe: C =1,0 + S* (9/L)*2 of C = 1,0 + S* (9/L)* 0,5 (met S de behaalde score en L de lengte van de scoreschaal) S = 0 → C = 1,0. Dus bij 0 scorepunten heb je altijd een 1,0.
Stel nu dat de N-term 1,0 zou zijn, heeft het CvTE dan al een idee hoe ze kunnen waarborgen dat de examens een standaard niveau hebben?	Twee examens zijn bijna nooit precies even moeilijk. Een jaar later kan het examen voor hetzelfde vak en schooltype moeilijker zijn en een N-term van 1,3 krijgen of makkelijker met een N-term van 0,4. De examens van opeenvolgende jaren verschillen dan in moeilijkheidsgraad, maar wat een leerling moet kennen en kunnen voor een 5,5 is van jaar tot jaar hetzelfde, daar zorgt de N-term voor.

Vraag	Antwoord
Waarom is er pas sinds 2000 (invoering/eerste afname van het Havo-examen) een N-term? Zou het kunnen zijn dat dit instrument is ontwikkeld om de politiek te beschermen tegen onwelgevallige uitslagen van eerste examens, en dat hierna voortschrijdend inzicht heeft geleid tot de mogelijkheid om hiermee de uitslag van een examen zodanig te sturen dat deze acceptabel is??	De N-term is ingevoerd om beter recht te kunnen doen aan het internationaal erkende ervaringsgegeven dat examenmakers er vrijwel nooit in slagen om een examen precies de vooraf beoogde moeilijkheidsgraad mee te geven. Iedere docent weet dit ook van zijn eigen schriftelijke overhoringen, proefwerken en schoolexamens. Het eerste blok uit het webinar laat zien dat we vóór 2000 minder goed recht deden aan dit principe. Voor een examen dat moeilijker was dan de vooraf beoogde moeilijkheidsgraad werd wel gecompenseerd. Voor een examen dat makkelijker was uitgepakt niet. Sinds de invoering van de N-term kan dat laatste ook.
Heeft de uitkomst van de examenvergadering van het NVON biologie invloed op de n-term?	Ja, in het artikel de normeringssystematiek van de centrale examens vo is beschreven dat de N-term in 3 stappen wordt bepaald. https://www.examenblad.nl/document/de-normeringssystematiek-van-de-ce/2017/f=/De_normeringssystematiek_van_de_CEs_vo_30_november_2016.pdf 1. Het technisch normeringsadvies van Cito; 2. Het advies van de vaststellingscommissie (vc) van het CvTE; 3. De vaststelling van de N-term door de CvTE-leiding. Bij stap 2 kijkt de vaststellingscommissie behalve naar stap 1 ook naar alle reacties die (via het Examenloket) zijn doorgestuurd naar de Examenlijn van het CvTE en ook naar de verslagen van de examenbesprekingen van de vakvereniging. In stap 2 worden behalve de kwantitatieve gegevens van stap 1 dus ook kwalitatieve reacties meegewogen.
Hoe komt het dat voor één vak elk jaar ongeveer dezelfde N-term wordt vastgesteld?	In dat geval verschilt de moeilijkheidsgraad van de examens over de jaren heen maar erg weinig. Als de N-term jaarlijks in de buurt ligt van de vooraf beoogde N-term (bij de meeste centrale examens is dat 1,0 of de N-term van het referentie-examen), dan slagen de examenmakers er goed in om hun examen ook daadwerkelijk de moeilijkheidsgraad mee te geven die vooraf beoogd was.
Gaat de vaststelling van de vakken met een klein aantal leerlingen op dezelfde manier? Het uitgangspunt is dan misschien niet meer juist.	Een deel van de CE's wordt hard geëquivaleerd (= heeft een pretest, posttest of anker-in-package). Bij de overige CE's wordt bij de bepaling van de N-term in eerste instantie gekeken naar het percentage onvoldoendes. Het aantal leerlingen dat het examen heeft afgelegd moet dan wel tenminste 200 zijn. Is het aantal leerlingen < 200, dan is het percentage onvoldoendes niet zo geschikt. Bij de normering wordt gewerkt met normeringstabellen waarin je kunt zien wat bij alle N-termen tussen 0,0 en 2,0 het percentage onvoldoendes en het gemiddeld cijfer is (zie ook dia 17 in het webinar). Bij 'kleine vakken' (vakken met < 200 leerlingen) kan het percentage onvoldoendes bij N=0,8 en bij N=0,9 sterke verschillen vertonen. Statistisch is het dan beter om bij de bepaling van de N-term uit te gaan van het gemiddeld cijfer.
Ik weet nog steeds niet hoe het berekend wordt, de N term dan	Stel dat we aannemen dat de leerlingen van jaar tot jaar (ongeveer) even vaardig zijn, dan ligt het percentage onvoldoendes van jaar tot jaar ook op (ongeveer) hetzelfde niveau.

Vraag	Antwoord
	Bij de normering wordt gewerkt met normeringstabellen (zie webinar dia 17) waarin je kunt zien, wat bij alle N-termen tussen 0,0 en 2,0 het percentage onvoldoendes (en het gemiddeld cijfer) is. Deze tabel is door Cito samengesteld op basis van de afnamegegevens (die via Wolf zijn verzameld). Voor het CE van dat jaar kiezen we dan de N-term waarbij het percentage onvoldoendes het dichtst ligt bij dat van de voorgaande jaren. Dit werkt goed als er geen vaardigheidsverschillen met eerdere examenjaren geconstateerd worden en het CE is afgelegd door minstens 200 leerlingen. Bij < 200 leerlingen: zie de vorige vraag. De N-term wordt dus niet berekend, maar uit een tabel gehaald op basis van vooraf gestelde uitgangspunten.
Is het niet 'handiger' om standaard de gemiddelde score (berekend via Wolf) een cijfer (bijvoorbeeld) 6.0 toe te kennen en op basis hiervan de N-term vast te stellen?	Dat is een fundamentele vraag. Alleen gaat het niet zo zeer om handiger maar om de vraag of je niet beter kunt ijken aan het gemiddeld cijfer dan aan het percentage onvoldoende. Voor beiden is wat te zeggen. Deze vraag is momenteel onderwerp van studie in het zogeheten normeringsoverleg tussen het CvTE en Cito. De vraag is een onderdeel van de systematische evaluatie die we toepassen: 'Doen we het goed of kan het beter en zo ja, wat zijn daarvan dan de consequenties?'
we kijken dus alleen naar het aantal onvoldoendes, stel er wordt bij een referentie-examen 20% onvoldoende gescoord en bij het echte examen hebben we bij een N term van 1 geen onvoldoendes dan moeten we dus naar beneden bij stellen om het referentie niveau te komen.	Klopt. Alleen willen wij dit geen bijstellen noemen. Bijstellen zou betekenen dat er iets mis is. We willen de N-term die het best past bij de (achteraf geconstateerde) moeilijkheidsgraad van het examen. Als bij een N-term van 1,0 het percentage onvoldoendes nul is, is het examen waarschijnlijk (veel) makkelijker uitgepakt dan vooraf was beoogd. Dat kan nu eenmaal gebeuren.
Is dat niet een cirkelredenering? We passen de N-term aan en we zien dat de verschillen minder fluctueren!	We passen geen N-termen aan. We stellen ze achteraf vast op basis van de moeilijkheidsgraad die daadwerkelijk is geconstateerd via de normeringstechnieken van Cito. Het ligt voor de hand dat een grote groep examenkandidaten zoals bij het centraal examen van jaar tot jaar even vaardig is en dus ook een zelfde gemiddeld cijfer haalt. Dat dat in de jaren '80 en daarvoor niet het geval was, blijkt te wijten te zijn aan het feit dat het onmogelijk is om ieder jaar examens van gelijke moeilijkheid te maken. Was het examen moeilijk, dan haalde de populatie een gemiddeld cijfer dat lager was dan daarvoor. Is dat terecht? Wij denken van niet. Door de N-term in te voeren konden we wel compenseren voor die verschillen in moeilijkheid. Hierdoor kreeg een moeilijk examen een hogere N-term en bleef het gemiddeld cijfer wel ongeveer gelijk aan het jaar ervoor. Logisch als er geen substantiële wijzigingen zijn in de exameneisen. In de jaren na 2012 zagen we juist wel weer fluctuerende percentages onvoldoendes en gemiddelde cijfers a.g.v. de aangescherpte uitslagregels. Als we geen N-term hadden gehad, was dat ook niet gebeurd en zouden we de leerlingen naar ons oordeel benadeeld hebben.

Vraag	Antwoord
	Het is mogelijk om de N-term zo te bepalen dat er jaarlijks hetzelfde gemiddeld cijfer uitkomt. Of een nauwelijks schommelend percentage onvoldoendes. Maar de crux is dat we dat nu juist niet doen, om recht te kunnen doen aan het principe dat als leerlingen beter presteren het gemiddeld cijfer hoort te stijgen.
Waarom kan de N term niet bv 0,15 zijn zodat de percentages onvoldoende dichter bij elkaar liggen?	De cijfers voor het CE kunnen tussen 1 en 10 liggen en worden afgerond op één decimaal. Zo heeft de overheid dat vastgelegd in het Eindexamenbesluit VO. Met een N-term van 0,15 zou een leerling ook het cijfer 5,95 kunnen behalen. In principe verzet het systeem van de N-termen zich daar niet tegen, maar een cijfer van 5,95 suggereert wel een heel grote meetnauwkeurigheid. Nu al kun je je afvragen of een leerling die een 6,0 behaalt wel vaardiger is dan zijn klasgenoot met een 5,9. Die vraag is niet te beantwoorden voor de leerling met een 5,95 t.o.v. zijn klasgenoot met een 5,96. Het antwoord op de vraag is: We hebben er in Nederland niet voor gekozen om te werken met CE-cijfers met meer dan één decimaal.
Staat, voor het maken van het examen, de N-term altijd op 1? En is er daarbij verschil tussen vakken?	Bij de meeste CE's wordt vooraf ingezet op een examen dat lijkt op het referentie-examen omdat dat gezien wordt als een goede operationalisering van de syllabus. Daarmee kun je de N-term van het referentie-examen zien als het richtpunt. Dat zal vaak niet exact 1,0 zijn. Wel wordt 1,0 gezien als een soort ideale N-term. Maar bij sommige CE's ligt dat anders. Bij CE's waar veel met meerkeuzevragen gewerkt wordt, is de N-term waarop wordt gemikt lager, omdat de gokkans daar een rol speelt. Bij de centraal schriftelijke en praktische examens (cspe's) is de N-term veelal ook lager. Dat komt doordat de beoordelingsaspecten van praktische opdrachten anders uitpakken dan bij vragen uit een schriftelijk examen.
welke invloed hebben de resultaten en zelfs eventuele fouten van de voorgaande jaren nog op de N-term?	In principe wordt een N-term alleen vastgesteld op basis van de beschikbare gegevens in dat examenjaar. Wel is het zo dat de keuze van het referentie-examen wordt bepaald door de resultaten van voorgaande jaren. Indirect hebben de resultaten van voorgaande jaren, en dan vooral die van het referentiejaar, dus wel invloed op nog vast te stellen N-termen. Dat is ook logisch, omdat we de prestatie-eis over de jaren heen gelijk moeten houden. En dat kan alleen als je het heden met het verleden vergelijkt. Bij de keuze voor een referentie-examen zorgen we er overigens voor dat we een jaar kiezen waarin geen fouten in het examen voorkwamen, om te voorkomen dat die nog blijven doorwerken in volgende jaren.
hoe kun je met de N-term behalve het gemiddelde ook tegelijk de breedte van de verdeling corrigeren ?	Net zoals de moeilijkheidsgraad kan verschillen, geldt dat ook voor de spreiding van resultaten van de leerlingen. De N-term kan wel compenseren voor verschillen in moeilijkheidsgraad, maar niet voor verschillen in spreiding. Als we bij de centrale examens zouden werken met vaardigheidsscores (zoals bij de rekentoets gebeurt), dan kan wel voor verschillen in spreiding worden gecompenseerd.
Is er sprake van absolute of relatieve normering?	Bij vakken waarbij we de beschikking hebben over aanvullende gegevens zoals een pretest of een posttest is sprake van absolute normering. Bij vakken waar we die gegevens niet hebben en we als uitgangspunt hanteren dat we het percentage onvoldoendes gelijk houden, is in beginsel sprake van relatieve normering.

Vraag	Antwoord
	Door het toepassen van de Fishermethode (blok 5 in het webinar) kunnen we op basis van vakken met aanvullende gegevens ook uitspraken doen over de vaardigheid van kandidaten op vakken waar we die gegevens niet hebben. Daarmee is ook bij vakken waar in beginsel sprake was van relatieve normering toch een absolute normering toegepast. Zie hiervoor https://www.examenblad.nl/document/de-normeringssystematiek-van-de-ce/2017/f=/De_normeringssystematiek_van_de_CEs_vo_30_november_2016.pdf
Waarom wordt er gekeken naar het % onvoldoende en niet naar het gemiddelde cijfer?	Dit is een keuze. Normaal gesproken blijven ze allebei gelijk over de jaren heen, omdat de vaardigheid van examenpopulatie niet verandert over de jaren heen (tenzij er sprake is van een substantiële verandering in exameneisen). Voor de continuïteit is het goed om dan van één van beide uit te gaan. Omdat men in het verleden veel waarde hechtte aan de grens tussen voldoende en onvoldoende is het % onvoldoendes als ijkpunt genomen. Je zou er ook voor kunnen kiezen om het gemiddeld cijfer als grens te nemen, maar bijvoorbeeld ook de grens tussen een 7 en een 8 (7,5).

Meerdere vragen gaan over examens waarvoor een N-term dichtbij, gelijk aan of hoger dan 2,0 is vastgesteld of juist rond 0,0. Kun je dan conclusies trekken over de kwaliteit van dat examen? Welke stappen neemt het CvTE dan voor de komende examens?
Een N-term > 2,0 betekent in het algemeen dat het CE echt te moeilijk is geweest. Aan één kwaliteitscriterium is dan niet voldaan. Andere criteria zijn onder meer: goede spreiding van de scores, juiste lengte, goede verdeling van makkelijke, gemiddelde en moeilijke opgaven, geen vragen die door goede leerlingen fout en door zwakke leerlingen goed gemaakt zijn, goede dekking van de leerstof, foutloos. Een te moeilijk of te makkelijk examen is dus niet zonder meer een slecht examen.

Het risico bestaat wel dat leerlingen in paniek raken door een te moeilijk examen. Bij het vaststellen van de N-term houden we daar wel rekening mee. Als docent kun je je leerlingen hier ook op voorbereiden: Als je tijdens je examen merkt dat je er minder goed uitkomt dan bij de examens waarmee je geoefend hebt, denk dan niet 'O, ik heb een black out'. Ga gewoon door, haal er uit wat er in zit. Jouw examen van dit jaar kan gewoon pittiger zijn. De kans is groot dat dat examen dan een hogere N-term krijgt en zo haal je toch een vergelijkbaar cijfer als voor je oefenexamens. En omgekeerd natuurlijk: Een examen kan ook makkelijker zijn dan de examens waarmee je hebt geoefend.

Als een examen te moeilijk of te makkelijk is gebleken, bekijken de vc en de td wat zij kunnen doen om te voorkomen dat het CE van de volgende jaren opnieuw een N-term > 2,0 krijgt. De toets- en itemanalyse, die Cito op basis van de afnamegegevens (verzameld via Wolf) heeft gemaakt, is hen daarbij behulpzaam. Als de vc en de td de vragen met die kennis nog eens tegen het licht houden, kunnen zij meestal goed zien welke vragen hebben bijgedragen aan de te hoge moeilijkheidsgraad.
Deze benadering wordt ook toegepast als een examen echt te makkelijk is geweest. Dat examen had dan eigenlijk een N-term < 0,0 moeten krijgen, maar dat doen we niet. De leerlingen hebben dan 'mazzel'.

5 Normering op basis van aanvullende gegevens

Tijdens het webinar is gesproken over normering op basis van **aanvullende gegevens**. Hierover zijn diverse vragen binnen gekomen waaruit bleek dat niet volledig duidelijk was wat bedoeld werd met aanvullende gegevens en onder welke groepen leerlingen deze gegevens verzameld worden. In onderstaande toelichting gaan we op deze vragen nader in.

De standaardgegevens worden gevormd door de scores van de leerlingen op het examen. Aanvullende gegevens worden buiten de afnameperiode om verzameld. Het gaat daarbij vrijwel altijd om informatie over de relatieve moeilijkheid van de vragen. Deze informatie is nodig om te kunnen duiden of een hoge score moet worden toegeschreven aan een grote vaardigheid van de leerling of aan het feit dat de vragen gemakkelijk waren.

Wij gebruiken hier vier methoden voor:
- pretest, posttest, anker-in-package. Het idee hierbij is dat je leerlingen vragen uit twee verschillende examens laat maken. Daarmee kun je de vaardigheid van de leerling verdisconteren en informatie krijgen over de relatieve moeilijkheid van vragen.
- standaardsetting. Hierbij wordt informatie over de relatieve moeilijkheid verkregen door experts (over het algemeen docenten die lesgeven aan examenklassen in het betreffende vak) vragen uit verschillende examens met elkaar te laten vergelijken.

Anker-in-package wordt toegepast bij digitale examens. Daarin worden oude en nieuwe opgaven in één examen afgenomen. Hierdoor kan de moeilijkheidsgraad van de nieuwe opgaven worden vergeleken met die van oude opgaven. Groot voordeel van deze methoden is dat beide opgaven onder dezelfde omstandigheden en door dezelfde kandidaten worden gemaakt.

Bij papieren examens (die openbaar zijn) is dat niet mogelijk. Daarom gebruiken we daar pretests en posttests. Deze worden afgenomen bij groepen leerlingen die zodanig onderwijs hebben genoten dat zij in staat zijn om de examenvragen te kunnen maken. Zo kan bij bv. Engels een posttest worden afgenomen bij leerlingen in 5 vwo. Uiteraard zijn zij gemiddeld nog wat minder vaardig dan leerlingen in 6 vwo, maar omdat de methode wordt gebruikt om de relatieve moeilijkheid van de vragen vast te stellen is dat verschil in vaardigheid niet relevant. Voordeel van deze methode is dat de posttest kan worden afgenomen na afname van het feitelijke examen en de geheimhouding daarmee dus niet in het geding is.
Bij bv natuurkunde is het niet mogelijk om alle examenvragen te laten maken door leerlingen in 5 vwo, omdat dan nog niet alle examenstof behandeld is. In dat geval is alleen een pretest mogelijk. Deze wordt 2 jaar van te voren afgenomen bij leerlingen in 6 vwo. Voordeel van deze methode is dat het leerlingen betreft van hetzelfde niveau als de uiteindelijke examenkandidaten. Nadeel is uiteraard dat er risico is op schending van geheimhouding. Overigens weten leerlingen niet dat zij opgaven uit toekomstige examens maken.

Feitelijk krijgen leerlingen bij een pretest en een posttest niet de vragen uit het referentie-examen voorgelegd, aangezien het referentie-examen een openbaar examen is en zij dus voorkennis zouden kunnen hebben over deze vragen in tegenstelling tot de vragen uit het nieuwe examen. Om die reden is er een zogenaamd anker ontwikkeld. Jaren geleden, toen het referentie-examen in ontwikkeling was (toen was nog niet bekend dat het een referentie-examen ging worden) is de moeilijkheid van het referentie-examen t.o.v. het anker bepaald. Via deze omweg van het anker kan nagegaan worden hoe de moeilijkheid van het huidige examen is, in vergelijking met het referentie-examen.

Doordat we vele jaren achter elkaar door leerlingen exact dezelfde vragen voorleggen (in pre- en posttests) kunnen we ook de prestatie van leerlingen op deze vragen over de jaren heen met elkaar vergelijken. Als we zien dat leerlingen op deze vragen in de loop van de jaren steeds beter gaan scoren én omdat we de vragen geheim houden kunnen we de toegenomen score toewijzen aan een gestegen vaardigheid (en niet aan meer voorkennis over de

vragen). Hiermee weten we echter alleen nog maar iets over de vaardigheidsontwikkeling van de pre- en posttestpopulatie over de jaren heen. Maar door de koppeling van het anker aan de feitelijke examenopgaven kennen we de moeilijkheidsgraad van het afgenomen examen ook. Als leerlingen nu op een even moeilijk examen beter gaan presteren, is dat dus volledig toe te wijzen aan de gestegen vaardigheid. Dit leidt dan tot een N-term met een hoger gemiddeld cijfer dat passend is bij de gemeten vaardigheidsstijging.

6 Prestatieverandering over de jaren heen

Sinds de invoering van de aangescherpte uitslagregels is vaardigheidsverandering van kandidaten een veel besproken onderwerp geweest. Het verbaasde dus niet dat hierover de nodige vragen zijn gesteld. De vragen die in het webinar niet aan de orde zijn geweest, beantwoorden we hier. Ook het onderwerp 'normering op basis van aanvullende gegevens' hangt hier nauw mee samen.

Vraag	Antwoord
Hoe bewaak je nu dat het niveau over langere tijd gelijk blijft. Ik gaf in de jaren 80 les (economie), daarna ongeveer 25 jaar niet (in de financiële wereld gewerkt), en sinds een paar jaar geef ik weer les. Voor mijn gevoel is het niveau van vragen stellen toch echt lager. Of is mijn referentiekader veranderd?	In principe mag je ervan uitgaan dat als je van jaar tot jaar de lat gelijk houdt, je die over langere tijd ook gelijk houdt. Om dit te bewaken houden we ook in de gaten of er trends waarneembaar zijn die zouden kunnen verraden of er onbedoeld toch sprake is van een daling van het niveau. Zo is bijvoorbeeld bij het vak biologie op item-niveau geconstateerd dat leerlingen steeds slechter scoorden op kennisvragen terwijl daar geen redenen voor waren in wijzigingen van bv exameneisen. Na een zorgvuldige analyse is besloten om op basis hiervan de norm enigszins aan te passen en dus een mogelijke niveaudaling te voorkomen. Daarnaast ligt het voor de hand dat in die lange tijd het programma wel eens is gewijzigd. Met de wijziging van een programma moet ook de prestatie-eis opnieuw worden vastgesteld voor dat programma. Het is mogelijk dat dat feitelijk een verhoging of verlaging van het niveau betekent.
Als de normering 2016 bij Engels 1,4 is terwijl die voor Duits 0,4 is, dan is het hogere gemiddelde (Duits 6,1 en Engels 7,0) toch niet alleen aan een prestatievertering toe te wijden?	Uit de aanvullende gegevens is duidelijk geworden dat de (lees)vaardigheid bij Duits de laatste jaren niet toeneemt. Om deze reden heeft de keuze van de N-term tot gevolg dat de gemiddelde cijfers ongeveer gelijk zijn gebleven. Bij Engels hebben we wel waargenomen dat leerlingen op een (geheime) set opgaven die zij buiten het examen hebben gemaakt, steeds beter scoren. Deze vergrote vaardigheid is ook zichtbaar in de gemiddelde cijfers. Dat Engels een hogere N-term heeft dan Duits is een toeval dat niets te maken heeft met een vaardigheidsverschil. Een N-term is alleen een instrument om te compenseren voor verschillen in moeilijkheid tussen examens. Zo zal, gezien de N-term, het examen Engels relatief moeilijker zijn geweest dan het examen Duits (voor zover je vakken met elkaar kunt vergelijken uiteraard).
Wat gebeurt er als het onderwijs landelijk verbeter/verslechtert?	Als het onderwijs verbetert, zullen de prestaties van de leerlingen verbeteren. De leerlingen zullen dan relatief meer vragen goed beantwoorden. Hierdoor zullen de N-termen gemiddeld genomen niet anders worden, maar de gemiddelde cijfers van de leerlingen wel.

7 Vakspecifieke vragen

Er zijn meerdere vragen binnen gekomen over de situatie bij **Frans vwo** in 2016 t.o.v. de jaren ervoor:
Het klopt dat de N-termen voor de examens Frans havo en vwo vaak laag zijn. Dit heeft te maken met de moeilijkheidsgraad van de examens, voor een hogere N-term hadden de examens moeilijker moeten zijn. Wij kunnen ons voorstellen dat dit vreemd overkomt, omdat regelmatig ook het gemiddelde cijfer voor deze examens vrij laag is.
Wat is hier aan de hand?
In de afgelopen tien jaar hebben Cito en het CvTE kunnen vaststellen dat de leesvaardigheid van kandidaten Frans havo en vwo is gedaald. Omdat wij echter gehouden zijn aan de afspraak om van jaar tot jaar dezelfde eisen aan kandidaten te stellen, leidt dit bij de centrale examens Frans havo en vwo regelmatig tot lagere gemiddelde cijfers. Bij het examen Frans vwo 2016 was dit niet het geval, omdat dit examen niet de vereiste moeilijkheidsgraad had. Uitgaande van de moeilijkheidsgraad van het examen had bij dit examen een negatieve N-term moeten worden vastgesteld; iets wat het CvTE niet doet. Daardoor konden we in 2016 voor Frans vwo niet de gelijke eisen aan examenkandidaten stellen als in de jaren ervoor. Dit is dan ook de reden dat het gemiddeld cijfer bij dit examen zo hoog was.
Wij zijn van mening dat deze situatie onwenselijk was. Uiteraard zijn wij nagegaan hoe dit is ontstaan. Voor de examens 2017 hebben wij maatregelen genomen in de hoop een dergelijke situatie te voorkomen.

Vraag	Antwoord
In hoeverre is TTO onderwijs van invloed op het gemiddelde cijfer van **Engels**. Zijn niet-tto scholen/leerlingen in het nadeel?	Voor de normering van de centrale examens Engels havo en vwo wordt de moeilijkheidsgraad van het nieuwe te normeren examen door een posttest vergeleken met een referentie-examen. Deze werkwijze maakt het mogelijk om populatie-onafhankelijk te normeren. Het CvTE heeft de opdracht met de normering de prestatie-eis van jaar tot jaar gelijk te houden. TTO-onderwijs heeft hier geen invloed op. Er is daarom ook geen sprake van een nadeel voor niet-TTO-leerlingen. Als de totale populatie voor Engels vaardiger zou worden doordat er meer TTO-scholen zijn, dan zou dat resulteren in een hoger gemiddeld cijfer voor het vak Engels.
In het jaar met het eindexamen Seneca voor **Latijn** had ik een leerling met 1 fout, ze kreeg als cijfer een 9,5. Een jaar (of twee) later had ik een leerling met 4 fouten, zij kreeg als cijfer een 9,7, dat voelt onrechtvaardig. Om die reden vind ik dat de N-term geen recht doet aan de prestaties van de leerlingen. Vooral de goede leerlingen krijgen soms niet het cijfer dat ze verdienen. Is daar iets aan te doen? (Overigens heb ik de cijfers niet meer gecontroleerd, maar deze cijfers stonden zo in mijn herinnering)	Blijkbaar was het examen in het 2e jaar moeilijker, waardoor een leerling met meer fouten een zelfde of zelfs een hoger cijfer kon halen. Het zou naar ons oordeel onterecht zijn als een leerling bij een moeilijk examen evenveel punten moet halen voor een zelfde cijfer. Dat is wat we in de jaren ’70, ’80 en ’90 deden. Wel is het zo dat ons systeem van normeren bij de centrale examens als ijkpunt de grens tussen voldoende en onvoldoende heeft. Dat heeft tot gevolg dat de nauwkeurigheid van het toegekende cijfer afneemt naarmate je verder van die grens komt. Op dit moment is dat een punt van studie in het zogeheten normeringsoverleg van het CvTE en Cito.

Vraag	Antwoord
Zien jullie een verband tussen het verplicht stellen van **wiskunde** en de verandering van de N-term voor dit vak sinds het verplicht is gesteld?	Op het vmbo is wiskunde niet verplicht. Het profiel havo C&M kent ook geen verplichte wiskunde. De overige profielen in havo en vwo wel. Recentelijk is er in deze verplichtstelling niets veranderd. De N-term weerspiegelt de schommeling in moeilijkheid tussen de examens van jaar tot jaar en heeft niets te maken met de vaardigheid van de leerlingen.
Hoe kan het dat **natuurkunde** hetzelfde gemiddelde had in de 3 gegeven jaren (voor de N-correctie?), terwijl de N-term toegepast heel anders werd?	De N-term is een instrument om te compenseren voor verschillen in moeilijkheid. Blijkbaar waren de examens in de genoemde jaren verschillend van moeilijkheidsgraad. Als het gemiddeld cijfer over die jaren wel constant is, betekent dat dat de groep examenkandidaten over die jaren een constante vaardigheid had. De N-term en het gemiddeld cijfer geven ieder verschillende informatie.
In het **natuurkunde**-examen van 2014 is er wel een vraag niet meegeteld die 5 punten waard was. Dit heeft geleid tot een oneerlijke beoordeling van leerlingen die deze moeilijke vraag (waarbij de formule gegeven had moeten worden maar ook in BINAS te vinden was) helemaal goed hadden.	U heeft gelijk dat deze situatie niet de schoonheidsprijs verdiende. Dit is dan ook een situatie die wij grondig hebben geanalyseerd en waar we van geleerd hebben. De uitgangspunten voor het schrappen van een vraag hebben we daardoor ook deels aangepast.
Het examen **economie** vmbo-tl is ingrijpend veranderd. Er is geen referentie examen hiervoor (naar mijn gevoel). Een nieuw programma dat ingevoerd werd zonder rekening te houden hoe wij dat als docenten moesten oppakken. Wordt dit ook meegenomen in de vaststelling van de n-term?	Bij economie GL/TL werken we op basis van een pretest, volgens vakinhoudelijke deskundigen kunnen de resultaten van deze pretest gehandhaafd worden. Bij het vaststellen van de N-term houden we – indien nodig – altijd rekening met een gewijzigde syllabus. Het examenprogramma is ongewijzigd, alleen de syllabus is aangepast, zie ook de vakspecifieke informatie in de Septembermededeling en pagina vernieuwing syllabus economie op Examenblad.nl.
hoe komt het dat er een groot verschil zat in de N-term voor **wiskunde** bij VMBO -BB examens 2016?	Net zo min als het bij papieren CE's uit opeenvolgende jaren altijd lukt om verschillen tussen de N-termen te voorkomen, lukt dat bij de varianten van een digitaal CE. Bij de digitale CE's wiskunde pakt een cluster van opgaven soms moeilijker uit dan verwacht of juist makkelijker. Alle varianten zijn via overlap van opgaven aan elkaar gelinkt. Daarom kunnen we zien of een bepaalde variant moeilijker of makkelijker was dan de andere. De eventuele verschillen in moeilijkheidsgraad vertalen zich in verschillen in N-termen tussen de varianten. Maar wat een leerling moet kennen en kunnen om een voldoende te behalen, is daardoor niet afhankelijk van de variant die hij heeft afgelegd. Net zo min als het in die zin uitmaakt of een papieren CE in 2015 of 2016 is afgelegd.

8 Normering 2e tijdvak

Tijdens het webinar was de tijd te kort om nog uitgebreid in te gaan op de normering van het 2e tijdvak. Vragen die hierover nog zijn gesteld, worden hieronder beantwoord.

Vraag	Antwoord
Hoe kan het dat er vanuit gegaan wordt dat tijdvak 1 en 2 dezelfde moeilijkheidsgraad hebben, maar het jaar erop is een andere moeilijkheidsgraad?	Examens van 1 jaar worden in 1 cyclus vastgesteld door dezelfde mensen. Dat vergroot de kans dat ze gelijk zijn. Omdat we daar niet zonder meer vanuit kunnen gaan, controleren we deze aanname op de manier zoals uitgelegd in het webinar.
De prestatie van kandidaten in het 2e tijdvak kan beïnvloed worden doordat leerlingen alleen maar herkansen voor een hoger cijfer, of zich er juist helemaal niet meer voor in zetten. Dan is de vooronderstelling toch ook weer anders?	Die mogelijke meetfout vermijden we door bij de normering van het 2e tijdvak ons alleen te baseren op de leerlingen die in het 1e tijdvak een onvoldoende gehaald hebben. Het aandeel leerlingen dat zijn best niet doet voor de herkansing is in die groep zo klein mogelijk. Door jaar in jaar uit naar de prestatie van deze zelfde groep te kijken, hebben we een duidelijk beeld van de prestatie van deze groep in tijdvak 2 t.o.v. tijdvak 1 en kunnen we afwijkingen in dat patroon toewijzen aan de verschillen in moeilijkheid tussen het 1e en 2e tijdvak.
Hoe “eerlijk” is een N-term voor een “klein” vak in het tweede tijdvak? (oftewel waar komt de statistiek in de knel met de kleine populatie)	Kleine vakken kennen in het 2e tijdvak vaak een afname door de staatsexamencommissie. Die examens kennen ook hun eigen normeringssystematiek. In andere gevallen proberen we de zwakke statistische gegevens op basis van kleine aantallen aan te vullen met andere gegevens zoals bijvoorbeeld de inschatting van de moeilijkheidsgraad door experts.
Is het dan eerlijk om de N-term voor tweede en derde termijn hetzelfde te laten zijn als in de eerste termijn? Volgens mij wordt er alleen afgeweken als er fouten in het examen zijn gemaakt.	Zie hiervoor de uitleg in het webinar. Behalve voor fouten, hogen we de N-term in het 2e tijdvak ook op als het examen aantoonbaar moeilijker is gebleken dan het 1e tijdvak. Dat gebeurt ieder jaar bij meerdere vakken. De afname van examens in het 3e tijdvak wordt gedaan door de staatsexamencommissie. De N-term voor die examens is niet direct gerelateerd aan die voor tijdvak 1 en 2 in dat jaar.

9 Overige vragen

Tot slot is er nog een grote hoeveelheid vragen die niet direct onder de voorgaande onderwerpen te vatten zijn. Deze worden hieronder apart beantwoord.

Vraag	Antwoord
Elk examen 15 % of meer onvoldoende, dat is een soort streven? Mag je dan verwachten dat gemiddeld elke leerling bijna 2 onvoldoendes heeft (ca. 10 vakken op de lijst)? Als	Minstens 15% onvoldoendes per CE is geen streven. Maar als er landelijk gezien duizenden kandidaten zijn, is een examen met 0% onvoldoendes toch vreemd: Hoe kan het nu dat iedereen een voldoende haalt? Omgekeerd kun je je ook afvragen of het met duizenden kandidaten zou kunnen gebeuren dat minder dan de helft een voldoende haalt. Tussen die 0% onvoldoendes en 50% onvoldoendes hebben wij de percentages 15% en 35% gemarkeerd: minder dan 15% vinden wij weinig, meer dan 35% veel.

Vraag	Antwoord
dat op mijn school niet zo is (of mij lijkt) zijn er dan scholen met grote aantallen onvoldoendes ?	Vooraleer we N-termen vaststellen waarbij het percentage onvoldoendes kleiner is dan 15 of groter is dan 35, vragen we ons altijd af wat daarvan de reden kan zijn; noem het een reality check. Maar percentages onvoldoendes onder 15% of boven 35% komen wel voor. Verder is het ons bekend dat de verschillen tussen scholen groot zijn. Het is dus goed mogelijk dat het aantal leerlingen dat op uw school een onvoldoende voor een vak (en/of voor het totaal van vakken) haalt, kleiner is dan het landelijk gemiddelde.
Zit er verschil in hoe goed een leerling een papieren examen maakt en hoe goed een leerling een digitaal examen maakt?	In de beginjaren (2005 – 2010) van de digitale CE's BB en KB evalueerden we die systematisch met docenten, examensecretarissen, systeembeheerders en de leerlingen zelf. Zowel docenten als leerlingen waren in overgrote meerderheid voorstander van digitale CE's. Minder dan 10% van de leerlingen zei: 'Geef mij maar papier.' Docenten gaven als voordeel aan dat de leerling bij digitaal – anders dan bij papier – maar één vraag tegelijk zag. Uit de enquêtes tot dusverre werden voor BB en KB overwegend voordelen gezien. Bij de digitale CE's BB en KB is geen systematisch onderzoek verricht naar hoe leerlingen presteren bij de papieren en de digitale examenvorm. Bij de rekentoets heeft dit papier/digitaal onderzoek wel plaatsgevonden. Daar zijn geen significante verschillen tussen papier en digitaal geconstateerd. Het is altijd mogelijk dat een individuele leerling beter presteert op een digitale dan op een papieren toets, en omgekeerd. Maar generiek gezien hebben wij daar geen aanwijzingen voor.
Is het geen optie om invloeden op de N-term van uitgevoerde correcties bekend te maken via het verslag van de examencampagne i.p.v. nabellen via de examenlijn?	Op dit moment doen we dat niet. Het is de moeite hierover door te denken met in acht neming van de tijdsfactor en de werklast.
Leidt de N-term in combinatie met het feit dat leerlingen nooit de dupe mogen zijn niet tot gemakzucht?	Daar is ons nog nooit iets van gebleken.
Jullie spreken telkens over "moeilijk" en "makkelijke" examens, maar hoe definieert u "moeilijk" en/of "makkelijk"?	Moeilijk is een relatief begrip. Wanneer alle leerlingen in Nederland gemiddeld 60% van de maximale score weten te behalen dan vinden we dat normaal. Als de leerlingen slechts 55% of nog minder weten te scoren dan wordt al snel het woord moeilijk gebruikt. Na afloop van de examens voeren de docenten de scores in het programma Wolf in. Ook hier wordt gevraagd of de docenten het examen (te) moeilijk vonden. Het is een gevoelswaarde die door CvTE en Cito niet is vastgelegd in waarden. Grosso modo zou je kunnen zeggen dat een examen met een N-term van 1,5 of hoger moeilijk gevonden wordt. De N-term wordt (zie uitleg in het webinar) bepaald door de scores van de leerlingen en niet op basis van de subjectieve waarneming.

Vraag	Antwoord
Wordt de normering ook gebruikt als politiek instrument?	Nee. Het CvTE heeft als opdracht om het niveau over jaren gelijk te houden om zo het civiel effect van de diploma's te behouden. Het CvTE krijgt dus niet op enig moment vanuit 'Den Haag' een telefoontje om nu maar eens wat meer of juist wat minder mensen te laten slagen.
Het gaat er toch om dat aan de prestatie van de leerling recht wordt gedaan. Daar past toch niet "mazzel hebben" bij, zoals een paar keer is voorbijgekomen ?	Correct. Sprake van mazzel is er dan ook alleen als er sprake is van een fout in het examen. Wij geven dan de voorkeur aan 'false positive' boven 'false negative'. Verder kan een kandidaat 'mazzel' hebben als het examen in het 2e tijdvak makkelijker is dan het examen in het 1e tijdvak.
Bij terugtrekken van een vraag wordt aan ALLE kandidaten de maximale score toegekend. Dat is wat mij betreft niet eerlijk t.o.v. leerlingen die deze vraag hebben beantwoord en daarvoor ZELF ook punten hebben gescoord. Zij hebben tijd verloren door met de opgave te stoeien en hadden dus minder tijd voor het beantwoorden van de overige vragen en zijn daardoor mijns inziens ernstig benadeeld.	Ik begrijp dat dit een gevoel van onrechtvaardigheid oproept. Het argument dat leerlingen aan deze opgave tijd hebben verloren, gaat alleen op als er sprake is van tijdnood bij het examen of als een leerling onevenredig veel tijd aan juist deze vraag heeft besteed. In het eerste geval zullen wij hiervoor ook een compensatie toepassen, los van het besluit voor deze specifieke vraag. In het tweede geval betreft dat ook de vaardigheid van een leerling om zijn tijd gelijkmatig over het examen te verdelen. Los van het aspect tijdnood gaan we ervan uit dat een vraag die uit het examen geschrapt wordt niet deugt en dat het gegeven antwoord volgens het cv misschien goed is, maar feitelijk niets zegt over de vaardigheid van de kandidaat. De behaalde score zegt dus ook niet of de kandidaat het goed begrepen heeft of niet. Is dat niet het geval en zegt het gegeven antwoord en dus de behaalde score wel iets over zijn vaardigheid (bv. omdat je de vraag zoals bedoeld kunt opvatten, maar ook op een andere manier) dan is het schrappen van de vraag niet de juiste oplossing óf is naast het schrappen extra compensatie via de N-term nodig.
Mis je bij het onderverdelen van een examen in blokken niet de totaalbelasting van het examen? Soms blijken bij een SE bij ons, de combinatie van een aantal vragen juist de moeilijkheid op te leveren (ik geef biologie)	Bij een pretest wordt alleen de moeilijkheid van de vragen ten opzichte van elkaar gemeten. Het klopt dat het niet alleen afhangt van de moeilijkheid van de individuele vragen of een examen een goed examen is. De lengte en onderlinge samenhang is ook van belang. Hier wordt in de uiteindelijke samenstelling van het examen dan ook nog specifiek op gelet.
Is het wenselijk dat je een 3 krijgt als je het examen heel slecht maakt? En niet een 1 of 2?	Ik neem aan dat u refereert aan het systeem van normering dat we in de jaren '70 hanteerden. Een leerling kon toen zonder iets goed te hebben al een 3 halen als er (fors) opgehoogd was vanwege een moeilijk examen en/of een onvolkomenheid. In die tijd werd het als ongewenst ervaren dat de zwakke leerlingen er onevenredig van profiteerden als er een ophoging werd toegepast. Uiteindelijk is het natuurlijk persoonlijk of je dit wenselijk acht of niet, maar wij werken al sinds jaar en dag met een cijferschaal van 1-10 en het ligt dus voor de hand dat je met niets goed een 1 krijgt.
Wanneer het examen meerdere vragen bevat die door veel leerlingen fout worden gemaakt, heeft dit dan invloed op de	Hier staat in feite dat het examen meerdere moeilijke vragen bevat. Als dat het geval is, zal de gemiddelde score relatief laag uitvallen en zal de N-term hoger worden om te compenseren voor deze moeilijke vragen. Met andere woorden: de gemiddelde score is direct gelinkt aan de N-term. Als er

Vraag	Antwoord
N-term? Of wordt juist naar de gemiddelde score voor het examen gekeken?	enkele moeilijke vragen zijn, maar daarnaast zijn er ook enkele heel erg makkelijke vragen dan zal de N-term niet extra hoog uitvallen.
Hoe kun je de N-term-aanpak toepassen op je eigen schoolonderzoeken?	De N-termaanpak kan wel, onder bepaalde voorwaarden, gebruikt worden. De formules van de CvTE methode zijn bruikbaar. Je kunt de moeilijkheid van de toets zelf bijsturen als je het gevoel hebt dat het schoolonderzoek moeilijk of makkelijk was. Je moet bij het evalueren of het een moeilijke of makkelijke toets was wel meenemen dat in een klassensituatie vaak t.o.v. het centraal examen kleine aantallen leerlingen zitten. Daarbij moet ook worden bedacht dat het schoolonderzoek vragen bevat van een zodanige moeilijkheid dat wanneer ongeveer de helft van de maximale score wordt gehaald, dit met een voldoende gewaardeerd kan worden. Als u een docent moderne vreemde talen bent, kunt u gebruik maken van mix en meet. Dit is een applicatie die op www.cito.nl is terug te vinden. Voor de andere vakken zou een berekening als volgt kunnen geschieden: zoek in de oude toets en itemanalyses op www.cito.nl de p-waarde van de vragen die in de toets zitten. Bereken hiermee de verwachte gemiddelde score. Deel deze door de maximale score. Dit levert de verwachte gemiddelde p 'waarde. Bepaal het gemiddeld cijfer dat je wilt krijgen. Het gewenste gemiddeld cijfer minus 9 maal de gemiddelde p 'waarde levert de gewenste N-term.
Hoe is te verklaren dat docenten in het algemeen een dalende trend in het niveau van hun leerlingen menen te constateren, terwijl de CE-resultaten die trend niet weergeven?	U stelt een moeilijk te beantwoorden vraag. Waarschijnlijk speelt hier dat er sprake is van verschillende waarnemingen. U constateert de daling op basis van de prestaties van de leerlingen die u concreet voor u heeft gezien in de afgelopen jaren. De CE-resultaten zijn gebaseerd op de gehele populatie, landelijk, eveneens over jaren. Het is de vraag of die metingen één-op-één met elkaar vergelijkbaar zijn. Zie hierover ook de eerste vraag bij het onderwerp 'prestatieverandering over de jaren heen'.
De ervaring die ik met centraal examens heb, is dat de cijfers altijd dicht bij elkaar liggen. Ook goede leerlingen scoren (bij Kunst algemeen) zelden boven een 7,5. Kan dit mede veroorzaakt worden door de manier van het berekenen van de N-term?	De voornaamste reden dat er weinig kandidaten boven de 7,5 scoren op een examen zit hem in de samenstelling van het examen en de (geringe) spreiding die het examen onder leerlingen verzorgt.
Is het mogelijk dat er gedetailleerde informatie wordt verstrekt aangaande de uitslag? Tot nu toe wordt alleen per domein of per opgave informatie gegeven en hoe je daarop gescoord hebt. Ik wil graag weten of mijn leerlingen vraag 6 (bijvoorbeeld) ook zo goed / slecht gescoord hebben als de rest van het land.	Ja; er wordt gedetailleerde informatie teruggekoppeld. Naast de groepsrapportage worden sinds 2013 de resultaten van de toets- en itemanalyse op de site van Cito bij de examendocumenten getoond. Hierin is te zien hoe de landelijke steekproef per vraag gescoord heeft. Deze kunt u dan vergelijken met de scores van uw leerlingen. http://www.cito.nl/onderwijs/voortgezet%20onderwijs/centrale_examens/normering_alg.aspx

Vraag	Antwoord
heeft u een uitleg over de gebruikte termen in de TIA's. Het gaat me om: blad 1 - MISSING / P /Sd / RSK / Rit / Rir / Ar	Cito publiceert de toets- en itemanalyses van de eerste tijdvak examens op de site bij de overige examendocumenten. Boven elke tabel staat een link: toelichting op de tabel en op de documenten. Deze link verwijst naar een pagina waar een afzonderlijke passage aan TIA (toets- en item-analyse) wordt gewijd. Deze passage bevat weer links naar twee bijlagen waarvan er een ingaat op de begrippen uit de tia : Begrippen uit de TIA en het andere document over de opbouw van de TIA: http://www.cito.nl/onderwijs/voortgezet%20onderwijs/centrale_examens/normering_alg.aspx