College voor Toetsen en Examens

DE NORMERINGSSYSTEMATIEK VAN DE CENTRALE EXAMENS VO

30 november 2016

Inhoud

Inleiding

In deze notitie zetten wij de normeringssystematiek voor de centrale examens vo uiteen. De normeringssystematiek leidt er toe dat wij na de afname voor ieder centraal examen een zogeheten normeringsterm of N-term vast stellen.
Wie ervaring heeft met centrale examens is bekend met de N-termen.
Voor wie niet bekend is met de N-termen en toch kennis wil nemen van deze notitie volgt hier een korte introductie. Het fenomeen N-term wordt in het vervolg van deze notitie bekend verondersteld.

kader 1

Introductie: wat is de N-term?

Examenmakers slagen er niet vaak in een examen precies de beoogde moeilijkheidsgraad mee te geven. Dit is een wereldwijd erkend ervaringsgegeven. Het geldt ook voor makers van de centrale examens. Dat zijn de leden van de constructiegroepen van Cito, de toetsdeskundige van Cito en de voorzitter en leden van onze vaststellingscommissies. Iedere individuele leraar heeft deze ervaring: ook schriftelijke overhoringen, proefwerken en schoolexamens kunnen achteraf te moeilijk of te gemakkelijk blijken.

De N-term is een maat voor de moeilijkheidsgraad van een centraal examen.
Bij de meeste centrale examens wordt vooraf gemikt op een N-term van 1,0.

Bij een N-term van 1,0 geldt: cijfer = 9 x (behaalde score/maximale score) + 1,0.
Een score van 53 punten op een examen waarvoor maximaal 90 scorepunten behaald kunnen worden, levert dan als cijfer: 9 x 53 / 90 + 1,0 = 5,3 + 1,0 = 6,3.

Een examen dat achteraf gezien makkelijker bleek dan was beoogd, krijgt een lagere N-term mee, bijvoorbeeld: 0,7.
Een score van 53 van de 90 punten levert dan als cijfer: 9 x 53 / 90 + 0,7 = 6,0.

Als een centraal examen voor een bepaald vak in 2015 een N-term van 1,0 had en in 2016 een N-term van 0,7, dan was het examen van 2016 0,3 cijferpunt makkelijker [1] dan dat van 2015.
Om een 5,5 te behalen had een leerling die in 2015 eindexamen deed 45 van de 90 scorepunten nodig (9 x 45 / 90 + 1,0 = 5,5). Voor diezelfde 5,5 waren in 2016 48 scorepunten nodig (9 x 48 / 90 + 0,7 = 5,5).
Dat in 2016 meer [1] scorepunten nodig waren voor een voldoende dan in 2015 is billijk omdat het centraal examen voor dit vak in 2016 makkelijker [1] bleek dan dat van 2015.

De N-termen stellen ons in staat om voor verschillen in moeilijkheidsgraad tussen opeenvolgende centrale examens te compenseren.
Wat een leerling moet kennen en kunnen om een voldoende te behalen voor het centraal examen is daardoor niet afhankelijk van het jaar waarin hij eindexamen doet.

Reacties op deze notitie

Wij willen meer bekendheid geven aan de normeringssystematiek van de centrale examens. Wij hopen dat deze notitie daar een bijdrage aan levert.
De notitie kan ongetwijfeld nog worden verbeterd.

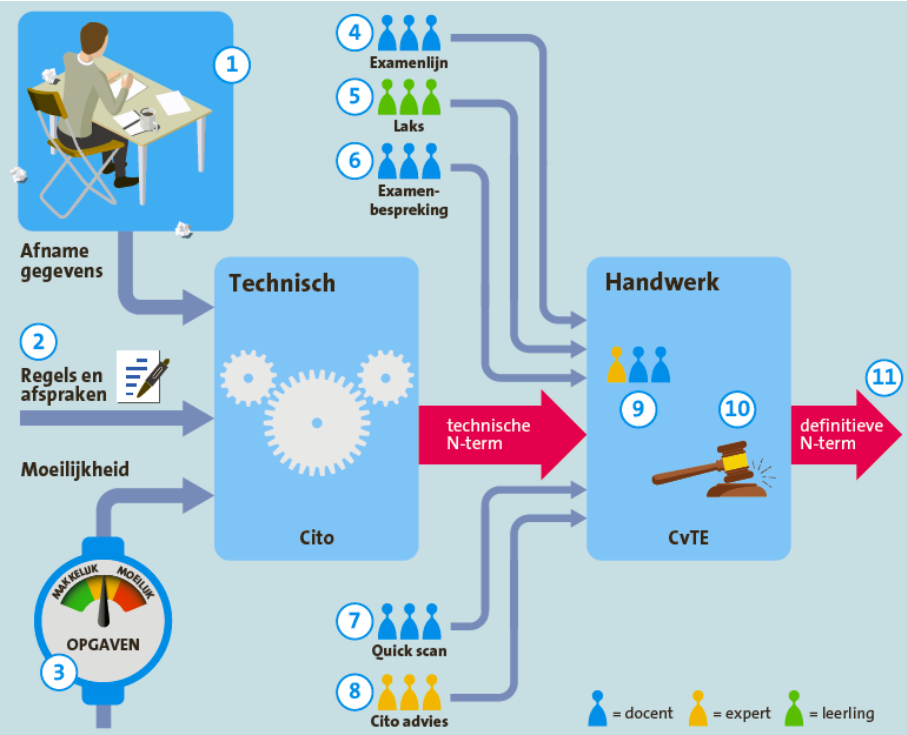
[1] rode tekst: rectificatie t.o.v. de versie van 28 november

Wie onduidelijkheden signaleert of verbetersuggesties heeft, is van harte uitgenodigd om ons daarvan in kennis te stellen via info@hetcvte.nl. Dit kan de leesbaarheid van volgende versies van deze notitie alleen maar ten goede komen.

Samenvatting

Als College voor Toetsen en Examens (CvTE) hebben wij ons als taak gesteld om *per centraal examen* 'de lat' van jaar tot jaar even hoog te leggen. Concreter: wat een leerling moet kennen en kunnen om voor het centraal examen biologie in de gemengde en theoretische leerweg (gl/tl) een voldoende te halen, is in 2015 hetzelfde als in 2010 en 2020.
Met een normeringssystematiek die relatieve of populatie-afhankelijke normering als uitgangspunt heeft, lukt dat niet. Daarom kiezen wij voor het uitgangspunt van absolute of populatie-onafhankelijke normering.

Het normeringsproces van de centrale examens vo kan als volgt gevisualiseerd worden:



1 Docenten geven scores per vraag van hun leerlingen door aan Cito m.b.v. Wolf.

2 Cito doet voorstellen voor regels en afspraken voor de normering. CvTE stelt deze vast. Ze zijn te vinden op toetswijzer.nl. Het technisch normerings-advies volgt zodoende rechtstreeks uit de afnamegegevens.

3 Cito verzamelt extra informatie over de moeilijk-heid van de vragen voorafgaand aan de normering.

4 Docenten kunnen via de examenlijn vragen, opmerkingen en klachten doorgeven.

5 Leerlingen kunnen via LAKS vragen, opmerkingen en klachten doorgeven.

6 Examenbespreking door vakvereniging.

7 Het oordeel van docenten over het examen wordt in Wolf ingevuld.

8 Cito geeft advies over de normering aan CvTE.

9 Vaststellingscommissie bekijkt of er een aanpassing van de technische N-term noodzakelijk is afgaande op input uit het veld (4, 5, 6 en 7).

10 CvTE-leiding zorgt dat over vakken heen consistent wordt omgegaan met de gegevens en de geluiden uit het veld en stelt de definitieve N-term vast.

11 De definitieve N-term wordt gepubliceerd. Bij elke score kan nu het CE-cijfer bepaald worden.

1 Doel

Als College voor Toetsen en Examens (CvTE) hebben wij ons als taak gesteld om *per centraal examen* 'de lat' van jaar tot jaar even hoog te leggen. [2] Concreter: wat een leerling moet kennen en kunnen om voor het centraal examen biologie in de gemengde en theoretische leerweg (gl/tl) een voldoende te halen, is in 2015 hetzelfde als in 2010 en 2020.
Jaarlijks zijn er ongeveer 30.000 examenkandidaten biologie gl/tl. Binnen deze groep loopt het vaardigheidsniveau sterk uiteen. Er zijn leerlingen die gemakkelijk een acht halen en leerlingen die hun uiterste best moeten doen voor een zes. Maar stel nu dat de 30.000 kandidaten van 2020 *als groep* net zo vaardig zijn als de groep van 2015, dan zal het gemiddelde CE-cijfer voor biologie gl/tl in 2020 (nagenoeg) gelijk zijn aan dat van 2015.
Maar stel dat de examenkandidaten biologie gl/tl van 2020 *als groep* vaardiger zijn dan de groep van 2015, dan zal het gemiddeld CE-cijfer voor biologie gl/tl in 2020 hoger liggen dan het gemiddelde van 2015.

1.1 Absolute normering

Wordt een groep kandidaten vaardiger dan behoort het gemiddeld CE-cijfer te stijgen. Neemt de vaardigheid van de groep van examenkandidaten af, dan daalt ook het gemiddeld CE-cijfer. Een normeringssystematiek waarbij de hoogte van de lat onafhankelijk is van het vaardigheidsniveau van de groep van examenkandidaten als geheel, wordt populatie-onafhankelijke of absolute normering genoemd.

1.2 Relatieve normering

Tegenover absolute normering staat relatieve of populatie-afhankelijke normering. Bij relatieve normering wordt de lat per centraal examen zo gelegd dat het gemiddeld cijfer (bij voorbeeld) een 6,3 is of het percentage voldoende (bij voorbeeld) 80%.
Als wij relatieve normering als uitgangspunt genomen zouden hebben, dan zou de groep biologie gl/tl van 2020 hetzelfde gemiddelde CE-cijfer krijgen als de groep van 2015, *ook al zou deze groep wel eens vaardiger kunnen zijn*. Een examenkandidaat die in 2020 voor het centraal examen biologie gl/tl het cijfer 5,3 behaalt, zou dan net zo vaardig kunnen zijn als een kandidaat van 2015 met cijfer 5,5. Bij relatief normeren is de hoogte van de lat dus afhankelijk van de vaardigheid van de populatie examenkandidaten van dat jaar en kan die hoogte van jaar tot jaar verschillen.

1.3 Het CvTE kiest voor absolute normering

Wij vinden dat alle leerlingen voor ieder centraal examen dat zij afleggen het cijfer moeten krijgen dat zij verdienen en daarom is gekozen voor het uitgangspunt van absolute – populatie-onafhankelijke – normering. [3]

[2] Behoudens aanpassingen, zoals bij voorbeeld in 2015 bij Nederlands vwo. Om te kunnen voldoen aan het referentieniveau 4F is de lat bij het centraal examen Nederlands 0,4 cijferpunt hoger gelegd dan in de voorafgaande examenjaren. Dat dit zou gebeuren is aan het begin van het schooljaar 2014-2015 aan de directies en de docenten Nederlands vwo bekend gemaakt.
[3] De normeringssystematiek is vastgelegd in de Regeling Omzetting scores in cijfers van het CvTE. In de Wet Cvte is geregeld dat voor de inwerkingtreding van deze regeling de goedkeuring van de minister nodig is.

Een nadeel van die keuze is dat de werkwijze bij de normering technisch ingewikkeld is en daardoor veel moeilijker is uit te leggen. Het heeft daardoor de schijn niet transparant te zijn.

In paragraaf 2 wordt eerst de werkwijze bij relatief normeren beschreven, omdat die veel begrijpelijker is. Aan het eind van paragraaf 2 wordt beschreven dat er in 2012, toen de aangescherpte uitslagregels werden ingevoerd, het slaagpercentage 8% lager gelegen had, als de werkwijze van relatieve normering zou zijn toegepast.
Daarna wordt in paragraaf 3 de werkwijze beschreven die daadwerkelijk wordt toegepast: de werkwijze waarbij absolute normering het uitgangspunt is.

2 De werkwijze bij relatieve normering

Stel dat als uitgangspunt zou gelden dat ieder centraal examen een gemiddeld cijfer van 6,3 heeft.

2.1 Drie stappen

Per centraal examen worden bij de normering drie stappen doorlopen:
- een technisch normeringsadvies van Cito;
- een advies van iedere CvTE-vaststellingscommissie;
- de vaststelling van de N-term (zie kader 1)

kader 2

Voor de N-term: zie eerst de introductie in de inleiding van deze notitie. Een *normeringsterm* of *N-term* is een getal dat voor alle mogelijke scores voor een centraal examen bepaalt met welk cijfer die score gewaardeerd wordt. Bij N = 1,0 moet de leerling 50% van de maximale score behalen voor een voldoende cijfer. Bij N = 0,0 is dat ongeveer 60% en bij N = 2,0 ongeveer 40%.

Stap 1: het technisch normeringsadvies
Na de afname maakt Cito van ieder centraal examen een toets- en itemanalyse. Dit gebeurt op basis van de scores die de correctoren per vraag aan de kandidaten hebben toegekend. Via het programma Wolf (Webbased optisch leesbare formulieren) hebben de correctoren deze afnamegegevens aan Cito doorgegeven.[4] In een toets- en itemanalyse is zichtbaar hoeveel scorepunten de kandidaten gemiddeld hebben behaald, zowel per vraag als voor het gehele examen. Op het niveau van de toets blijkt daarnaast onder meer hoe de scorepunten gespreid zijn en hoe betrouwbaar het examen als meting was. De toets- en itemanalyse is daarmee de kwantitatieve graadmeter voor hoe het examen 'gevallen' is.
Cito stelt daarnaast bij ieder centraal examen een normeringstabel op waarin per N-term van 0,0 tot en met 2,0 zichtbaar is wat het gemiddeld CE-cijfer is en wat het percentage (on)voldoende is.

De N-term waarbij het gemiddelde cijfer op 6,3 uitkomt, zou dan – bij relatieve normering – het technisch normeringsadvies van Cito zijn.

Een andere manier om het uitgangspunt van relatief normeren te operationaliseren is dat ieder centraal examen een vast percentage voldoende, zeg 80 %, krijgt.
In dat geval wordt het technisch normeringsadvies van Cito bepaald door in de normeringstabel de N-term te zoeken, waarbij het percentage voldoende het dichtst bij 80 ligt.

Stap 2: het advies van de CvTE-vaststellingscommissie
Een of twee dagen voordat de N-termen aan de scholen bekend gemaakt worden komen al onze vaststellingscommissies bijeen voor de normeringsvergadering.
Iedere vaststellingscommissie beschikt dan over de toets- en itemanalyse en het technisch normeringsadvies van Cito en wordt bijgestaan door een toetsdeskundig vakmedewerker van Cito.
De vaststellingscommissie beschikt bovendien over de – kwalitatieve – reacties op het

[4] Bij de digitale centrale examens BB en KB is het doorgeven van de toegekende scores geautomatiseerd.

examen die zijn binnengekomen via de CvTE-examenlijn en over de verslagen van examenbesprekingen die door vakverenigingen georganiseerd zijn. Daarnaast heeft Cito via Wolf (zie stap 1) in de zogeheten quickscan ook kwalitatieve reacties verzameld. [5] Ook de resultaten van de quickscan liggen voor.
Onze vaststellingscommissie weegt alle reacties en brengt een advies N-term uit. Stel bij voorbeeld dat vakdocenten aangeven dat veel leerlingen in tijdnood gekomen zijn en de toets- en itemanalyse bevestigt dat doordat naar het einde toe vragen frequenter niet beantwoord zijn, dan zal de vaststellingscommissie dit in haar advies over de N-term verdisconteren. Een ander voorbeeld: als uit de toets- en itemanalyse zou blijken dat een bepaalde opgave door de minst vaardige kandidaten juist goed en door de vaardige kandidaten slecht is gemaakt, zal de vaststellingscommissie die vraag nog eens extra onder de loep nemen. Achteraf is dan meestal wel duidelijk hoe deze vraag beter gesteld had kunnen worden.
Bij haar advies over de N-term houdt de vaststellingscommissie rekening met eventuele minder valide vragen en voor toekomstige examens wordt daar lering uit getrokken. Voor de vaststellingscommissies en de toetsdeskundigen van Cito is de normeringsvergadering daardoor een belangrijk evaluatiemoment.
Bij een vakinhoudelijke fout is er een procedure om via de N-term te compenseren. Als een onvolkomenheid kort na de afname aan het licht komt, sturen wij een aanvulling op het correctievoorschrift naar alle correctoren, waardoor de correctie zo wordt uitgevoerd dat kandidaten niet de dupe zijn van een fout in het examen. Als een onvolkomenheid niet via een aanvulling op het correctievoorschrift kon worden rechtgezet, wordt een hogere N-term vastgesteld dan zonder de fout het geval zou zijn geweest.

Stap 3: de vaststelling van de N-term
Onze directeur en sectormanagers stellen op basis van deze stappen, na advisering door psychometrici van Cito, de N-termen vast. Voorafgaand aan de normeringsvergaderingen hebben wij, samen met Cito, ook overlegd met het LAKS, zodat behalve van de vakverenigingen, ook de bevindingen van de leerlingenorganisatie bekend zijn.
’s Ochtends om 8 uur van de dag na de normeringsvergaderingen worden alle N-termen bekend gemaakt via Examenblad.nl. Dit stelt de VO-scholen in staat om van hun kandidaten de cijfers voor de centrale examens en de uitslag van het eindexamen vast te stellen.

2.2 Samenvatting werkwijze bij relatief normeren

Samengevat komt de werkwijze bij relatieve normering er op neer dat, behoudens onvolkomenheden, per centraal examen de N-term zo gekozen wordt, dat het gemiddeld cijfer (bij voorbeeld) 6,3 is of het percentage voldoende (bij voorbeeld) ± 80% is.
Zo eenvoudig werkt relatief normeren.

Zijn er bij relatief normeren minder verschillende N-termen?
Nee, ook bij relatief normeren zal het niet zo zijn dat het centraal examenbiologie gl/tl jaarlijks een vaste N-term krijgt.

[5] In de quickscan worden 4 enquêtevragen aan de correctoren gesteld over: de moeilijkheidsgraad, de lengte, de aansluiting op het gegeven onderwijs, een waarderingscijfer.

Ook als de groepen van 2020 en 2015 even vaardig zijn, zou de N-term van biologie gl/tl in 2020 0,7 kunnen zijn, terwijl in 2015 als N-term 1,2 is vastgesteld. Onder de aanname van gelijke vaardigheid van de populaties van 2020 en 2015, zou dat verschil ontstaan als het examen in 2020 (0,5 cijferpunt = 1,2 – 0,7) moeilijker is dan het examen van 2015.
Ieder examen van tevoren precies de bedoelde moeilijkheidsgraad meegeven is praktisch ondoenlijk.

2.3 Relatieve normering en de aanscherping van de uitslagregels in 2012

Alvorens de overstap te maken naar de werkwijze die wij daadwerkelijk hanteren, gaan wij eerst nog in op wat er in 2012, het jaar waarin de uitslagregels werden aangescherpt, gebeurd zou zijn indien met relatieve normering gewerkt zou zijn.
In 2012 gold voor het eerst als extra voorwaarde om te kunnen slagen dat de CE-cijfers van een kandidaat gemiddeld tenminste 5,5 moeten zijn. Vanaf 2013 kwam daar voor het vwo en havo nog de kernvakkenregel bij. [6]
Vooraf was berekend wat de gevolgen geweest zouden zijn als de nieuwe uitslagregel was toegepast op de resultaten van de examenkandidaten van 2011. Het slaagpercentage over het gehele vo zou dan zo'n 8% lager gelegen hebben.
Als per centraal examen de N-term van 2012 zo gekozen zou zijn, dat de groep van 2012 hetzelfde gemiddelde cijfer behaalt als de groep 2011, met andere woorden als wij relatief zouden hebben genormeerd, zou het slaagpercentage over 2012 daadwerkelijk ongeveer 8% lager gelegen hebben dan in 2011. 'Ongeveer' omdat ook de schoolexamencijfers medebepalend zijn voor het slaagpercentage.
Dat het werkelijke slaagpercentage van 2012 slechts 0,9% lager was dan in 2011, kwam doordat absolute normering als uitgangspunt is gehanteerd bij de centrale examens. Op grond daarvan is geconstateerd dat de populatie van 2012 vaardiger was dan in 2011. De belangrijkste oorzaak van die toegenomen vaardigheid was waarschijnlijk dat docenten en leerlingen een tandje hebben bijgezet. De staatssecretaris heeft hen daarmee gecomplimenteerd in zijn brief waarmee hij de Examenmonitor VO 2012 aan de Tweede Kamer aanbood. Bij relatieve normering zou aan de betere prestatie van de populatie van 2012 geen recht gedaan zijn.

[6] Onder de eindcijfers voor de kernvakken Nederlands, Engels en wiskunde mag slechts één onvoldoende zijn en die onvoldoende mag geen vier of lager zijn.

3 De werkwijze bij absolute normering

Bij absolute normering moet de lat van jaar tot jaar op gelijke hoogte liggen. 'Hoe wordt die hoogte dan bepaald?' is daarbij een logische eerste vraag. Bij relatieve normering speelt die vraag niet, omdat de hoogte van de lat afhankelijk is van de vaardigheid van de populatie.

3.1 De hoogte van de lat

Wij bepalen de hoogte van de lat door per vak een referentie-examen aan te wijzen.

kader 3

Een *referentie-examen* is een centraal examen uit een eerder examenjaar dat gezien wordt als een goede operationalisering van de wat een leerling op grond van het examenprogramma en de syllabus [7] voor het desbetreffende vak moet kennen en kunnen.

Criteria voor een goede operationalisering zijn onder meer: foutloos, positieve reacties vanuit de docenten (rechtstreeks, via de quickscan met Wolf, via uitgebreide vragenlijsten achteraf, via de internetforums, de vakvereniging docenten), de leerlingen (via het LAKS) en – indien voorhanden – andere kanalen (zoals de media), evenwichtige representatie van leerinhouden en vaardigheden, toetstechnisch goed (goede spreiding van de scores, juiste lengte, juiste moeilijkheidsgraad, goede verdeling van makkelijke, gemiddelde en moeilijke opgaven, geen vragen die door goede leerlingen fout en door zwakke leerlingen goed gemaakt zijn, et cetera).

Bij absolute normering is het doel om het niveau van het referentie-examen te handhaven. Onder de hoogte van de lat wordt verstaan: het gevraagde beheersingsniveau om voor het referentie-examen (met de daarbij behorende N-term) een voldoende te behalen.

3.2 Niveauhandhaving

Examens verschillen in moeilijkheidsgraad, populaties kunnen verschillen in vaardigheid. Niveauhandhaving, hoe doe je dat?
Dat gebeurt doordat Cito, in opdracht van ons, equivaleringsmethoden, zoals pretest, posttest en anker in package, toepast.

kader 4

In een *pretest* worden combinaties van opgaven van een toekomstig centraal examen en een referentie-examen afgenomen.
Omdat de pretest-kandidaten in één zitting zowel opgaven van het toekomstige centraal examen als van het referentie-examen hebben afgelegd, kan Cito de moeilijkheidsgraad van deze beide examens vergelijken. Stel dat blijkt dat het gepreteste centraal examen 0,3 cijferpunt makkelijker blijkt dan het referentie-examen en het referentie-examen heeft een N-term van 1,1, dan zou het gepreteste examen een N-term krijgen van 0,8 (= 1,1 – 0,3).
In theorie kan de N-term van een gepretest centraal examen dus al bepaald worden vóórdat de afname heeft plaatsgevonden.
Door jaarlijks te pretesten ontstaat een cijfermatig beeld van eventuele veranderingen in de vaardigheid van examenkandidaten over de jaren heen:

[7] De syllabi voor de centrale examens worden door het CvTE bij regeling vastgesteld. Deze regeling treedt pas in werking na goedkeuring van de bewindspersoon.

- De hoogte van de lat is het vereiste beheersingsniveau om op het referentie-examen een voldoende te halen bij de aangewezen referentienorm, (in het bovenstaande voorbeeld is dat een N-term van 1,1).
- Bij een N-term van 0,8 is het vereiste beheersingsniveau voor het gepreteste examen hetzelfde. Daarom ligt de lat voor het gepreste examen en het referentie-examen op dezelfde hoogte als het gepreteste examen een N-term van 0,8 krijgt.
- Stel dat het referentie-examen in 2010 is afgenomen en het gemiddelde cijfer daarvoor 6,2 was. Stel bovendien dat het gepreteste examen in 2016 wordt afgenomen en dat na de afname blijkt dat het gemiddelde cijfer een 6,4 is, dan kan gesteld worden dat de populatie van 2016 0,2 cijferpunt vaardiger is dan de groep van 2010.
- Van ieder examenjaar waarbij gepretest is, kan zo het vaardigheidsheidsverschil met de populatie die het referentie-examen heeft afgelegd worden bepaald.

Een *posttest* werkt net als een pretest, maar vindt kort *na* het centraal examen plaats. Analoog aan de pretest signaleert de posttest vaardigheidsverschillen en kan de N-term van het geposttеste examen worden afgeleid van de N-term van het referentie-examen.

Anker in package wordt gebruikt bij de flexibele en digitale centrale examens in de basis- en kaderberoepsgerichte leerweg (bb en kb). Verspreid over de verschillende digitale varianten (de package) die er van deze CE's zijn, komen geheim gehouden opgaven uit eerdere examenjaren (het anker) voor.
Ook bij anker in package ontstaat een beeld van eventuele vaardigheidsverschillen van groepen kandidaten over de jaren heen. Bij anker in package kunnen de N-termen van de actuele varianten van het digitale CE worden afgeleid uit de N-term van ankerexamen, die op zijn beurt weer is afgeleid uit de N-term van het referentie-examen.

Centrale examens mét equivalering
Niveauhandhaving kan worden gerealiseerd bij alle centraal examens met equivalering, dat wil zeggen pretest, posttest of anker in package.
De equivaleringsmethoden zijn bewerkelijk (zie kader 4) en uiteraard brengen ze kosten mee.
Daarom worden niet alle centrale examens geëquivaleerd.

kader 5

Een pretest heeft nogal wat voeten in de aarde. Zo moet iedere opgave van zowel het gepreteste examen door minimaal 200 kandidaten zijn gemaakt. Hetzelfde geldt voor de opgaven uit het referentie-examen, waaraan het gepreteste examen wordt geankerd. Voor de kandidaten die aan de pretest meedoen, moet sprake zijn van een serieuze examensituatie: het resultaat dat zij voor de toets behalen weegt mee in hun schoolexamen. Bovendien moet zeker gesteld worden dat de kandidaten niet al kennis hebben genomen van het referentie-examen. Als de kans groot is dat zij het referentie-examen wel al gezien hebben, wordt niet het referentie-examen zelf, maar een ander examen aan het pretest-examen geankerd. Dat andere examen moet op zijn beurt dan weer eerder met het referentie-examen zijn geankerd. Via een tussenstap kan het gepreteste examen dan toch geankerd worden aan het referentie-examen.

Het bovenstaande geldt ook voor een posttest.

Een pretest kent daarnaast altijd een geheimhoudingsrisico. Dat is onvermijdelijk omdat de pretest-afname enkele jaren eerder plaats vindt. Bij een posttest speelt dit niet, maar daar geldt weer dat de omlooptijd erg kort is, omdat de posttest moet plaatsvinden tussen de

afname en de normering. Voor een posttest komen daarom alleen vakken in aanmerking met veel gesloten vragen (en dus weinig correctietijd) en die vakken moeten ook nog eens voor in het examenrooster worden geplaatst.

Voor anker in package moet per opnieuw te gebruiken opgave zorgvuldig worden bekeken of deze niet verouderd is.

Centrale examens zonder equivalering
Wij zouden het liefst zien dat alle centrale examens geëquivaleerd worden, maar zoals hierboven is aangegeven, is dat niet haalbaar. Vóór 2012, het eerste jaar van de aangescherpte uitslagregels, gingen wij er bij centrale examens zonder equivalering vanuit dat de groep kandidaten even vaardig was als de populatie die het referentie-examen had afgelegd.
Bij onveranderde condities vinden wij dat op zichzelf een valide aanname. Er is immers weinig reden om te veronderstellen dat de leerlingen bij bijvoorbeeld geschiedenis havo van het ene op het andere jaar veel beter of veel zwakker worden. Dat kan op een bepaalde school misschien gebeuren, maar niet als landelijk effect. Het bepalen van de N-term bij geschiedenis havo gebeurde op dezelfde wijze als hierboven in 2.1 is aangegeven. In stap 1 wordt met behulp van de normeringstabel bepaald bij welke N-term het percentage voldoende en het gemiddeld cijfer het dichtst bij het percentage voldoende en het gemiddeld cijfer van het referentie-examen liggen. De stappen 2 en 3 zijn letterlijk hetzelfde als onder 2.1 beschreven. De aanname dat de populatie voor geschiedenis havo even vaardig is als de populatie die het referentie-examen heeft afgelegd, wordt dus geoperationaliseerd door te stellen dat de groep geschiedenis havo van jaar x een percentage voldoende en een gemiddeld cijfer moet behalen, dat ongeveer gelijk zijn aan die van de referentiepopulatie uit een eerder examenjaar. Door de aanscherping van de uitslagregels was vanaf 2012 echter niet langer sprake van onveranderde examencondities.

De Fisher-methode
Voor de centrale examens zonder equivalering is toen als volgt een schatting gemaakt van het vaardigheidsverschil met 2011:
Als de cijfers laten zien dat de havo-leerlingen van 2012 voor de centrale examens mét equivalering, te weten Engels, Duits, Frans, aardrijkskunde, management en organisatie, wiskunde A, wiskunde B, natuurkunde, scheikunde [8], vaardiger zijn dan de groepen van 2011, is het niet aannemelijk dat dat voor centrale examens zonder equivalering, zoals Spaans-havo, geschiedenis-havo en biologie-havo niet geldt.
De psychometrici van Cito hebben berekend dat het meest waarschijnlijke vaardigheidsverschil tussen de havo-populaties van 2012 en 2011 +0,1 bedroeg. Voor het vwo was dit +0,1, voor de algemene vakken in het vmbo +0,2 en voor de beroepsgerichte vakken +0,1.
De methode die Cito hiervoor gebruikt staat in de vakliteratuur van de psychometrie bekend als Fisher's combined probability test, kortweg de Fisher-methode.
Wij hebben deze uitkomsten overgenomen als geconstateerde vaardigheidsverschillen tussen de populaties 2012 en 2011.
Voor geschiedenis havo bijvoorbeeld, een centraal examen zonder equivalering, is in 2012 de N-term van het centraal examen als volgt bepaald:

[8] Behalve pretest, posttest en anker in package is ook standard setting gebruikt.
standard setting: experts hebben de moeilijkheid van het CE 2012 vergeleken met het referentie-examen (of een daarmee vergelijkbare aanpak)

In stap 1 is met behulp van de normeringstabel bepaald bij welke N-term het percentage voldoende en het gemiddeld cijfer het dichtst bij het percentage voldoende en het gemiddeld cijfer van het referentie-examen liggen. Bij die N-term is 0,1 opgeteld, omdat +0,1 als vaardigheidsverschil uit de Fisher-methode kwam. Als gevolg van de Fisher-methode was het technisch normeringsadvies dus 0,1 hoger, dan wanneer geschiedenis havo genormeerd was op basis van de aanname van gelijke populaties. Het gevolg is ook dat het gemiddeld cijfer 0,1 hoger ligt en dat ook het percentage voldoende hoger is.

Analoog zijn de populaties van 2013, 2014, 2015 en 2016 vergeleken met de groep van 2011, de populatie van het laatste jaar vóór de nieuwe uitslagregels, het laatste jaar waarin de aanname van gelijke vaardigheden nog valide was.
De geconstateerde vaardigheidsverschillen met de populatie van 2011 zijn weergegeven in de onderstaande tabel.

	vaardigheidsverschil populatie [examenjaar …] t.o.v. ≤ 2011				
cluster	2016	2015	2014	2013	2012
beroepsgerichte vakken	+0,1	+0,1	+ 0,1	+ 0,1	+ 0,1
algemene vakken bb	+0,4	+0,4	+ 0,3	+ 0,3	+ 0,2
algemene vakken kb	+0,1	+0,0	+ 0,1	+ 0,1	+ 0,2
algemene vakken gl/tl	+0,2	+0,2	+ 0,2	+ 0,2	+ 0,2
havo niet-kernvakken	+0,1	+0,0	+ 0,0	+ 0,3	+ 0,1*
havo-kernvakken Eng, wisB	+0,6	+0,7	+ 0,3	+ 0,6	
havo-kernvakken Ned, wisA	+0,1	+0,0			
vwo niet-kernvakken	+0,1	+0,2	+ 0,0	+ 0,2	+ 0,1*
vwo-kernvakken Eng, wisA, wisB	+0,6	+0,7	+ 0,3	+ 0,6	
vwo-kernvakken Ned, wisC	+0,1	+0,2			

* Omdat de kernvakkenregeling in 2012 nog niet van kracht was, betreft het hier alle vakken.

3.3 Nogmaals de drie stappen

Bij absolute normering worden dezelfde drie stappen doorlopen als bij relatieve normering:

- een technisch normeringsadvies van Cito;
- een advies van iedere CvTE-vaststellingscommissie;
- de vaststelling van de N-term

Stap 1, het technisch normeringsadvies, is aanmerkelijk ingewikkelder dan bij relatieve normering.
Bij de centrale examens mét equivalering wordt de N-term van het technisch normeringsadvies afgeleid van de N-term van het referentie-examen. Dat gebeurt op de manier zoals in 3.2 beschreven is in kader 3.
Bij centrale examens zonder equivalering wordt de N-term bepaald, nadat met behulp van de Fisher methode is bepaald in hoeverre er sprake is van een vaardigheidsverschil

met de populatie van 2011. Dat gebeurt op de manier zoals in 3.2 beschreven is voor geschiedenis havo.

De beschrijving van stap 2, het advies van de CvTE-vaststellingscommissie, is identiek aan die voor stap 2 bij relatieve normering (zie 2.1), zij het dat het technisch normeringsadvies, dat voor de vaststellingscommissie het uitgangspunt vormt, anders tot stand gekomen is. Ook bij absolute normering worden alle individuele opgaven tegen het licht van de toets- en itemanalyse gehouden, weegt de vaststellingscommissie de kwalitatieve reacties en kijkt zij nogmaals of er geen onvolkomenheden aan het licht komen waarvan kandidaten niet de dupe mogen zijn.

Stap 3: de vaststelling van de N-term
Ook bij stap 3 is er geen verschil tussen absoluut en relatief normeren. Het is de CvTE-leiding die de N-termen vaststelt. Voordien heeft er overleg plaatsgevonden met het LAKS.

De stappen 2 en 3 verschillen niet van de werkwijze bij relatieve normering, maar in stap 1 is er een fundamenteel verschil tussen absolute en relatieve normering. Voor ons maakt stap 1 bij absolute normering het mogelijk om voor ieder centraal examen een passende norm vast te stellen.
In onze visie is dat een norm die recht doet aan de prestatie-eisen die gelden voor het vak, zoals neergelegd in het examenprogramma, de syllabus en het referentie-examen, en waarbij de kandidaten het cijfer krijgen waarop zij recht hebben.

4 Relatieve of absolute normering?

Wij hanteren absolute normering als uitgangspunt, omdat de 5,5 van het ene jaar ook de 5,5 van het andere jaar hoort te zijn. Het moet niet zo zijn dat de je als examenkandidaat meer moet doen voor een 5,5 dan de leerlingen uit een vorig jaar, omdat jouw lichting vaardiger is dan de vorige lichtingen. Of minder hoeft te doen omdat jouw jaargang zwakker is. Ook dat is onwenselijk.
Wij nemen daarbij voor lief dat absolute normering moeilijker is uit te leggen, maar hoopt dat de bovenstaande uiteenzetting bijdraagt begrip van en voor de werkwijze bij de normering van de centrale examens vo.

College voor Toetsen en Examens
info@hetcvte.nl

www.hetcvte.nl