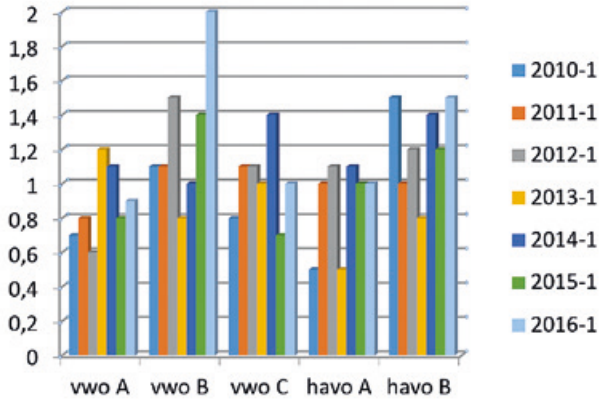
N-TERM IN WERKING

CvTE
Jacqueline Wooning

Elk jaar wachten docenten en leerlingen met spanning af wat de uiteindelijke N-term zal zijn. Op fora en in blogs wordt daar vaak al een voorschot op genomen: 'het examen was moeilijk, dus het zal wel een hoge N-term worden', of 'mijn leerlingen hebben hoog gescoord, ik vrees voor een lage N-term.' Andere uitspraken zijn soms gebaseerd op het idee dat examens in moeilijkheid variëren en daarmee de door de kandidaten behaalde cijfers ook. Hier is sprake van een misvatting: uiteraard zijn niet alle examens precies even moeilijk, maar bij het bepalen van de cijfers wordt gecompenseerd voor de geconstateerde moeilijkheidsverschillen. Om dit te realiseren spelen statistische technieken en de N-term een belangrijke rol. Hoe dit werkt, wordt in dit artikel besproken. We besteden daarbij speciaal aandacht aan de totstandkoming van de N-termen bij vwo wiskunde B en vwo wiskunde C.

N-term

Een normeringsterm of N-term is een getal waarmee voor alle mogelijke scores voor een centraal examen bepaald wordt met welk cijfer die score gewaardeerd wordt.
De N-termen fluctueren van jaar tot jaar zoals in figuur 1 is te zien. Een hogere of lagere N-term betekent niet dat het examen slechter of beter is, alleen dat het relatief moeilijk of makkelijk is.



figuur 1 N-termen CE wiskunde 1e tijdvak

Van score naar cijfer

Bij het omzetten van een score S naar een cijfer C wordt vaak gebruik gemaakt van de formule
$C = 9 \cdot (S / L) + N$ (waarbij L de schaallengte is, S de behaalde score en N de normeringsterm). Voor $N = 1$ krijgt een kandidaat een 5,5 als hij 50% van de scorepunten behaalt: de cesuur is 50%. Door N *hoger* of *lager* te kiezen, wordt deze cesuur *lager* respectievelijk *hoger* dan 50%. Om te voorkomen dat de maximale score leidt tot een cijfer boven de 10 (als $N > 1{,}0$) en een score van 0 tot een cijfer lager dan 1 (als $N < 1{,}0$), kun je de formule verfijnen.[1]

Relatief normeren

Bij relatief normeren wordt de N-term zo bepaald dat het percentage voldoendes (bijvoorbeeld 80%) of het gewenste gemiddelde cijfer (bijvoorbeeld 6,3) door de jaren heen steeds gelijk blijft. Als relatieve normering toegepast wordt bij eindexamens, heeft dit tot gevolg dat er geen rekening gehouden wordt met mogelijke verschillen tussen populaties eindexamenkandidaten over de jaren heen. Gezien de grootte van de populaties eindexamenkandidaten is het ook heel aannemelijk dat deze van jaar tot jaar even vaardig zijn.
Als echter de populatie van 2015 toch vaardiger is dan die van 2011, is het dus mogelijk dat kandidaten in 2015 geen diploma krijgen, terwijl ze dat in 2011 wel zouden hebben gekregen. Bij relatief normeren wordt de cesuur dus bepaald in samenhang met de vaardigheid van de populatie eindexamenkandidaten.

Absoluut normeren

Bij absoluut normeren worden elk jaar dezelfde eisen aan de kandidaten gesteld. Het toegekende cijfer is dan dus alleen afhankelijk van de door de kandidaat geleverde prestatie en onafhankelijk van eventuele verschillen in vaardigheid van de gehele populatie eindexamenkandidaten ten opzichte van voorgaande jaren. Omdat in Nederland is afgesproken een normeringssystematiek te hanteren, waarbij de gelijkwaardigheid van diploma's over de jaren heen uitgangspunt is, worden de examens zoveel mogelijk absoluut genormeerd. Het College voor Toetsen en Examens (CvTE) heeft de taak om per centraal

examen 'de lat' van jaar tot jaar even hoog te leggen, zodat de prestatie die een kandidaat moet leveren om een voldoende te halen steeds gelijk is. Absoluut normeren vereist dat de moeilijkheidsgraad van opeenvolgende examens moet kunnen worden vergeleken, zodat bij de normering voor verschillen in moeilijkheid kan worden gecorrigeerd. Men spreekt dan van equivalering. Bij vakken die met een nieuw programma starten (zoals in 2017 havo wiskunde A en B) of waarvan het aantal deelnemers erg laag is (zoals bij vwo wiskunde C) kan equivalering niet worden toegepast. In dergelijke gevallen wordt relatief genormeerd.

A Absoluut normeren bij examens met equivalering

Fase 1: equivalering

De eerste fase van het vaststellen van de N-term vindt bij deze examens al plaats tijdens het constructieproces. Voor wiskunde A en B worden combinaties van toekomstige opgaven en die van een referentie-examen[2] ruim tevoren *gepretest* bij kandidaten uit (voor)examenklassen. Doel is de moeilijkheidsgraad van beide examens te vergelijken en vóór afname van het nieuwe examen een inschatting te maken van de te verwachten N-term. Stel dat met behulp van berekeningen na afloop van de *pretest* het nieuwe examen 0,3 cijferpunt moeilijker wordt geschat dan het referentie-examen en het referentie-examen een N-term van 1,5 heeft, dan is de eerste schatting van de N-term van het toekomstige examen: 1,5 + 0,3 = 1,8.
Als de eindexamenpopulatie van het toekomstige examen even vaardig is als die van het referentie-examen, zal de N-term van 1,8 bij het toekomstige examen tot een gelijk gemiddelde en een even hoog percentage voldoendes leiden als de N-term 1,5 bij het referentie-examen: de lat bij deze twee examens ligt op gelijke hoogte. Mocht de populatie echter vaardiger zijn, dan zal de N-term van 1,8 tot een hoger percentage voldoendes leiden.

Fase 2: bepaling N-tech

Na afname van elk nieuw examen maakt Cito een toets- en itemanalyse (TIA)[3] op basis van de scores die de correctoren hebben toegekend (Wolf). Daaruit blijkt hoe de scorepunten gespreid zijn en hoe betrouwbaar het examen als meetinstrument was. Ook kan de TIA een indicatie geven of kandidaten in tijdnood zijn gekomen en welke vragen een goed onderscheidend vermogen hebben.
Met behulp van de resultaten van de *pretest* en de ingezonden scores van de afname wordt een nieuwe schatting gemaakt van de N-term. Deze noemen we N_{pt} (pt staat voor *pretest*). Hierbij worden de afnamegegevens vooral gebruikt om de relatieve moeilijkheid van de vragen van het nieuwe examen ten opzichte van elkaar nauwkeuriger te kunnen schatten. Bij het schatten van de N-term, N_{pt}, is uiteraard sprake van een zekere mate van onnauwkeurigheid. Het 90%-betrouwbaarheidsinterval voor de berekende N-term heeft, afhankelijk van onder andere de omvang van de *pretest*, vaak een breedte van 0,3 tot 0,7. In ons voorbeeld zou bij een breedte van het betrouwbaarheidsinterval van 0,4 het 90%-betrouwbaarheidsinterval van N_{pt} dus zijn: BI = [1,6 ; 2,0]. Op basis van de TIA wordt, rekening houdend met onnauwkeurigheden, vervolgens statistisch getoetst of de aanname dat de populaties (van het referentie-examen en het huidige examen) even vaardig zijn, houdbaar blijft (met een hypothesetoets). Hiertoe wordt gezocht naar de N-term waarmee bij het nieuwe examen het percentage voldoendes en het gemiddeld behaalde cijfer zoveel mogelijk overeenkomen met die van het referentie-examen (relatief normeren). Deze N-term wordt aangeduid met N_{gp}, waarbij gp staat voor gelijke populaties.
Als N_{gp} binnen het betrouwbaarheidsinterval BI van N_{pt} ligt, is het aannemelijk dat de nieuwe eindexamenpopulatie even vaardig is als de referentiepopulatie en is de meest waarschijnlijke N-term gelijk aan N_{gp}. Deze N-term wordt aangeduid als N-tech.
Mocht echter blijken dat je niet kunt uitgaan van even vaardige populaties (N_{gp} ligt dan buiten BI), dan is er een statistisch algoritme waarmee het meest waarschijnlijke vaardigheidsverschil tussen beide populaties wordt bepaald en daarmee ook de technische N-term (N-tech). Dat algoritme werkt als volgt:

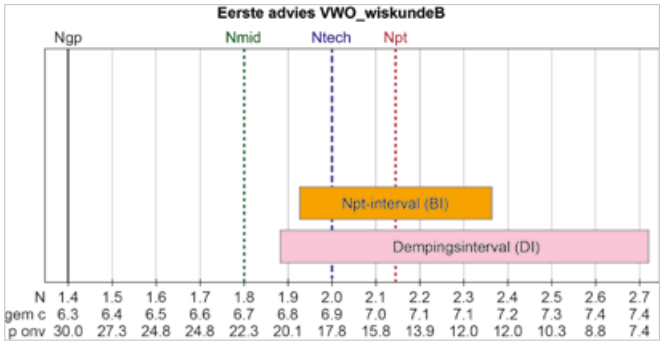
a Als eerste wordt gekeken of het gemiddelde N_{mid} van N_{gp} en N_{pt} binnen het betrouwbaarheidsinterval ligt. Is dat het geval, dan is N-tech = N_{mid}.
b Als ook N_{mid} buiten BI ligt, dan is N-tech de N-waarde die net binnen het betrouwbaarheidsinterval BI ligt aan de kant van N_{gp}.
c Tot slot wordt getoetst of de gemiddelde cijfers van opeenvolgende populaties (onder gelijke condities) niet onwaarschijnlijk veel van elkaar verschillen. De aanname die hierbij gehanteerd wordt is dat het gemiddeld behaalde cijfer van de huidige populatie niet zomaar meer dan 0,3 cijferpunt mag verschillen van die van de populatie van het voorgaande jaar. Als dit wel het geval is, wordt het gemeten verschil gedempt, waarbij gebruik wordt gemaakt van een dempingsinterval. De volgende werkwijze wordt hierbij gehanteerd: als rand BI binnen het dempingsinterval ligt, dan is N-tech = rand BI. Zo niet, dan wordt de N-term 'gedempt'.

In kader 1 wordt deze procedure toegelicht voor het examen vwo wiskunde B 2016-I.

kader 1

Fase 1 en 2 vwo wiskunde B 2016-I
Bij het referentie-examen behaalde 70% van de kandidaten een voldoende. Bij de aanname dat de populaties eindexamenkandidaten van 2016 en die van het referentiejaar even vaardig zijn, zou dit voor het examen wiskunde B van 2016 een N-term 1,4 opleveren met een bijbehorend percentage voldoendes van 70%:

N_{gp} = 1,4.
Op basis van de *pretest* is de volgende N-term vastgesteld: N_{pt} = 2,145 met een bijbehorend betrouwbaarheidsinterval BI = [2,0 ; 2,3].
N_{gp} ligt buiten dit interval dus de aanname van gelijke populaties wordt verworpen. N_{mid} = 1,8 ligt niet in BI dus N-tech wordt de waarde die nét binnen BI ligt: N-tech = 2,0.
Het gemiddelde cijfer bij N = 2,0 is 6,9 en wijkt niet meer dan 0,3 af van het examen van 2015 (7,1). Er is dus geen demping nodig.
Resultaat fase 1 en 2: N-tech = 2,0. Zie figuur 2.



figuur 2 Bepaling N-tech vwo wiskunde B 2016-I

Fase 3: vaststelling definitieve N-term
De dag vóór de N-termen bekend worden gemaakt, vindt de normeringsvergadering plaats, waar de definitieve N-termen worden vastgesteld. De vaststellingscommissie, die het examen heeft vastgesteld, formuleert een advies voor de definitieve N-term op basis van de TIA, daarbij meewegend de meldingen die via de examenlijn zijn binnengekomen, waaronder die van het LAKS, de door docenten ingevulde Quickscan, eventuele aanvullingen en de bevindingen uit de centrale examenbesprekingen. Op basis van dit advies stellen de directeur en sectormanagers van het CvTE, alle overwegingen gewogen hebbend, de definitieve N-term vast: N-def.

kader 2

Fase 3 vwo wiskunde B 2016
Er is geen reden gevonden om af te wijken van N-tech. De definitieve N-term wordt dus 2,0.

2,0 is hoog ...
Als er niet absoluut genormeerd was, maar uitgegaan zou zijn van even vaardige populaties, zou voor vwo wiskunde B een N = 1,4 zijn vastgesteld. Daarmee zou de vaardigheidstoename (zoals berekend op basis van de *pretest*) niet in de cijfers zichtbaar zijn en zouden de kandidaten tekort zijn gedaan. De N-term van 2,0 duidt dus op een moeilijk examen.
Daarbij speelt ook de lengte een rol: uit de TIA blijkt dat de beschikbare tijd krap was. Middels nadere analyse door Cito en CvTE probeert men duidelijkheid te krijgen over wat dit betekent voor de constructie van toekomstige examens. Ondanks het feit dat elk examen met de grootst mogelijke zorgvuldigheid wordt samengesteld, betekent dit dat de toekomstige examens nogmaals tegen het licht zijn en worden gehouden. Voor 2017 heeft dit voor vwo wiskunde B geleid tot een kortere versie van het examen dan gebruikelijk (zie ook de *Maartaanvulling* op examenblad.nl). Ook het pilotexamen vwo wiskunde B werd als lang ervaren. Pilotdocenten gaven aan dat het examen goed paste bij het onderwijs, maar dat het sterk denkactieve karakter van dit examen eraan bijdroeg dat het een moeilijk examen was. Vanwege het experimentele karakter kon dit examen niet absoluut genormeerd worden. Ook voor dit soort situaties worden vaste procedures gehanteerd, waarbij analyses van de resultaten van de overlapvragen met het reguliere examen in ogenschouw worden genomen, evenals de resultaten van de pilotexamens van voorgaande jaren en het geringe aantal van slechts 107 kandidaten. De gehanteerde procedure wees uit dat met een N-term van 2,4 de kandidaten recht werd gedaan. Alhoewel het ongebruikelijk is om een N-term vast te stellen die boven de 2,0 ligt, is hier wel voor gekozen: kandidaten mogen niet de dupe worden van het star vasthouden aan een richtlijn. Een dergelijke situatie zal zich overigens eerder voordoen in pilotsituaties dan bij reguliere programma's, waarmee de examenmakers veel langer ervaring hebben.

B Examens zonder equivalering en nieuwe programma's

Nieuwe programma's en vwo wiskunde C
Absoluut normeren is voor de reguliere examens havo 2017 en vwo 2018 niet mogelijk: het nieuwe programma wordt voor het eerst geëxamineerd, zodat het niet mogelijk is geweest de eindexamenopgaven te *pretesten* bij eindexamenkandidaten in eerdere jaren en equivalering toe te passen. Door het geringe aantal wiskunde C-kandidaten, is bij vwo wiskunde C absoluut normeren ook niet mogelijk, omdat aan de *pretest*-resultaten geen echt betrouwbare statistische conclusies ontleend kunnen worden. Het *pretesten* van wiskunde C-opgaven geeft vooral ondersteuning aan het werk van de constructiegroep en de vaststellingscommissie. Bij deze examens gaat men er in eerste instantie van uit dat de huidige populatie en de referentiepopulatie even vaardig zijn. Door middel van relatief normeren wordt de N-term in de eerste fase bepaald.

Fase 1
Het percentage voldoendes (bijvoorbeeld 80%) vormt het uitgangspunt.
Met behulp van een normeringstabel wordt nagegaan bij welke N-term voor het huidige examen het percentage

voldoendes zo dicht mogelijk bij dat van het referentie-examen ligt.

kader 3

Fase 1 vwo wiskunde C 2016
Bij het referentie-examen was het percentage voldoende 74.
Bij een N-term van 0,9 is voor het examen van 2016 het percentage voldoende 75. Deze waarde ligt het dichtst bij het percentage van het referentie-examen.
$N_{gp} = 0{,}9$ met een gemiddeld cijfer van 6,2.

Fase 2
Jaarlijks wordt onderzocht of er sprake is van een vaardigheidsverschil tussen de populatie uit dat jaar en die van 2011, het laatste jaar voor de invoering van de nieuwe uitslagregels (5,5 gemiddeld op het CE in 2012 en kernvakkenregel in 2013). Als de populatie voor de vakken mét equivalering 'gemiddeld' vaardiger is geworden, ligt het voor de hand dat dit ook het geval is voor de vakken zonder equivalering. De methode die gebruikt wordt voor de bepaling van de vaardigheidsverandering heet de Fisher methode (Fisher's combined probability test). De generieke vaardigheidsverandering[4] wordt verdisconteerd in de N-term.

kader 4

Fase 2 vwo wiskunde C 2016
Wiskunde C vormt een groep met twee kernvakken waar de vaardigheidsontwikkeling gelijk blijkt te zijn aan die van de groep niet-kernvakken.
Het vaardigheidsverschil voor de niet-kernvakken bij de populatie van 2016 t.o.v. 2011 is + 0,1
N-tech = $N_{gp} + 0{,}1 = 0{,}9 + 0{,}1 = 1{,}0$.

Fase 3
Fase 3 is hetzelfde als die van de examens met equivalering.

kader 5

Fase 3 vwo wiskunde C 2016
Er is geen reden gevonden om af te wijken van N-tech. De definitieve N-term wordt dus 1,0. Dit levert een gemiddeld cijfer van 6,3 met 75% voldoende op.

De systematiek van het normeren is niet uitputtend behandeld. Hopelijk geeft dit artikel wel iets meer duidelijkheid over de normering van examens en de werkwijze die het mogelijk maakt te compenseren voor de onontkoombare variatie in moeilijkheid van de examens.

Noten

[1] Zie *www.toetswijzer.nl/html/normering/CvEmethode.pdf*.
[2] Een referentie-examen is een reeds afgenomen examen dat gezien wordt als goede operationalisatie van het examenprogramma en de syllabus. Dit is mede gebaseerd op de afnamegegevens en het oordeel van docenten.
[3] De TIA van de examens 2016 zijn te vinden op *www.cito.nl/onderwijs/voortgezet%20onderwijs/centrale_examens/schriftelijke_examens_havovwo/examens_havovwo_2016*. Kies in het menu: havo of vwo en het tijdvak. Kies het document in de kolom TIA.
[4] Er wordt onderscheid gemaakt tussen de vaardigheidsverandering bij verschillende groepen van vakken. Voor centrale examens zonder equivalering maakt men een schatting van het vaardigheidsverschil met 2011. Uitgangspunt hierbij is dat het aannemelijk is dat populaties die vaardiger zijn op de centrale examens met equivalering (voor havo: Engels, Duits, Frans, aardrijkskunde, management en organisatie, wiskunde A, wiskunde B, natuurkunde en scheikunde) dan de populatie van 2011, ook vaardiger zullen zijn op de centrale examens zonder equivalering. Via de Fishermethode kan berekend worden wat het meest waarschijnlijke vaardigheidsverschil is.

Webinar

In maart 2017 heeft het webinar 'Hoe komt de N-term tot stand' plaatsgevonden. Dit webinar is via examenblad.nl te volgen (zoek op 'webinar n-term').

Over de auteur

Jacqueline Wooning is clustermanager exacte vakken havo/vwo bij het College voor Toetsen en Examens. Dit artikel is tot stand gekomen in samenwerking met de voorzitters van de vaststellingscommissies wiskunde AC en wiskunde B.
E-mailadres: *info@cvte.nl*