Paul van der Molen

Normering 2021: Anders dan anders

In *Euclides* 92-6 is in het artikel 'N-term in werking' uitgelegd hoe de N-term tot stand komt. Die werkwijze kon in het bijzondere examenjaar 2021 niet gehanteerd worden. Paul van der Molen legt uit hoe dit jaar de N-termen zijn vastgesteld.

Inleiding

De normering van de centrale examens in 2021 is anders verlopen dan in het verleden gebruikelijk was. In de aanloop naar de centrale examens heeft de minister besloten dat leerlingen een extra herkansing kregen en de resultaten van een vak mochten wegstrepen. Deze maatregelen kwamen voort uit een behoefte om rekening te kunnen houden met de ongewone, bij sommige leerlingen gebrekkige, voorbereiding op de centrale examens. Bij het nemen van deze extra maatregelen heeft de minister ook gesteld dat de eisen bij elk vak zoveel mogelijk gehandhaafd moesten worden. De normering moest dus op zo'n manier worden uitgevoerd dat de norm uit het verleden, gegeven de omstandigheden, zo goed mogelijk kon worden gehandhaafd. Dit artikel vertelt hoe de N-termen bij het vak havo wiskunde A tot stand zijn gekomen. De andere wiskundevakken hebben hetzelfde proces doorlopen.

Uitzonderlijke situatie

De toevoeging 'gegeven de omstandigheden', zoals hierboven vermeld, laat al doorschemeren dat we in 2021 te maken hadden met een uitzonderlijke situatie. Er is een aantal redenen waarom de werkwijze uit het verleden dit jaar niet goed paste.

Om te beginnen zou in 2021 de populatie die in mei examen zou doen wel eens minder representatief kunnen zijn voor de hele populatie dan in voorgaande jaren.

Dit zou met name opgaan wanneer substantieel minder leerlingen, of minder scholen, aan het eerste tijdvak zouden deelnemen. De examenpopulatie van 2021 zou daardoor niet goed te vergelijken zijn met de populaties in eerdere jaren.

Ook kunnen we in 2021 niet voor alle vakken onze normhandhavingsinstrumenten zoals pre- of posttesten inzetten. De reden hiervoor is dat deze instrumenten er last van hebben als onderdelen van het examenprogramma niet in dezelfde mate aandacht in de voorbereiding hebben gehad als in voorgaande jaren, met name als één onderdeel relatief minder aandacht heeft gehad. Ook als de vaardigheidsontwikkeling bij vakken sterk verschilt ten opzichte van andere vakken, en die verschillen zijn niet in lijn met eerdere jaren, dan geeft dat problemen bij de normhandhaving. Deze factoren veroorzaken een grotere onzekerheid in de uitkomsten.

De basis van de normering

In de winter en het voorjaar hebben normeringsspecialisten van Cito en CvTE beschreven welke informatie beschikbaar zou moeten zijn tijdens de normering en hoe deze gebruikt zou worden. Hiermee werd de basis gelegd voor de normering van 2021. Deze basis is in januari in een docent-informatiewebinar gepresenteerd. Zie *Examenblad.nl*[1], waar je meer gedetailleerde informatie vindt over de normering inclusief een informatieve animatie.

De basis voor de normering bestond uit vier stappen: in stap 1 werd de voorlopige technische N-term zodanig bepaald dat een 'representatieve steekproef' een vergelijkbaar gemiddeld cijfer kreeg als in de jaren 2014 - 2019.

In stap 2 werd het oordeel van docenten gebruikt om na te gaan of de norm uit stap 1 wel goed overeenkomt met de norm uit het verleden. Docenten geven niet allemaal hetzelfde oordeel. Daarom werden de docenten in drie gelijke groepen verdeeld en was het interval van de middelste groep leidend voor de vergelijking in stap 2 (er werd dus gewerkt met het 33/67-percentiel-interval).

Ten slotte werd in stap 3 vergeleken of de voorlopige N-term wel in lijn was met de gebruikelijke moeilijkheid van het examen van dat vak (historische N-term). De gedachte hierachter is dat het erg onwaarschijnlijk is dat de moeilijkheid van het examen in 2021 opeens erg ver afwijkt van deze waarden. Daartoe is een 90% betrouwbaarheidsinterval gemaakt op basis van het gemiddelde en de standaarddeviatie van de N-termen in de afgelopen zes

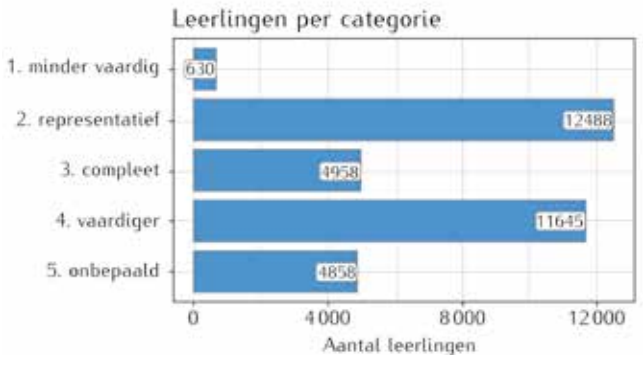
jaar. Bij vakken waar de voorlopige N-term na stap 2 buiten het historisch N-termeninterval ligt, werd de N-term aangepast tot in dit historisch interval. De N-term mocht daarbij maximaal worden opgeschoven tot de grenzen van het 10/90-percentiel-interval van de docentoordelen. Tot slot werd in stap 4 nog gekeken of er sprake was van fouten of andere onvolkomenheden waarvoor compensatie via de N-term nodig was.

De docentvragen

Op voorhand was niet duidelijk welke leerlingen in mei examen zouden doen. Stel dat van elke klas de zwakste leerlingen pas in juni voor het eerst examen zouden doen, dan is de mei-populatie sterker dan het landelijk gemiddelde. Daarom hebben we docenten via het programma Wolf de vraag gesteld of de groep die in mei examen had gedaan, wel representatief was voor de hele klas.
De andere vraag die we via Wolf aan de docenten voorlegden was: '... Cito en CvTE willen graag weten waar volgens u de grens tussen voldoende en onvoldoende prestatie ligt op dit examen, wanneer u het vergelijkt met andere jaren. Bij welke totaalscore past dan volgens u op dit examen het cijfer 5,5?'. Hiermee probeerden we een link met de norm in het verleden te maken. De score die de docent had doorgegeven hebben we omgerekend naar een N-term. Uit de resultaten werd duidelijk dat sommige docenten het examen makkelijker inschatten dan andere docenten. Omdat we (zeker bij wiskunde) van honderden docenten het oordeel hebben ontvangen, ontstond een goed beeld van de verdeling van de oordelen.

Tijdvak 1

In Wolf hebben 501 scholen de gegevens van 34 579 leerlingen doorgegeven die het examen havo wiskunde A gemaakt hadden. Volgens de vraag over representativiteit zaten 12 488 leerlingen in een representatieve groep en zaten 4 958 leerlingen in een klas die voor 100% deelnam in mei, zie figuur 1. Er werden dus 17 446 leerlingen in de steekproef opgenomen.



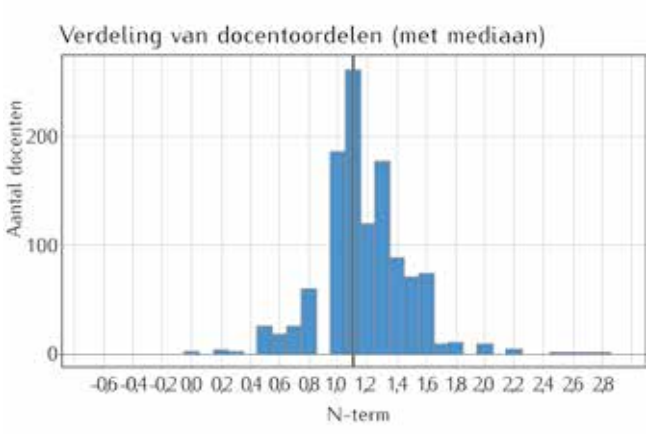
figuur 1

De 17 446 leerlingen zaten op scholen die in het verleden gemiddeld even vaardig waren als het landelijk gemiddelde. Het gemiddelde cijfer 2017 – 2019[2] was een 6,4 volgens een intern afgesproken berekeningswijze. Met behulp van een zogenaamde normeringstabel kan gevonden worden bij welke N-term het gemiddeld cijfer een 6,4 is, zie tabel1.

N-term	Gemiddeld cijfer	Percentage onvoldoende
0,8	6,2	27,8
0,9	6,3	25,2
1,0	6,4	25,2
1,1	6,5	22,8
1,2	6,6	20,5

tabel 1

Hieruit volgt dat na stap 1 de voorlopige technische N-term een 1,0 was. Vervolgens werd gekeken of dit overeenkwam met de docentoordelen. De verdeling van de docentoordelen is weergegeven in figuur 2.

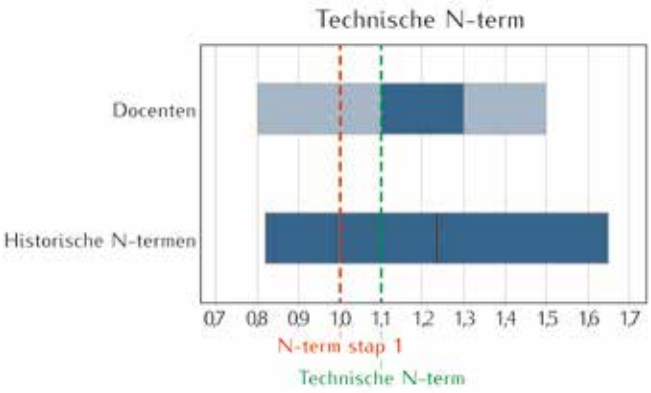


figuur 2

Hieruit valt af te lezen dat de docenten (over het algemeen) het examen moeilijker inschatten dan een examen met N = 1,0. Het frequentiediagram van figuur 2 is omgezet naar een soort boxplot. Zie de bovenste balk in figuur 3. Het 33/67-percentiel-interval is donkerblauw weergegeven en het 10/90-percentiel-interval lichtblauw. De voorlopige technische N-term na stap 1 ligt niet in het 33/67-percentiel-interval dat loopt van 1,1 tot en met 1,3. We kiezen nu voor een N-term die wel in dit interval ligt

en wel zo dicht mogelijk bij de N-term na stap 1:
De N-term na stap 2 werd daarmee een 1,1.

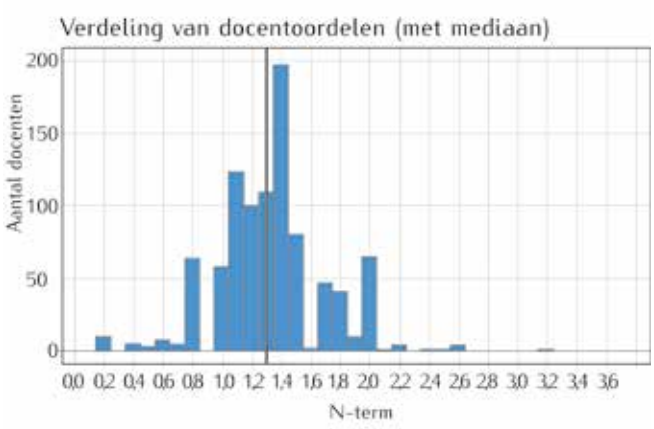
Vervolgens is in stap 3 gekeken naar de historische N-termen. Bij havo wiskunde A betrof dit dus alleen 2017, 2018 en 2019. De N-termen waren toen 1,5; 1,2 en 1,0. Omdat er slechts drie jaren werden meegenomen, werd deze stap met enige voorzichtigheid uitgevoerd. Bij dit vak viel de technische N-term na stap 2 vrijwel middenin het 90% betrouwbaarheidsinterval en was er dus geen reden tot verdere bijstelling van de technische N-term. De hele procedure die leidt tot de technische N-term van 1,1 is zichtbaar in figuur 3. Ten slotte (stap 4) is op basis van signalen van docenten en op basis van de toets- en itemanalyse besloten om de N-term met 0,1 te verhogen in verband met mogelijke tijdnood. Opgemerkt wordt dat de techniek al voorziet in compensatie voor tijdnood. Doordat groepen leerlingen die de leerstof wel voldoende beheersen meer dan gemiddeld last gehad kunnen hebben van de tijdnood, is hier toch besloten voor een (kleine) extra compensatie. Dit alles leidde ertoe dat voor het mei-examen van havo wiskunde A geldt:
N = 1,2. Het gemiddeld cijfer dat de leerlingen in tijdvak 1 behaalden was daarmee een 6,6 en 20% van de leerlingen haalde een onvoldoende.



figuur 3

Tijdvak 2

Ook in tijdvak 2 keken we naar de 'representatieve scholen'. Representatieve scholen zijn de scholen die in tijdvak 1 hadden aangegeven dat hun groep leerlingen die in mei examen deed, representatief was voor de hele klas. Het ligt dan voor de hand dat de groep leerlingen die in juni voor het eerst examen deed ook representatief was voor de hele klas. Er waren 615 leerlingen die in juni voor het eerst examen deden die aan deze steekproef-criteria voldeden. Op dezelfde manier als in het eerste tijdvak werd de N-term na stap 1 berekend. Omdat de scores op dit examen heel erg laag waren, werd dit een 2,3. De docenten beoordeelden het examen wel als iets moeilijker dan het examen in tijdvak 1, maar niet als heel veel moeilijker. De verdeling van de docentoordelen is te zien in figuur 4. Dit leverde een 33/67-percentiel-interval op van [1,2 ; 1,4]. De N-term na stap 2 werd daarmee 1,4. Stap 3 gaf geen bijstelling en dus werd de technische N-term op basis van de scores van de 'eerstekansers' 1,4.



figuur 4

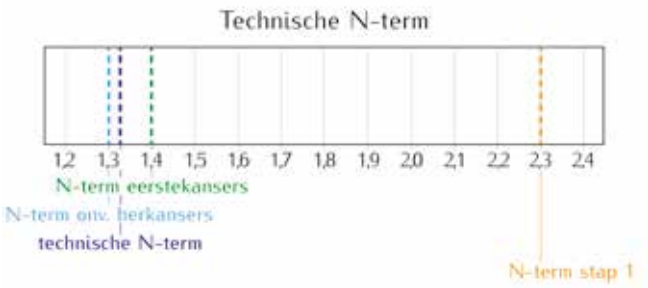
In tijdvak 2 waren er ook herkansers. Ook op basis van hun scores kan een N-term worden geschat. Dit werd gedaan op een vergelijkbare manier die we in het verleden gebruikten voor het tweede tijdvak.

Tweede-tijdvakvergelijking

Bij de tweede-tijdvakvergelijking wordt gekeken naar de procentuele scores van de leerlingen die op het examen in mei een onvoldoende scoorden en in juni herkansten (onvoldoende herkansers). Op basis van regressie naar het gemiddelde en een kleine leerwinst verwacht je bij twee examens die precies even moeilijk zijn, een lichte verhoging van de procentuele score. Elke afwijking van deze lichte verbetering is een indicatie van een verschil in moeilijkheid tussen de twee examens. Bijvoorbeeld, bij een daling van de procentuele score is dit zeer waarschijnlijk het gevolg van het feit dat het tweede tijdvakexamen moeilijker is.

Er waren in het tweede tijdvak 1585 onvoldoende herkansers. De scores van deze leerlingen lieten zien dat het tweede tijdvak 0,2 cijferpunt moeilijker was dan het eerste tijdvak. Deze 0,2 werd bij de technische N-term van

tijdvak 1 opgeteld. Dit leidde tot een N-term van 1,3. We hebben nu op basis van twee verschillende datasets een N-term voor het examen geschat. Deze twee schattingen werden gecombineerd door een gewogen gemiddelde te nemen op basis van het aantal kandidaten. De hele normering van het examen in tijdvak 2 is samengevat in figuur 5.

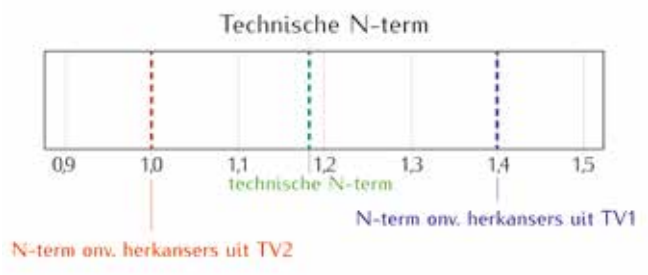


figuur 5

Omdat er meer herkansers waren dan eerstekansers weegt deze uitkomst zwaarder en werd een N-term van 1,3 vastgesteld. Het gemiddelde cijfer van de leerlingen die tijdvak 2 maakten werd hiermee een 5,2 en 56% van de leerlingen haalde een onvoldoende.

Tijdvak 3

In tijdvak 3 waren er 370 onvoldoende herkansers die hun eerste kans in tijdvak 1 hadden gedaan en 445 onvoldoende herkansers die hun eerste kans in tijdvak 2 hadden gedaan. Op basis van de scores van deze twee groepen is de N-term op twee manieren geschat.
In figuur 6 zijn de resultaten in een overzicht te zien.



figuur 6

Het gewogen gemiddelde van de twee groepen was (afgerond) N = 1,2. Deze N-term leidde tot een gemiddeld cijfer 4,7 en 72% onvoldoende.

Nabeschouwing

Goed normeren is niet eenvoudig. Dit heeft vooral te maken met de onzekerheid van de uitkomsten van de schattingen. Door het betrekken van meerdere informatiebronnen kan deze onzekerheid verkleind worden. Het toevoegen van het docentoordeel, zoals we dat dit jaar voor het eerst gedaan hebben, is waardevol gebleken. De normeringen geven nog geen volledig beeld van de vaardigheid van de populatie 2021. Om dit beeld volledig te beschrijven zijn aanvullende analyses nodig. In november zullen CvTE en Cito deze analyses afronden. We verwachten dat deze analyses belangrijke informatie opleveren die meegenomen zullen worden bij de evaluatie van de centrale examens 2021.

Noten

[1] zie: *https://www.examenblad.nl/nieuws/20210219/normering-centrale-examens-2021/2021*

[2] Alleen de jaren 2017, 2018 en 2019 werden meegenomen in de berekening omdat het examen havo wiskunde A vanaf 2017 gebaseerd is op een nieuw examenprogramma.

Over de auteur

Paul van der Molen is oud-docent wiskunde, oud-examenmaker wiskunde en sinds 2014 manager normering (verantwoordelijk voor het technisch normeringsadvies van Cito bij de centrale examens VO).
E-mailadres: *Paul.vandermolen@cito.nl*