Paul van der Molen / Cito, Arnhem

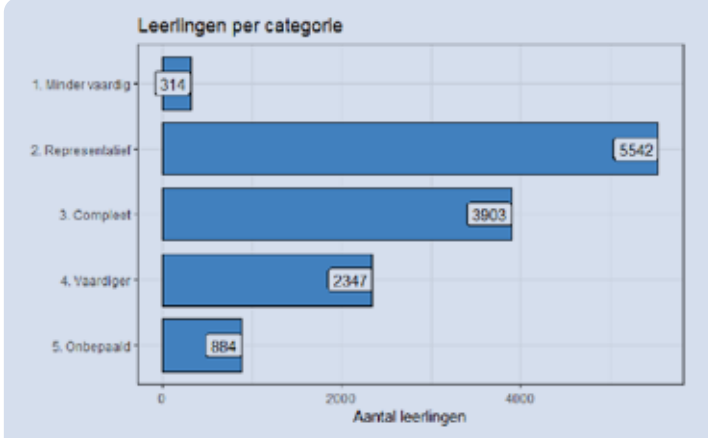
Normering 2021: anders dan anders

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 natuurkunde tot stand zijn gekomen. De vakken biologie en scheikunde hebben hetzelfde proces doorlopen.

De basis van de normering

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 oorzaken 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.
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-informatie-webinar gepresenteerd. Zie ook https://www.examenblad.nl/nieuws/20210219/normering-centrale-examens-2021/2021. Op deze pagina vind je meer gedetailleerde informatie 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 2015-2019. 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-in-



Figuur 1.

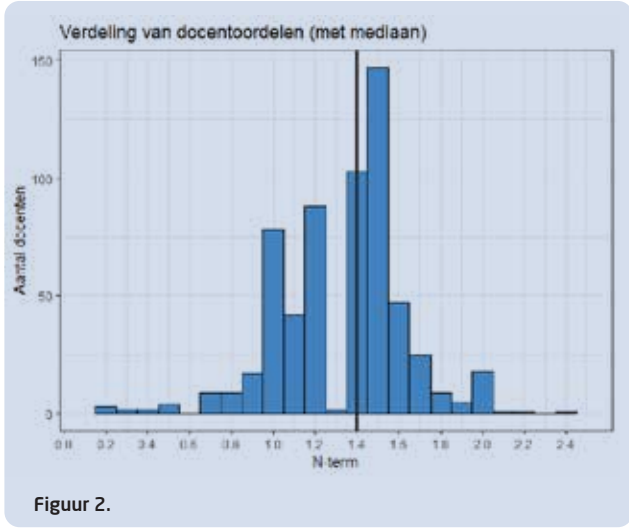
terval). Vervolgens werd in stap 3 vergeleken of de voorlopige N-term wel in lijn was met de gebruikelijke moeilijkheid van het examen van dat vak (historische N-term). De gedachte hierachter is dat het erg onwaarschijnlijk is dat de moeilijkheid van het examen in 2021 opeens erg ver afwijkt van deze waarden. Daartoe is een 90% betrouwbaarheidsinterval gemaakt op basis van het gemiddelde en de standaarddeviatie van de N-termen in de afgelopen 6 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.

Tijdvak 1

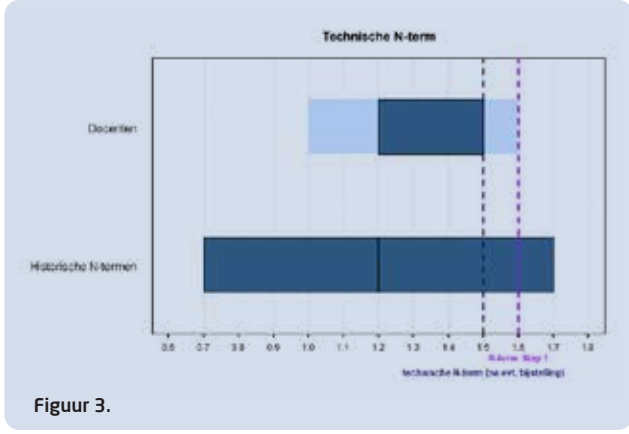
In Wolf hebben 499 scholen de gegevens van 12990 leerlingen doorgegeven die het natuurkunde-examen havo gemaakt hadden. Volgens de vraag over representativiteit zaten 5542 leerlingen in een representatieve groep en 3903 leerlingen zaten in een klas die voor 100% deelnam in mei (zie figuur 1). Er werden dus 9445 leerlingen in de steekproef opgenomen. Het gemiddelde cijfer 2015-2019 van de scholen in de steekproef was een 6,3 volgens een intern afgesproken berekeningswijze. Met behulp van een zogenaamde normeringstabel kan gevonden worden bij welke N-term het gemiddeld cijfer een 6,3 is:

N-term	Gemiddeld cijfer	Percentage onvoldoende
1,4	6,1	31,3
1,5	6,2	28,5
1,6	6,3	25,6
1,7	6,4	22,7
1,8	6,5	20,2

Hieruit volgt dat na stap 1 de voorlopige technische N-term een 1,6 was.
Vervolgens werd gekeken of dit overeenkwam met de docentoordelen. De verdeling van de docentoordelen is weergegeven in figuur 2.
Hieruit valt af te lezen dat de docenten (over het algemeen) het examen iets makkelijker inschatten dan een examen met N = 1,6. 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 1,2 tot en met 1,5. 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 1,5.



Figuur 2.



Figuur 3.

Vervolgens is in stap 3 gekeken naar de historische N-termen. Bij havo natuurkunde betrof dit alleen 2015 tot en met 2019. De N-termen waren toen 1,2 ; 1,2 ; 1,2 ; 0,6 en 1,3. Bij dit vak viel de technische N-term na stap 2 in 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 1,5 is zichtbaar in figuur 3. Ten slotte was in stap 4 geen aanleiding om voor tijdnood of fouten te compenseren via de N-term. Dit alles leidde er dus toe dat voor het mei-examen van havo natuurkunde geldt: N = 1,5. Het gemiddeld cijfer dat de leerlingen in tijdvak 1 behaalden was daarmee een 6,2 en 28,5% van de leerlingen haalde een onvoldoende.

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 494 leerlingen die in juni voor het eerst examen deden die aan deze steekproefcriteria 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 2,1. De docenten beoordeelden het examen als iets makkelijker dan het examen in tijdvak 1. De verdeling van de docentoordelen is te zien in figuur 4. Dit leverde een 33/67-percentiel-interval op van [1,1 ; 1,3]. De N-term na stap 2 werd daarmee een 1,3. Stap 3 gaf geen bijstelling en dus werd de technische N-term op basis van de scores van de 'eerstekansers' een 1,3.

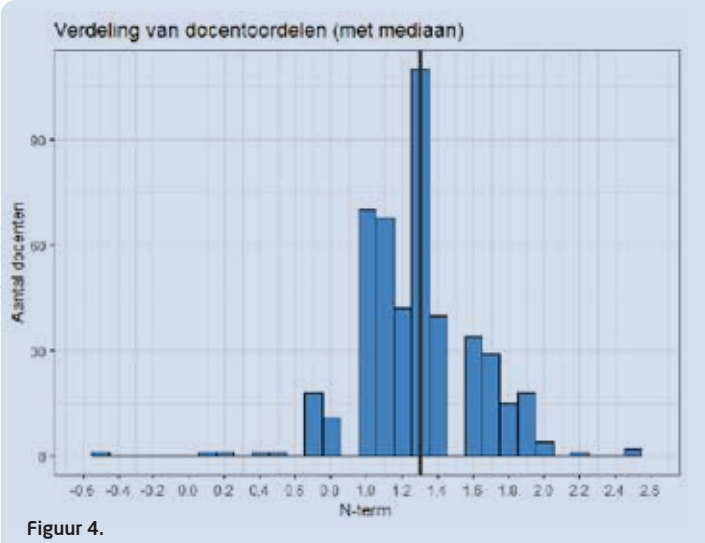
In tijdvak 2 waren er ook herkansers. Ook op basis van hun scores kan een N-term worden geschat. Dit werd gedaan op een vergelijkbare manier die we in het verleden gebruikten voor het tweede tijdvak.

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.

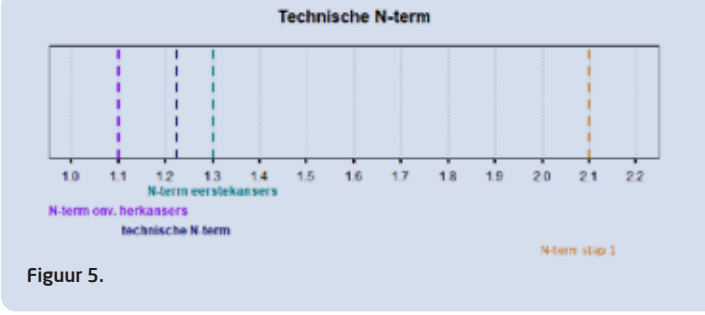
Er waren in het tweede tijdvak 305 onvoldoende herkansers. De scores van deze leerlingen lieten zien dat het tweede tijdvak 0,4 cijferpunt makkelijker was dan het eerste tijdvak. Deze 0,4 werd van de technische N-term van tijdvak 1 afgetrokken. Dit leidde tot een N-term van 1,1. We hebben nu op basis 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. Omdat er meer eerstekansers waren dan onvoldoende herkansers weegt deze uitkomst zwaarder. Het afgeronde gewogen gemiddelde leverde een N-term van 1,2. De hele normering van het examen in tijdvak 2 is samengevat in figuur 5.
Het gemiddeld cijfer van de leerlingen die tijdvak 2 maakten werd hiermee een 5,3 en 53% van de leerlingen haalde een onvoldoende.

Tijdvak 3

In tijdvak 3 waren er 65 onvoldoende herkansers die hun eerste kans in tijdvak 1 hadden gedaan en 42 onvoldoende herkansers die hun eerste kans in tijdvak 2 hadden gedaan. Deze aantallen zijn laag. De schattingen op basis van deze gegevens zijn wel indicatief maar niet betrouwbaar genoeg om direct te volgen. Van tevoren was afgesproken dat in dat geval de N-term van tijdvak 1 of tijdvak 2 wordt overgenomen. Omdat uit de analyse naar voren kwam dat het derde tijdvak makkelijker was dan het examen van zowel tijdvak 1 als van tijdvak 2, werd de laagste van die twee N-termen genomen. Dat was een 1,2. Echter, vanwege een onvolkomenheid in vraag 17 is besloten hiervoor te compenseren. De compensatie werd berekend en hieruit volgde dat de definitieve N-term 1,3 werd. Dit leidde tot een gemiddeld cijfer 5,6 en 42% onvoldoende.



Figuur 4.



Figuur 5.

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 gedaan en in een rapport beschreven. Het rapport 'vaardigheid examenkandidaten 2021' is te vinden op www.cito.nl