Anders dan anders

In juni ontvingen wij via de Examenlijn veel vragen van docenten over de vastgestelde N-term voor het centraal examen economie vwo. Dat was voor ons aanleiding om dit artikel te schrijven, over de totstandkoming van de N-termen bij het examen economie vwo. Bij het vak economie in de andere schoolsoorten is hetzelfde proces doorlopen.

Jacqueline Wooning
Paul van der Molen

De normering van de centrale examens in 2021 is anders verlopen dan in het verleden. 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 houden met de door de coronapandemie 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.

Minder representatief

De toevoeging 'gegeven de omstandigheden' laat al doorschemeren dat we in 2021 te maken hadden met een uitzonderlijke situatie. Er zijn 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 konden we in 2021 niet voor alle vakken onze normhandhavingsinstrumenten zoals pre- of posttesten[1] 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 niet in lijn zijn met eerdere jaren, dan geeft dat problemen bij de normhandhaving.

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 bestond uit vier stappen:

- In stap 1 werd de voorlopige technische N-term zodanig bepaald dat een 'representatieve steekproef van kandidaten' 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 overeenkwam met de norm uit het verleden. Docenten gaven 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-percentielinterval).
- 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 procent betrouwbaarheidsinterval gemaakt op basis van het gemiddelde en de standaarddeviatie van de N-termen in de afgelopen 6 jaar.
- Dat betekent dat de kans 5 procent is dat de N-term die bij dit examen past hoger uitvalt dan de N-termen in dit betrouwbaarheidsinterval en eveneens 5 procent dat de N-term lager zou moeten zijn. 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

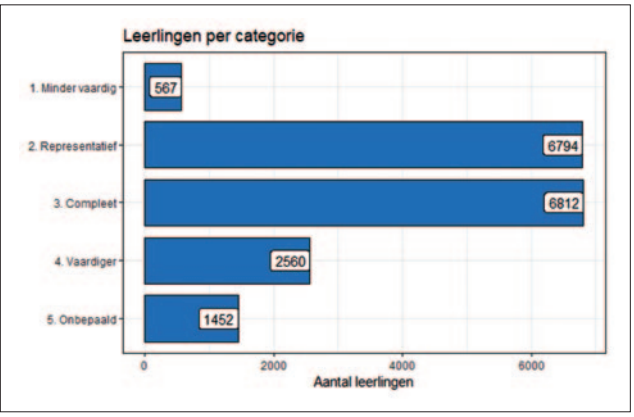
van het 10/90-percentiel-interval van de docentoordelen.
- Tot slot werd in stap 4 nog gekeken of er sprake was van fouten of andere onvolkomenheden waarvoor compensatie via de N-term nodig was.

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 dat Cito en CvTE graag willen weten waar de grens tussen voldoende en onvoldoende prestatie ligt op dit examen vergeleken 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. Dit levert dan een beeld van de verdeling van de oordelen van docenten over de moeilijkheidsgraad van het examen.

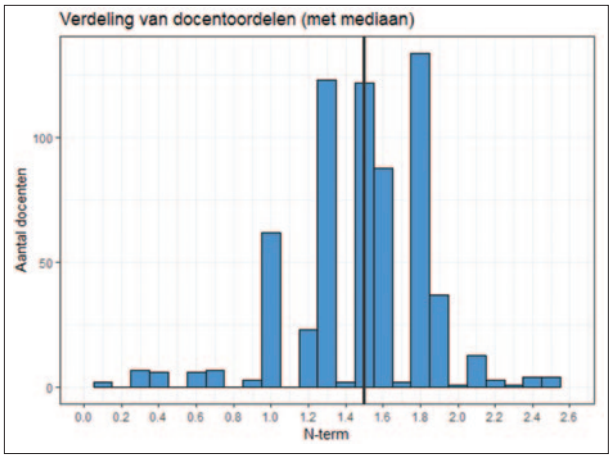
Tijdvak 1

In Wolf hebben 502 scholen de gegevens van 18.185 leerlingen doorgegeven die het examen vwo economie hebben gemaakt. Volgens de vraag over representativiteit zaten 6794 leerlingen in een representatieve groep en 6812 leerlingen in een klas die voor 100 procent deelnam in mei (zie figuur 1). Er werden dus 13.606 leerlingen in de steekproef opgenomen.
De 13.606 leerlingen zaten op scholen die in het verleden gemiddeld even vaardig waren als het landelijk gemiddelde. Het gemiddelde cijfer over de jaren 2014-2019 was voor economie vwo 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:



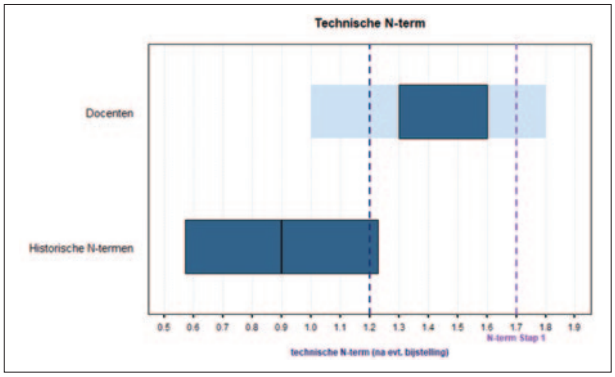
N-term	Gemiddeld cijfer	Percentage onvoldoende
1,5	6,2	27,7
1,6	6,3	23,5
1,7	6,4	20,0
1,8	6,5	20,0
1,9	6,6	16,6

Hieruit volgt dat na stap 1 de voorlopige technische N-term een 1,7 was. Vervolgens werd gekeken of dit overeenkwam met de docentoordelen. De verdeling van de docentoordelen is als volgt weer te geven:



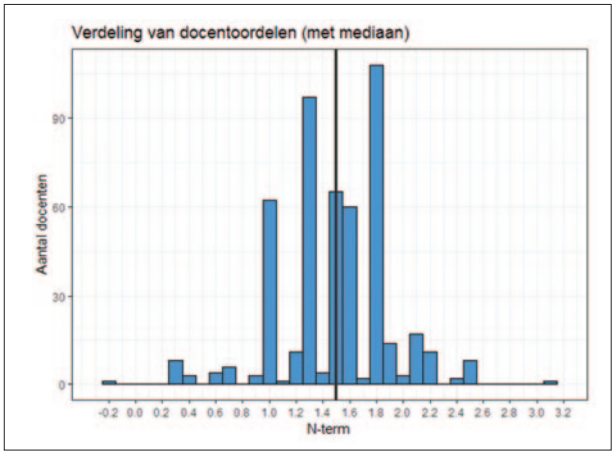
Hieruit valt af te lezen dat de docenten (over het algemeen) het examen makkelijker inschatten dan een examen met N = 1,7. Het frequentiediagram van figuur 2 is omgezet naar een soort boxplot. Zie de bovenste balk in. Het 33/67-percentiel-interval is donkerblauw weergegeven en het 10/90-percentiel-interval lichtblauw.
De voorlopige technische N-term na stap 1 ligt niet in het 33/67-percentiel-interval dat loopt van 1,3 tot en met 1,6. 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 1,6.
Vervolgens is in stap 3 gekeken naar de historische N-termen. Een N-term van 1,6 viel buiten het 90 procent-interval van historische N-termen, waardoor opgeschoven is naar de rand van dit interval. Dat leidde tot een N-term van 1,2. Deze N-term valt ruimschoots binnen het 90 procent-interval van de docentoordelen en dus is de technische N-term 1,2.
Ter controle is ook nog gekeken naar de uitkomsten van een standaardsetting, waarbij een groep docenten voorafgaand aan de afname van het examen volgens een vastgestelde systematiek een inschatting maakt van de moeilijkheidsgraad van het examen. Een N-term van 1,2 zat aan de bovenkant van het interval dat door hen was geschat. Dit gaf dus geen aanleiding om te twijfelen aan de N-term van 1,2.
In stap 4 was geen aanleiding om voor tijdnood of

fouten te compenseren via de N-term. Daarmee werd de N-term vastgesteld op 1,2. Het gemiddeld cijfer dat de leerlingen in tijdvak 1 behaalden was daarmee een 5,9 en 36 procent 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 392 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 2,9. De docenten beoordeelden het examen als ongeveer even moeilijk als in tijdvak 1.
De verdeling van de docentoordelen is te zien in de volgende figuur. Dit leverde een 33/67-percentiel-interval op van [1,3 ; 1,8]. De N-term na stap 2 werd daarmee een 1,8. In stap 3 werd deze door het historisch interval bijgesteld naar 1,2. De technische N-term op basis van de scores van de 'eerstekansers' werd daarmee een 1,2.



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[2].
Er waren in het tweede tijdvak 676 herkansers die voor het eerste tijdvak een onvoldoende hadden gehaald. De scores van deze leerlingen lieten zien dat het tweede tijdvak 0,6 cijferpunt moeilijker was dan het eerste tijdvak. Deze 0,6 werd bij de technische N-term van tijdvak 1 opgeteld.
Dit leidde tot een N-term van 1,8. 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 herkansers waren dan eerstekansers weegt deze uitkomst zwaarder en volgde een technische N-term van 1,6.
Omdat het verschil tussen de twee schattingen groter was dan 0,4 cijferpunt werd kritisch naar deze uitkomst gekeken. Uit de scoreverdeling bleek dat zeer veel leerlingen minder dan 25 procent van de punten hadden gehaald. Dat kan de schatting op basis van de scores van de onvoldoende herkansers onterecht hoog doen uitvallen.
Daarnaast was er een standaardsetting uitgevoerd. Deze gaf aan dat, gezien de moeilijkheidsgraad van het examen, een N-term tussen 1,2 en 1,4 zou passen. Tot slot gaven de docenten aan dat het examen ongeveer even moeilijk was als het examen in tijdvak 1 met een mediaan van 1,5. Alles wegende heeft dat geleid tot een N-term van 1,5.
Het gemiddeld cijfer van de leerlingen die tijdvak 2 maakten werd hiermee een 5,1 en 64 procent van de leerlingen haalde een onvoldoende.

Tijdvak 3

In tijdvak 3 waren er 209 onvoldoende herkansers. Uit de 'tweede-tijdvakvergelijking' bleek dat het examen makkelijker was dan het examen van het eerste tijdvak én het tweede tijdvak. Dat heeft geleid tot een N-term van 1,0. Deze N-term leidde tot een gemiddeld cijfer 5,0 en 58 procent onvoldoende.

Nabeschouwing

De normeringen 2021 geven 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 welke zullen worden meegenomen bij de evaluatie van de normering in 2021.

[1] Zie voor meer informatie over pre- of posttesten de volgende notitie bij www.toetsspecials.nl, bij 'normering': Equivaleringsprocedures.pdf
[2] Voor meer informatie over deze methode, kijk op examenblad.nl, bij 'veelgevraagd': Tweedetijdvakvergelijking - Examenblad

Jacqueline Wooning is programmamanager normering bij het College voor Toetsen en Examens,
Paul van der Molen is manager normering bij het Cito