TOELICHTING OP DE NORMERING

College voor Toetsen en Examens, februari 2024

De normeringssystematiek is altijd in ontwikkeling. De normeringssystematiek die we vanaf 2024 toepassen, is een doorontwikkeling van de systematiek die we voor de coronapandemie hanteerden en de tijdelijke systematiek die we in de coronajaren hanteerden. De normering is er altijd op gericht dat de prestatie-eis op een centraal examen gelijk is, ongeacht het jaar of tijdvak waarin de leerling het examen maakt. De prestatie-eis is de prestatie die een leerling moet leveren op een centraal examen om te laten zien dat die de leerstof beschreven in de syllabus in voldoende mate beheerst.

Als het centraal examen toevallig wat moeilijker is dan zorgt een wat hogere n-term ervoor dat het cijfer dat de leerling haalt hetzelfde is als bij een wat makkelijker examen. Dat betekent dat een prestatie die in het ene jaar een 6 waard is in een ander jaar ook een 6 waard is. Dit heet een absolute normering.

BRONNEN VOOR DE NORMERING

Om te bepalen hoe hoog de n-term voor een examen moet zijn, is het nodig om de moeilijkheidsgraad van het examen te kennen. Dit wordt op verschillende manieren gemeten. Elke manier levert een bron voor de normering op:

Bron E: Equivalering

Met een pretest, een posttest of een anchor-in-package (aip) kan het verschil in moeilijkheid tussen twee examens bepaald worden. Deze drie methoden zijn erop gericht dat een groep leerlingen behalve opgaven uit een nieuw examen ook oude opgaven maakt waarvan de moeilijkheidsgraad al bekend is (een anker). De prestatie op de nieuwe opgaven wordt vergeleken met de prestatie op de oude opgaven. Op deze manier kan bepaald worden of de moeilijkheidsgraad verschilt. Pre- en posttests worden toegepast bij diverse papieren examens voor vmbo gl/tl, havo en vwo en een aip bij vrijwel alle digitale examens voor vmbo bb en kb.

Als het verschil in moeilijkheidsgraad van twee examens bekend is, net alsde afnamegegevens van de examenpopulaties van beide examens, kan ook het verschil in vaardigheid van beide populaties bepaald worden. Bij vakken waarvoor geen pretest, posttest of aip uitgevoerd is, wordt de vaardigheidsverandering geschat op basis van de vaardigheidsverandering van vergelijkbare vakken. Dit kan dan weer gebruikt worden om de moeilijkheidsgraad te bepalen van de examens zonder pretest, posttest of aip.

Bron D: Docentenoordeel

Direct na de correctie van het centraal examen vragen we alle examendocenten om de moeilijkheidsgraad van het gecorrigeerde examen te vergelijken met een examen waarvan de moeilijkheidsgraad bekend is, een zogenoemd referentie-examen. Het gemiddelde van al deze oordelen omtrent het verschil in moeilijkheid wordt gebruikt om de moeilijkheidsgraad van het actuele examen te bepalen. Dit noemen we het docentenoordeel over de moeilijkheid.

Bron H: Historische N-term

Bij ieder vak is er sprake van een zekere mate van continuïteit waar het de examenontwikkeling betreft. Examens zijn ieder jaar nieuw maar moeten ook een zekere mate van voorspelbaarheid hebben zodat een leerling weet wat die kan verwachten. Dit heeft tot gevolg dat de moeilijkheidsgraad van verschillende centrale examens van een vak zich binnen een zekere bandbreedte bevinden. Hoewel n-termen in principe kunnen variëren van 0,0 tot 2,0 – en in de praktijk ook af en toe boven 2,0 – is de variatie binnen een vak veel kleiner. Om te voorspellen wat de n-term van het nieuwe examen zou moeten zijn, wordt daarom ook gekeken naar de n-termen die in het verleden voor een vak zijn vastgesteld, de zogenoemde historische n-termen.

Bron V: Vaststellingscommissie-oordeel
De leden van de vaststellingscommissie maken per examen een inschatting of het examen makkelijker, even moeilijk of moeilijker is dan een referentie-examen[1]. Het gemiddelde oordeel van de vaststellingscommissie over het verschil in moeilijkheid wordt gebruikt om de moeilijkheidsgraad van het actuele examen te bepalen.

Bron C: Cito-standaardbepaling
Bij een standaardbepaling uitgevoerd door Cito schat een groep experts (meestal docenten) de moeilijkheidsgraad van het nieuwe examen. Hiervoor wordt gebruik gemaakt van een vooraf bepaalde methode, zoals de Angoff- of 3DC-methode. Bij een standaardbepaling kan een verschil in moeilijkheid worden geschat tussen het nieuwe examen en het referentie-examen. In geval van een nieuw programma kan op basis van de inhoud van de syllabus voor een nieuw examen bepaald worden waar de grens tussen voldoende en onvoldoende op dat examen zou moeten liggen.

Bron T: Tweede tijdvakvergelijking (alleen beschikbaar in tijdvak 2)
Bij de normering van examens in tijdvak 2 beschikken we over afnamegegevens van leerlingen die het examen van zowel tijdvak 1 als tijdvak 2 hebben gemaakt. Voor de normering maken we gebruik van de scores van leerlingen die in tijdvak 1 een onvoldoende haalden.

Op basis van historische gegevens verwachten we dat de leerlingen zich verbeteren tussen tijdvak 1 en tijdvak 2. Als leerlingen zich minder verbeteren dan we op basis van die historische gegevens verwachten dan is het examen in tijdvak 2 blijkbaar moeilijker dan in tijdvak 1 en vice versa. Op basis van de verbetering die leerlingen laten zien kunnen we de moeilijkheidsgraad van het examen in tijdvak 2 vergelijken met de moeilijkheidsgraad van het examen in tijdvak 1.

Voor een uitgebreidere beschrijving van deze methode verwijzen we naar de bijlage in dit document.

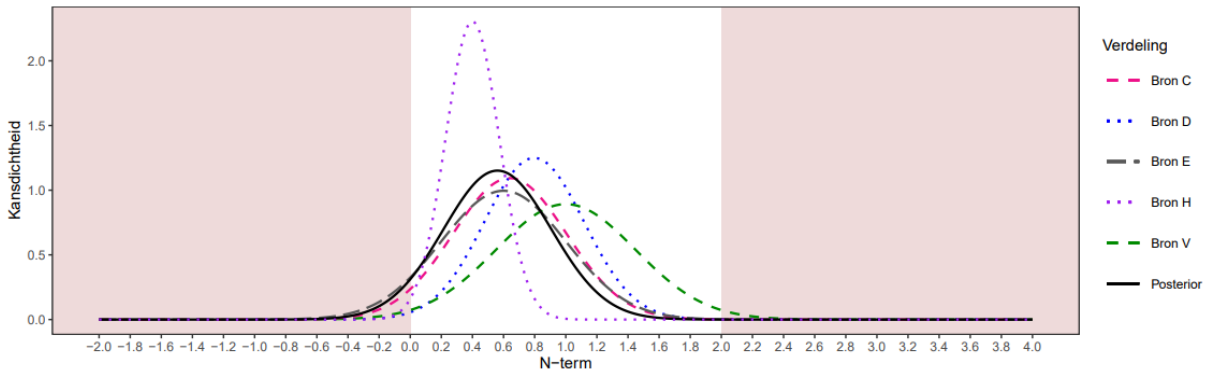
BEPALEN VAN DE TECHNISCHE N-TERM
Elk van bovengenoemde methoden om de moeilijkheidsgraad van het te normeren examen te bepalen leidt tot een schatting van de technische n-term het examen. Het kan per examen verschillen welke bronnen beschikbaar zijn. Zo voert Cito voor een examen met een aip geen standaardbepaling uit en zal bron C ontbreken. En als een vak bijvoorbeeld een nieuw examenprogramma heeft, zijn er geen historische n-termen beschikbaar.

Elk van de bronnen heeft zijn eigen betrouwbaarheid. Zo is bijvoorbeeld een aip in principe betrouwbaarder dan een pretest en is een pretest in principe weer betrouwbaarder dan een standaardsetting. Voor het docentenoordeel geldt dan weer dat het een betrouwbaardere schatting oplevert als meer docenten hun oordeel hebben gegeven. Zo zijn er verschillende factoren die de betrouwbaarheid van een bron kunnen beïnvloeden.

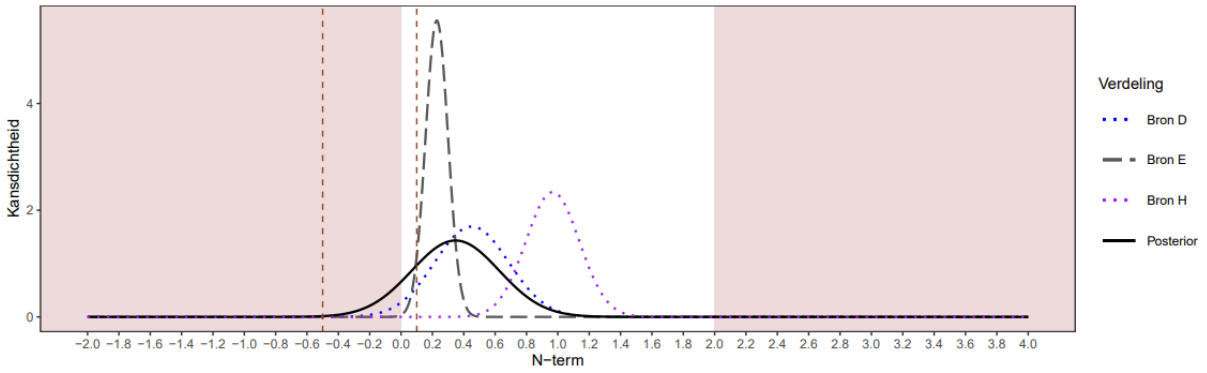
Voor het bepalen van de technische n-term worden alle beschikbare bronnen gewogen waarbij een betrouwbaarder bron een zwaarder gewicht krijgt dan een minder betrouwbare bron.

In de praktijk ziet dat er als volgt uit:

[1] Een eerder afgenomen examen waarvan de moeilijkheidsgraad, en dus de n-term, bekend is.



Elke beschikbare bron wordt weergegeven als een verdeling van n-termen. Een bron met grotere betrouwbaarheid heeft een hogere en smallere piek dan een bron met een lagere betrouwbaarheid. De doorgetrokken zwarte lijn, de posterior genoemd, is het resultaat van de weging van alle bronnen samen, waarbij een betrouwbare bron zwaarder meetelt dan een minder betrouwbare bron. In dit geval liggen de toppen van de verdelingen van de verschillende bronnen tussen 0,4 (bron H) en 1,0 (bron V). De top van de posterior ligt op 0,6. Dit is de meest waarschijnlijke uitkomst en dus is dit de technische n-term. In dit geval heeft bron H een grotere betrouwbaarheid dan de andere bronnen. Hierdoor ligt de technische n-term relatief dicht bij de historische n-term.

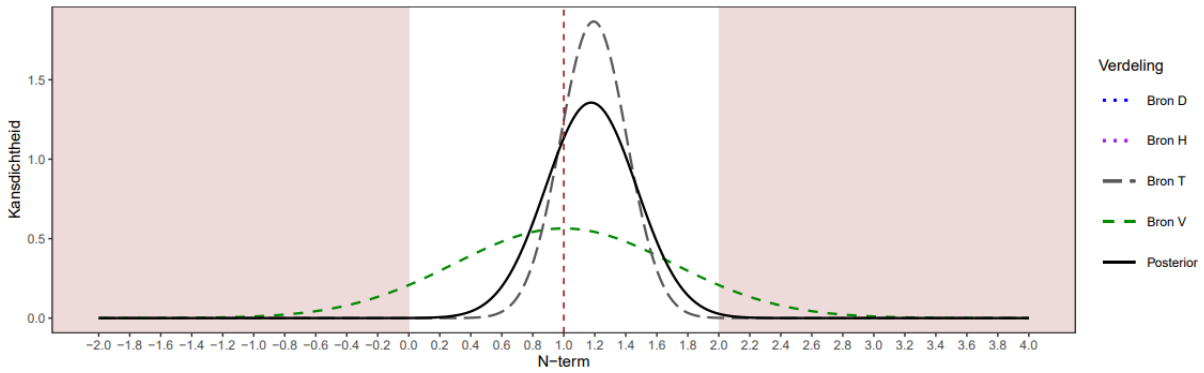


Niet bij ieder examen zijn alle bronnen beschikbaar. Hierboven is een examen te zien waarbij alleen de bronnen D, E en H beschikbaar zijn. De toppen van de verdelingen van de verschillende bronnen variëren van 0,2 (bron E) tot 1,0 (bron H). De top van de posterior ligt dicht bij de top van bron E die in dit geval het meest betrouwbaar is. De technische n-term is nu 0,3.

De verticale bruine stippellijnen geven de n-termen weer waarbij de actuele examenpopulatie een gemiddeld cijfer zouden halen dat 0,3 hoger of 0,3 lager is dan het gemiddeld cijfer dat de examenpopulatie van het jaar ervoor haalde. Normaal gesproken is de vaardigheid van examenpopulaties over de jaren heen redelijk stabiel. Het gemiddeld cijfer is dan ook min of meer constant over de jaren heen. Een afwijking van 0,3 cijferpunt is dan niet waarschijnlijk en kan het gevolg zijn van een statistische toevalligheid. De bruine lijnen dienen in zo'n geval als waarschuwing om nog eens kritisch te kijken naar de beschikbare informatie en kunnen als afkapgrens gehanteerd worden.

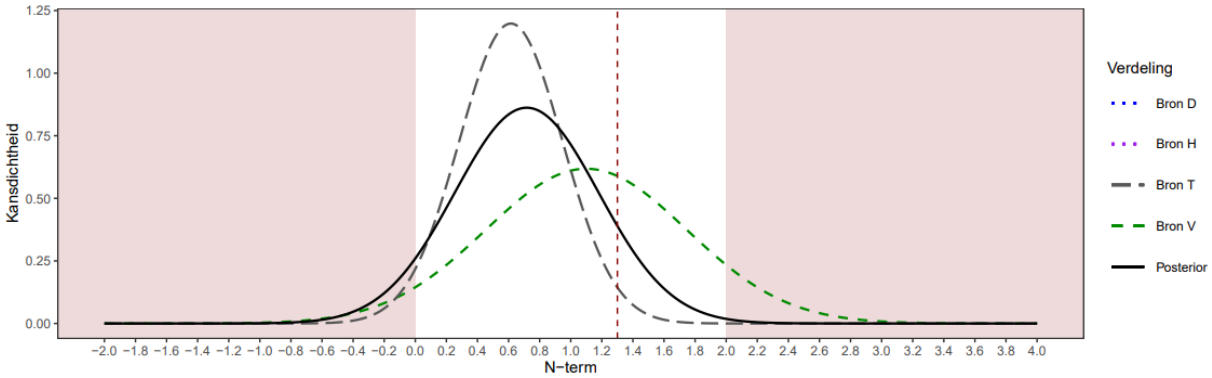
TIJDVAK 2

De normering in tijdvak 2 verloopt op dezelfde wijze als hiervoor beschreven met één verschil. Voor tijdvak 2 wordt na afloop van tijdvak 1 een voorlopige n-term gepubliceerd die een ondergrens voor de vast te stellen n-termen vormt. Hieronder wordt dat toegelicht aan de hand van twee voorbeelden.



We zien hier een examen uit tijdvak 2. De bruine stippellijn geeft hier de voorlopige n-term voor tijdvak 2 weer. Als de top van een verdeling rechts van deze bruine lijn ligt, wijst de bron op een moeilijker examen in tijdvak 2. Ligt de top links van de verdeling dan is het examen in tijdvak 2 waarschijnlijk makkelijker.

In dit voorbeeld is de doorgeschoven n-term 1,0. De twee bronnen V en T wijzen op een n-term van respectievelijk 1,0 en 1,2. De posterior verdeling heeft zijn top net op 1,2 liggen. Het examen in tijdvak 2 blijkt op basis van beide bronnen 0,2 cijferpunt moeilijker en de technische n-term voor het examen in tijdvak 2 is dus 1,2.



In dit voorbeeld is de doorgeschoven n-term gelijk aan 1,3. Beide beschikbare bronnen wijzen op een lagere n-term, namelijk 1,1 (bron V) en 0,6 (bron T). De top van de posterior verdeling ligt op 0,7, wat de best passende n-term voor dit examen is. Desondanks is de technische n-term 1,3 omdat de voorlopige n-term de ondergrens van vast te stellen n-termen is.

Mocht in dit examen een fout geconstateerd zijn waarvoor via de n-term gecompenseerd moet worden dan wordt de hoogte van de compensatie allereerst berekend middels de daarvoor geldende formules. De uitkomst hiervan wordt opgeteld bij de top van de posterior verdeling. Stel dat in dit voorbeeld de hoogte van de compensatie 0,2 cijferpunt zou zijn dan tellen we die op bij 0,7. Dat zou leiden tot een gecorrigeerde n-term van 0,9. Maar aangezien deze nog steeds lager is dan de voorlopige n-term zal de n-term worden vastgesteld op 1,3 na compensatie. Het examen is makkelijker gebleken dan het examen in tijdvak 1. Dit komt echter niet tot uiting in de n-term omdat hiervoor de voorlopige n-term als ondergrens geldt. De n-term voor het examen in tijdvak 2 is al hoger dan op grond van de moeilijkheidsgraad passend is.

VAN TECHNISCHE N-TERM NAAR DEFINITIEVE N-TERM

De technische n-term is niet altijd de definitieve n-term. In eerste instantie kijken we kritisch naar de totstandkoming van de technische n-term. We gaan na of er redenen zijn om aan de uitkomst te twijfelen omdat er bijvoorbeeld bijzondere omstandigheden waren waardoor een specifieke bron minder bruikbaar is. Zo is het denkbaar dat als een vaststellingscommissie vooraf de moeilijkheid van een examen schat maar in de campagne een vraag met veel punten wordt geneutraliseerd vanwege een fout in de opgave de

moeilijkheidsschatting van minder waarde is omdat hierbij geen rekening is gehouden met die neutralisatie.

Daarnaast kijken we samen met de vaststellingscommissie of de n-term nog moet worden aangepast vanwege fouten of tijdnood. Hierbij maken we gebruiken van de toets- en itemanalyse en de quickscan, waarin docenten direct na het afronden van de correctie middels korte vragen feedback geven op het examen. Ook wegen we de reacties die we over een centraal examen hebben ontvangen van leerlingen via het LAKS en van docenten via onze eigen examenlijn. Als blijkt dat er een fout in het examen zat of dat het examen te lang was, passen we een compensatie toe door de n-term op te hogen met een waarde die met vooraf vastgestelde formules wordt bepaald.

BIJLAGE: TWEEDE TIJDVAKVERGELIJKING IN MEER DETAIL (BRON T)

De tweede tijdvakvergelijking begint met het berekenen van het verschil in procentuele score tussen tijdvak 2 (tv2) en tijdvak 1 (tv1). We kijken daarvoor naar de groep herkansers die in tv1 een onvoldoende haalde. We vergelijken dus de procentuele score van deze groep in tv2 met de procentuele score die deze groep haalde in tv1. Nu zijn er drie factoren die van invloed zijn op dit verschil in procentuele score:

1) Verschil in moeilijkheid tussen tv2 en tv1
   Dit is wat we bij het normeren willen weten. De n-term compenseert tenslotte voor een verschil in moeilijkheid tussen verschillende examens.
2) Het regressie-effect
   Dit is een empirisch, statistisch bekend verschijnsel; leerlingen met lage cijfers zullen zich gemiddeld meer verbeteren dan leerlingen met hoge cijfers.
3) De leerwinst
   De leerwinst draagt bij aan een verschil in de procentuele score omdat de leerling in de voorbereiding op de herkansing nog even alles op alles zet. Dit zorgt ervoor dat dat deze groep in tv2 iets vaardiger is dan in tv1.

Om het verschil in moeilijkheid tussen tv2 en tv1 te kunnen bepalen moeten ook de andere twee factoren bekend zijn. Het verschil in procentuele score kan berekend worden aan de hand van de ingestuurde gegevens in Wolf. Het regressie-effect wordt berekend met behulp van een statistische methode. De leerwinst kan geschat worden door uitkomsten over heel veel jaren te middelen. Als we het verschil in procentuele score tussen tv2 en tv1 voor de onvoldoende herkansers compenseren voor het regressie-effect en de leerwinst, kennen we het verschil in moeilijkheid tussen tv2 en tv1. Als tv2 moeilijker blijkt wordt de voorlopige nterm tv2 opgehoogd met het zojuist berekende verschil in moeilijkheid. Dit levert de top van bron T.