Technische toelichting bij groepsrapportage digitale examens

Deze technische toelichting wil stap voor stap laten zien hoe de berekeningen zijn uitgevoerd die in de groepsrapportage zijn weergegeven. Zoals in de leeswijzer al is aangegeven, bestaat de groepsrapportage uit drie delen die eigenlijk alle drie dezelfde informatie geven. Dit zijn van links naar rechts:

a) tabel met geobserveerde groeps- en landelijke gemiddelden
b) grafische weergave van de geobserveerde gemiddelden
c) grafische weergave van de betrouwbaarheidsintervallen rondom de groepsgemiddelden

Het eerste deel van deze toelichting gaat in op de wijze waarop de berekeningen zijn uitgevoerd. In het tweede deel zullen we iets uitgebreider stil staan bij de lijntjes onder 'prestatie t.o.v. landelijk' (onderdeel c).

Bij de digitale examens is er sprake van meerdere varianten met een zekere overlap. Bij de constructie en vaststelling van de examens zijn alle vragen uit alle varianten toegewezen aan een van de drie domeinen die binnen wiskunde onderscheiden worden. Elke leerling heeft één variant gemaakt. Daarmee heeft hij een totaalscore bereikt die met behulp van een N-term voor die variant kan worden omgezet in een cijfer.
Op deze manier is het verschil in moeilijkheid tussen de varianten verdisconteerd in het examencijfer.
Maar zoiets bestaat nog niet op domeinniveau.
Om dit probleem op te lossen is bedacht om alle meetkunde-opgaven, alle rekenopgaven en alle algebra-opgaven bij elkaar te zetten, zodat drie grote sets van opgaven ontstaan. Aan de hand van de score van een leerling op de opgaven binnen een bepaald domein en met behulp van de kennis hoe moeilijk de opgaven van dat domein in zijn variant waren, kan berekend worden wat zijn score zou zijn geweest als hij alle opgaven van dat domein had gemaakt. Je kunt dit beschouwen als extrapolatie van zijn vaardigheid naar de andere vragen in het domein[1]. Het resultaat is dat leerlingen die een variant hebben gemaakt waarin het domein relatief moeilijk was, relatief gezien op alle vragen van dat domein beter zullen scoren.

Een voorbeeld: leerling A maakt variant x. Daarin zitten 20 scorepunten meetkunde. Hij scoort 10 punten.
Leerling B maakt variant y. Daarin zitten 22 scorepunten meetkunde. Hij scoort 11 punten.
Als leerling A alle meetkunde-opgaven uit alle varianten zou maken dan zou hij 55% goed maken.
Als leerling B alle meetkunde-opgaven uit alle varianten zou maken dan zou hij 45% goed maken.
Dat ze een verschillende 'berekende' totaalscore krijgen, terwijl ze binnen hun variant allebei 50% scoorden, komt dus omdat de meetkunde-opgaven in variant x iets moeilijker waren dan in variant y.
Leerling A scoorde 50% op een moeilijke set opgaven en zal dus meer dan 50% halen als hij ook de wat makkelijkere opgaven uit de andere varianten voorgelegd zou krijgen.

Vervolgens worden de berekende scores van alle leerlingen van Nederland samengenomen.
Het gemiddelde van deze score is het landelijk gemiddelde. Op deze manier is te zien dat heel Nederland voor rekenen beter scoorde dan voor meetkunde. Kennelijk zijn de meetkunde-opgaven moeilijker voor de leerlingen dan de rekenopgaven.
Voor de groepsrapportage zijn de gemiddelden per brin-dependancecode berekend en daarbinnen ook nog per corrector. Op deze manier kunnen de resultaten van alle leerlingen op een locatie, of van alle leerlingen van een corrector, vergeleken worden met de prestatie van alle leerlingen in Nederland.

Hiermee zijn de resultaten in het linker deel en het middendeel van de groepsrapportage besproken.

Nu willen we de lijnen in het rechterdeel van de rapportage nader toelichten. Daarvoor hebben we twee korte intermezzo's nodig.

Kort intermezzo 1: Wat is een percentielscore?
Als iemand een bepaalde percentielscore heeft behaald dan is dat het percentage leerlingen dat minder goed dan hem presteerde. Een percentielscore van 65 betekent dus dat 65% van de overige leerlingen in Nederland minder goed presteerde. Als alle leerlingen van Nederland op volgorde van score gaan staan, dan staat 65% aan de ene kant en 35% aan de andere kant. Dit getal zegt dus iets over de plaats waar je in de populatie staat.

[1] Hiervoor zijn psychometrische modellen nodig. Het gaat voor deze technische toelichting te ver om dit volledig uit de doeken te doen.

Bij de groepsrapportage is gekeken welk percentiel iedere school zou krijgen op basis van zijn gemiddelde prestatie, ten opzichte van alle andere scholen. Ook is van iedere corrector de relatieve positie ten opzichte van alle correctoren bekeken, gebaseerd op de gemiddelde prestatie van de kandidaten die onder de corrector vallen.

Kort intermezzo 2: wat is een betrouwbaarheidsinterval?
Ieder berekend gemiddelde is eigenlijk een schatting van het 'ware' gemiddelde. In de groepsrapportage gaat het om de 'ware' gemiddelde prestaties van kandidaten per school of per corrector. Ieder gemiddelde dat we schatten, kent een schattingsfout. Zo ook bij de schatting van groepsgemiddelden binnen de groepsrapportage[2]. Deze schattingsfout hangt af van de spreiding van de prestaties van kandidaten binnen scholen, van de spreiding van de gemiddelde prestaties tussen scholen en van het aantal kandidaten van een school of corrector. Dit aantal kandidaten verschilt tussen scholen of correctoren. De schattingsfout wordt kleiner naarmate het gemiddelde op basis van meer kandidaten wordt geschat: er is immers meer informatie beschikbaar.

Een 90%-betrouwbaarheidsinterval is een interval rondom het geschatte gemiddelde. De breedte ervan hangt af van de schattingsfout. De verschillen tussen scholen of correctoren ontstaan door verschillen in aantal kandidaten. Bij (theoretisch) oneindig vaak herhalen van de hele examenafname, met dezelfde kandidaten, correctoren en scholen (zonder leereffecten tussendoor), en steeds opnieuw berekenen van de betrouwbaarheidsintervallen, dan bevindt het 'ware' gemiddelde zich in 90% van de gevallen in dat interval. Vaak wordt daarom gezegd dat met 90% zekerheid het 'ware' gemiddelde van een school of corrector in dat betrouwbaarheidsinterval ligt.

Wat stellen de lijntjes nu precies voor?
In de groepsrapportage zijn de boven- en ondergrens van het 90%-betrouwbaarheidsinterval rondom het school- of corrector-gemiddelde omgezet in percentielscores. Deze boven- en ondergrens zijn de bolletjes aan de uiteinden van de lijntjes. Je mag deze lijntjes interpreteren als een indicatie van de kwaliteit van het onderwijs. Dat behoeft een nadere uitleg.

De lijn 'min' correspondeert met een percentielscore 0. De stippellijn in het midden correspondeert met een percentielscore van 50 en de lijn 'max' correspondeert met een percentielscore van 100.
We willen de interpretatie van een lijntje nader uitleggen aan de hand van een voorbeeld. In de tweede helft van pagina 2 van de leeswijzer staat een tabel met interpretaties. We nemen nu voor onze bespreking als voorbeeld het tweede lijntje van boven. Dit zou kunnen passen bij een klas met 25 leerlingen, die op het gehele examen een cijfer 6,6 scoorde terwijl het landelijk gemiddelde een 6,4 was. Het betrouwbaarheids-interval-lijntje loopt (ongeveer) van een percentielscore 40 tot 77. De gemiddelde percentielscore van deze klas zal ergens rond de 60 liggen. Het gemiddelde ligt wel boven de stippellijn (de klas is gemiddeld iets beter dan landelijk) maar je kunt niet zeggen dat deze klas beter presteert dan landelijk. Het 'ware' gemiddelde van de klas zou namelijk gelijk kunnen zijn (90% zeker) aan het landelijk gemiddelde omdat de stippellijn (percentiel 50 en dus gelijk aan het landelijk gemiddelde) binnen het betrouwbaarheidsinterval valt. Het verschil tussen het gemiddelde van deze klas en het landelijk gemiddelde is daarom niet significant.

Als het hele betrouwbaarheidsinterval boven percentiel 50 ligt, dan weten we met redelijke zekerheid (90%) dat de school of corrector bovengemiddeld heeft gepresteerd. Als het hele betrouwbaarheidsinterval onder percentiel 50 ligt, dan weten we met redelijke zekerheid (90%) dat de school of corrector ondergemiddeld heeft gepresteerd.

Het kan voorkomen dat een corrector maar 2 leerlingen had die gemiddeld een cijfer 8,2 scoorden.
De percentielscore van de bovengrens is dan wel heel hoog maar van de ondergrens niet. De gebruikte methode zorgt er namelijk voor dat het betrouwbaarheidsinterval breed wordt als de groep klein is of de spreiding in de groep groot is. Je mag het ook zo lezen: het betrouwbaarheidsinterval zegt iets over de kwaliteit van het onderwijs. Als je 2 heel erg goede leerlingen hebt dan wil dat niet 100% zeker zeggen dat het onderwijs heel goed was. Je had mogelijk toevallig twee erg goede leerlingen. Als je met 50 leerlingen een gemiddelde van 8,2 haalt dan is de kans veel groter dat dit komt door goed onderwijs.
Het betrouwbaarheidsinterval zal in dat geval smaller zijn en ook een stuk hoger liggen.

[2] Om precies te zijn wordt voor het schatten van het gemiddelde van een school of corrector de Kelly- of Empirical Bayes-schatter gebruikt. Dit is een algemeen geschikt bevonden schatter binnen een analyse met meer niveaus (leerlingen binnen groepen). Hierbij vindt een weging plaats tussen het landelijk gemiddelde en het geobserveerde groepsgemiddelde. Naar mate er meer leerlingen in de groep zitten, weegt het geobserveerde gemiddelde zwaarder.

Nog even de kenmerken op een rijtje:
Hoe groter het verschil tussen de groep en landelijk, hoe verder de lijn weg uit het midden schuift.
Hoe kleiner de groep, hoe langer de lijn.

Om een idee te krijgen hoe een rapportage gelezen zou kunnen worden, wordt verwezen naar een voorbeeldrapportage met daarbij een beschrijving hoe de resultaten in deze rapportage geduid zouden kunnen worden. Deze zijn te vinden op de Cito website.

Ten slotte nog een relativering. Bepaalde verschillen tussen jouw groep en landelijk lijken erg duidelijk. Toch kan het voorkomen dat het gegeven onderwijs niet de enige reden is voor het optreden van verschillen. Het lijkt daarom verstandig om de aanpassingen aan het onderwijs met beleid te doen. Als er meerdere jaren achtereen hetzelfde beeld uit de rapportage naar voren komt, is de noodzaak om in te grijpen evident. Zo kan een achterblijvend domein rekenen in het ene jaar optreden en het volgende jaar niet. Pas als meerdere jaren achtereen rekenen achterblijft bij de landelijke prestaties en/of bij de andere domeinen dan lijkt het verstandig om daar in het onderwijs echt actief op in te spelen.