Equivaleren en equivaleringsprocedures

Wat is equivalering?
Equivalering is een procedure waarbij op basis van statistische gegevens de norm (=standaard) die op een referentie-examen is vastgesteld, wordt overgebracht op een nieuw examen. Het equivaleren van twee examens is het vergelijken van de moeilijkheid van die examens om te kunnen bepalen welke score op de ene toets overeenkomt met een bepaalde score op de andere toets. Bij het equivaleren van examens gaat het erom op het nieuwe examen de score aan te wijzen die vergelijkbaar is met de score die op het referentie-examen de laagste voldoende opleverde (=5,5). Als dat bekend is, kan voor het nieuwe examen een equivalente N-term berekend worden.

Equivaleringsprocedures
Er zijn verschillende equivaleringsprocedures. De verschillen tussen de procedures hebben betrekking op de aannames die worden gedaan en zitten voor een deel ook in de logistieke uitwerking van de gegevensverzameling. Het equivaleren van examens is alleen zinvol als de beide te equivaleren examens dezelfde eigenschap meten. Een voorwaarde voor equivalering is dat de examens inhoudelijk gelijkwaardig zijn. Daarmee bedoelen we dat de beide examens betrekking hebben op dezelfde leerstof (examenprogramma) en inhoudelijk uitwisselbaar zijn. Dat bereiken we door de beide examens te ontwikkelen volgens eenzelfde toetsmatrijs. De examens hebben dan een vergelijkbare verdeling van vragen (en scorepunten) over de onderscheiden onderdelen uit de leerstof.

De volgende methoden worden onderscheiden:
1 Equivalering op basis van aanvullende gegevens
- Pretest
- Posttest

2 Equivalering bij gelijkwaardige populaties (of: equivalering op resultaat)
3 Equivalering op basis van overlap tussen examens
- Equivalering van experimentele examens
- Equivalering van toetsvarianten

4 Tweede tijdvakequivalering
5 Equivalering op basis van kwalitatieve expertoordelen (of: deskundigenoordeel)

Equivalering op basis van aanvullende afnamegegevens
Kenmerkend voor equivalering op basis van aanvullende afnamegegevens is dat de opgaven uit het referentie-examen en het nieuwe examen gecombineerd worden afgenomen. Zo kunnen de opgaven van beide examens ook daadwerkelijk op moeilijkheidsgraad met elkaar vergeleken worden. We onderscheiden bij deze vorm van equivaleren twee modaliteiten namelijk de pretest en de posttest. Het onderscheid heeft betrekking op het tijdstip waarop de 'aanvullende afnamegegevens' worden verzameld, namelijk voor of na de afname van het eigenlijke examen.

Pretest[1]
Bij de pretest worden de opgaven voor een nieuw examen voorgelegd aan een geschikte populatie, tezamen met opgaven van het referentie-examen. Zo kan de moeilijkheid van de vragen van het nieuwe examen ten opzichte van het referentie-examen worden bepaald. In de genoemde vakken bestaan de examens

[1] De pretest komt vooral voor bij vwo en havo. De pretest wordt toegepast bij wiskunde A en B, natuurkunde, scheikunde, biologie, economie en voor havo ook bij aardrijkskunde. In het vmbo wordt het toegepast bij wiskunde en economie.

voor vijftig tot tachtig procent uit gepreteste opgaven. Aan de hand van de gepreteste vragen uit het examen wordt na afname van het examen de equivalente N-term berekend. De uitkomsten van deze bepaling worden vervolgens getoetst aan algemene criteria voor de normering, zoals toegestane percentages onvoldoendes en niet te grote variatie van jaar op jaar.

Posttest
Deze methode komt alleen voor bij vakken met heel veel machinaal scoorbare antwoorden: de moderne vreemde talen. Meteen na het examen wordt het zojuist afgenomen examen tezamen met onderdelen uit het referentie-examen afgenomen bij een geschikte populatie. Voor vmbo-tl-leerlingen zijn bijvoorbeeld vwo/havo-leerlingen van leerjaar 3 een geschikte populatie.

De resultaten die deze leerlingen op de voorgelegde toetsen behalen, vormen de aanvullende data. Deze toetsen bevatten voor een deel oude examenopgaven of zogeheten ankeropgaven, waarvan de statistische eigenschappen bekend zijn. De rest van de toets die ze maken bestaat uit opgaven uit het nieuwe (te equivaleren) examen. Deze toetsen worden gemaakt op het moment dat het nieuwe examen net afgelopen is. De leerlingen bij wie de aanvullende data worden verzameld, kunnen dan nog niet bekend zijn met de opgaven uit het nieuwe examen. De techniek van de bepaling van de N-term verloopt verder net zo als bij vakken met een pretest.

Aan de equivalering volgens pre- of posttest ligt een modelmatige benadering ten grondslag met een eigen begrippen kader. Vertrekpunt voor de equivalering via een pretest of posttest vormt een statistisch model (zie hieronder) dat geschikt is om de examendata en de aanvullende data adequaat te beschrijven. In zo'n model wordt ervan uitgegaan dat de scores van leerlingen op de vragen verklaard kunnen worden door twee verschillende factoren, namelijk de vaardigheid van de leerling (de leerlingparameter) en de moeilijkheid van de vraag (de itemparameter).

Als we beschikken over beide parameters, komt het model tot voorspellingen over de resultaten die leerlingen behalen op de verschillende vragen. Het variëren van de parameters leidt tot andere voorspellingen. In een schattingsprocedure wordt bepaald welke parameterwaarden het best passen bij de geobserveerde data; deze parameterwaarden beschrijven zo goed mogelijk de beschikbare data. Als we de parameters van alle leerlingen en van alle vragen hebben bepaald, kunnen we het model gebruiken om van leerlingen resultaten te voorspellen op vragen die zij niet hebben gemaakt. Toepassing van deze techniek maakt het mogelijk om op basis van het model te voorspellen wat de resultaten zouden zijn van eenzelfde groep leerlingen op elk van beide examens. Met deze voorspelde resultaten kunnen we beoordelen of het ene examen moeilijker of makkelijker is dan het andere.

Samenvattend bestaat de procedure uit drie stappen:
1. schatten van het model
2. voorspellen van de resultaten op vragen die leerlingen niet gemaakt hebben
3. vergelijken van de examens

Statistisch model in de praktijk
Het statistisch model dat we gebruiken bij meerkeuzevragen en open vragen is een vrij eenvoudig kansmodel. Het gaat ervan uit dat de kans dat een bepaalde leerling een bepaalde vraag goed beantwoordt, afhankelijk is van de vaardigheid van die persoon en van de moeilijkheid van de opgave. Bij meerkeuzevragen kunnen de scores van de leerlingen in twee categorieën verdeeld worden, bijvoorbeeld 'goed' en 'fout', of 1 en 0. Bij het gehanteerde statistische model is

de kans op een goed antwoord 50 procent, indien de vaardigheid van een leerling precies even groot is als de moeilijkheid van de vraag. Is de vaardigheid kleiner dan de moeilijkheid, dan neemt de kans op een goed antwoord af. Is de vaardigheid groter dan de moeilijkheid, dan neemt de kans toe.

Voor een uitgebreidere toelichting op deze manier van equivaleren verwijzen we naar de publicatie 'Equivalering op basis van aanvullende data' die gedownload kan worden in de toetsspecial 'Normering' in de rubriek over equivalering.

Equivalering op resultaat (equivaleren bij even vaardige populaties)
Bij de vakken met een pretest en een posttest blijkt dat de vaardigheid van kandidaten in opeenvolgende jaren zelden grote sprongen maakt. De vaardigheid in opeenvolgende jaren is min of meer constant wanneer er geen ingrijpende wijzigingen in studielast, examenprogramma of bijvoorbeeld toegestane hulpmiddelen aan de orde zijn. Dat maakt het mogelijk om bij vakken zonder pretest, posttest of mogelijke koppeling aan een ander vak toch een equivalente norm vast te stellen. Dit kan door N zo te kiezen dat het resultaat bij het te normeren examen overeenkomt met dat van het referentie-examen (percentage onvoldoende en in bepaalde omstandigheden speelt ook het gemiddeld cijfer een rol). Dat dit kan heeft te maken met een belangrijk verschil tussen een individuele school en het hele stelsel. Op een individuele school kan de toevallige vaardigheid van een leerlingengroep van jaar tot jaar verschillen. Het zou dan raar zijn om op resultaat te equivaleren. Goede leerlingen zouden dan te lage en matige leerlingen te hoge cijfers krijgen. In het hele stelsel zijn er echter honderden scholen en middelen die verschillen uit. Ter illustratie enige resultaten voor het vak Engels op vwo-niveau. De vergelijking is tussen 2003 en 2004. De vergelijking is op schoolniveau en betreft het gemiddeld cijfer van alle leerlingen van elke school in die jaren. Op landelijk niveau was het gemiddelde schoolcijfer in beide jaren gelijk: 6,9 bij het SE, en 6,4 bij het CE. Het gemiddelde verschil per school tussen 2003 en 2004 bedroeg echter 0,2 voor het SE. En 0,4 voor het CE.

Bij examens waarvan de resultaten van voldoende kandidaten beschikbaar zijn, wordt de normering gebaseerd op een statistische analyse die het verschil in moeilijkheid tussen de examens van verschillende jaren in kaart brengt. Voor veel examens geldt, dat we alleen gebruik maken van de resultaten die verzameld zijn volgens deze variant. We passen dan een 'eenvoudige' analyse toe, waarbij we ervan uitgaan dat het prestatieniveau van de groep kandidaten in één bepaald jaar niet zal verschillen van de groep kandidaten in een volgend of voorgaand jaar. Het gevolg van deze aanname is dat de normering zo aangepast kan worden, dat elk jaar het percentage onvoldoende op het examen gelijk is. Een vergelijkbare manier is de normering zo aan te passen, dat het gemiddeld cijfer dat de populatie haalt elk jaar gelijk is. Door het grote aantal kandidaten bij centrale examens, is deze vorm van normering in veel gevallen een krachtige en goed verdedigbare methode, die eenvoudig en met relatief beperkte kosten is uit te voeren. Een belangrijke voorwaarde is natuurlijk wel dat er geen ingrijpende wijzigingen in het onderwijs zijn doorgevoerd.

Bezwaar
Een bezwaar van deze procedure is, dat er geen rekening wordt gehouden met de mogelijkheid dat de vaardigheid van de groep kandidaten van jaar tot jaar kan verschillen. Het is bijvoorbeeld denkbaar dat kandidaten over een reeks van jaren gemiddeld vaardiger worden in een bepaald vak. In dat geval zou de waardering van de prestatie van een individuele leerling mede afhangen van de prestaties van de andere examenkandidaten. Immers, als andere leerlingen beter presteren, wordt het resultaat van de individuele kandidaat met een lager cijfer gewaardeerd dan wanneer de anderen minder presteren. Doordat de groep

kandidaten van jaar tot jaar verschilt, kan het voor een leerling uitmaken in welk jaar hij examen doet.

Deskundigenoordeel
Bij sommige vakken is het aantal kandidaten zo klein dat met de beschikbare gegevens geen betrouwbare statistiek bedreven kan worden. In dat geval kan het College voor Examens al vóór het examen een N-term vaststellen. Die wordt – behoudens calamiteiten zoals een onjuistheid in een opgave – na het examen als definitieve N-term vastgesteld. De N-term die op basis van het deskundigenoordeel wordt vastgesteld komt tot stand door vergelijking van het te normeren examen met het referentie-examen.