HOE IS GELIJKWAARDIGHEID VAN TOETSEN TE GARANDEREN

# Het random construeren van toetsen uit een itembank

Bernard Veldkamp

**In het voorjaar van 2011 heeft de auteur een cursusbijeenkomst verzorgd voor de NVE over het onderwerp random toetsconstructie. In dit artikel geeft hij een samenvatting van de cursusstof, waarbij hij vooral gebruikmaakt van de methode die Gibson en Weiner (1998) ontwikkeld hebben.**

## Gelijkwaardige toetsen

Een van de fundamentele problemen waar toetsspecialisten in de praktijk tegenaan lopen, is het geautomatiseerd construeren van parallelle versies van toetsen uit een itembank (Gibson & Weiner, 1998). De centrale vraag daarbij is hoe gelijkwaardigheid van toetsen te garanderen is, terwijl ze slechts door een klein aantal personen, of soms zelfs maar door één enkele kandidaat, gemaakt worden. Voor dergelijke kleine aantallen kandidaten is het zinloos om bijvoorbeeld een betrouwbaarheid uit te rekenen. De vraag is nu over welk gereedschap een toetsspecialist beschikt om tot een optimaal resultaat te komen. Dit probleem is weer actueel geworden nu veel examenorganisaties beschikken over toetssystemen die het random selecteren van toetsen uit een itembank mogelijk maken. Voor elke kandidaat of groepje kandidaten kan een individuele toets geconstrueerd worden, hetgeen grote voordelen heeft wat betreft het efficiënt gebruikmaken van alle items uit de itembank en het voorkomen van fraude door het bekendraken van veelgebruikte items. De vraag is echter in hoeverre deze random toetsen gelijkwaardig zijn en hoe voorkomen kan worden dat een deel van de kandidaten ernstig benadeeld wordt door gebrekkige examinering.

## Geschiedenis

Eerst een stukje geschiedenis. Eén van de oudst bekende methodes om met dit probleem om te gaan is de Matched Random Subsets methode van Gulliksen (1950). De methode was oorspronkelijk bedoeld om een bestaande toets op te delen in twee vergelijkbare helften om daarmee de *split-half* betrouwbaarheid uit te kunnen rekenen. Maar deze methode bleek ook goed toepasbaar te zijn bij het random construeren van toetsen. De methode werkt op de volgende manier. Allereerst worden de beschikbare items, afhankelijk van het aantal toetsen dat men wil construeren, opgedeeld in paren, drietallen of viertallen van items, die zoveel mogelijk op elkaar lijken. Daarbij kan worden gekeken naar psychometrische eigenschappen als moeilijkheid en onderscheidend vermogen, maar ook naar inhoudelijke eigenschappen. Vervolgens worden de items uit de verschillende groepjes random toegewezen aan de verschillende toetsen. Op die manier worden verschillende elkaar niet overlappende toetsen geconstrueerd, die zo veel mogelijk op elkaar lijken. Hierbij wordt zowel rekening gehouden met de psychometrische als met de inhoudelijke eigenschappen van de toets.

## Parallelle toetsen

De methode van Gulliksen resulteert weliswaar in min of meer vergelijkbare toetsen, maar daarmee is de gelijkwaardigheid nog niet gegarandeerd. Om te mogen spreken van gelijkwaardige, of in psychometrische termen van parallelle, toetsen moet voldaan worden aan een aantal voorwaarden. Uit de literatuur zijn meerdere definities van parallelle toetsen bekend. In dit artikel wordt de definitie van Lord en Novick (1968) gebruikt. Er is volgens hen sprake van parallelle toetsen als de gemiddelde scores, de variantie van de scores en de betrouwbaarheid van de toetsen gelijk zijn voor de populatie van kandidaten. Vergelijkbaarheid van de inhoud van de toetsen is daarbij ook noodzakelijk, maar dat kan worden


# In hoeverre zijn random toetsen gelijkwaardig

gegarandeerd door een vergelijkbaar aantal items te selecteren voor de verschillende domeinen.

***Item Response Theorie***

Om aan de eisen te voldoen die Lord en Novick geformuleerd hebben, heeft het voordelen om te werken met een met item response theorie (IRT) gekalibreerde itembank. De psychometrische eigenschappen als moeilijkheid en onderscheidend vermogen kunnen onafhankelijk van de populatie geschat worden volgens het Rasch model, het 2-parameter logistisch model of het OPLM tijdens het pre-testen. Deze psychometrische parameters worden samen met andere itemkenmerken vastgelegd in de itembank. Birnbaum (1968) heeft beschreven hoe vervolgens toetsen kunnen worden geconstrueerd:

1. Leg het doel van de toets vast.
2. Formuleer het doel in termen van psychometrische parameters.
3. Selecteer telkens het beste item uit de bank, totdat een toets geconstrueerd is die aan de doelstelling voldoet. Tijdens deze stap kan men eventueel rekening houden met de inhoud van de items.

Als men bijvoorbeeld een examen af wil nemen dat resulteert in een zak/slaagbeslissing over de kandidaat is het doel (stap 1) het betrouwbaar vaststellen of de kandidaat hoger scoort dan de cesuur. Vertaalt naar psychometrische parameters betekent dit dat de toets betrouwbaar moet meten rond de cesuur, oftewel in IRT termen rond de cesuur veel informatie moet verschaffen (stap 2). In stap 3 worden tenslotte die items geselecteerd die telkens de meeste informatie geven rond de cesuur. Eventueel kan men hierbij per inhoudscategorie het benodigde aantal items selecteren.

Omdat bij deze methode telkens het beste item uit de itembank wordt geselecteerd, kan het gebeuren dat de bank uitgeput raakt of dat er meer items nodig zijn voor een toets om de geformuleerde doelen te bereiken. Dat is ongewenst en daarom zijn er alternatieve methodes ontwikkeld voor geautomatiseerde constructie van parallelle toetsen. Een inleiding hierin en een overzicht van beschikbare methodes voor parallelle toetsconstructie worden gegeven in Van der Linden (2005).

Omdat de IRT parameters niet afhangen van de populatie, is voor het afnemen van de toets de psychometrische kwaliteit te bepalen. Op deze manier kan zelfs voor random samengestelde toetsen worden bepaald of ze aan psychometrische kwaliteitseisen voldoen. Is dat niet het geval dan kan men een nieuwe random toets genereren, net zo lang tot er een toets is die aan de criteria voldoet.

***Klassieke testtheorie***

In de praktijk beschikken veel examenorganisaties niet over itembanken die met IRT gekalibreerd zijn. Ook wordt IRT niet door alle toetssystemen ondersteund. In plaats daarvan hebben toetsspecialisten de beschikking over klassieke itemparameters zoals de p-waarde en de rit-waarde. De vraag is nu hoe dergelijke parameters, die populatie- en toetsafhankelijk zijn, gebruikt kunnen worden om nieuwe toetsen te construeren. Een voorstel hiervoor wordt gedaan door Gibson en Weiner (1998). Zij gaan uit van een itembank die bestaat uit deugdelijke items die gereviewed zijn door inhoudsexperts en die gepre-test zijn bij voldoende kandidaten om tot een stabiele schatting van de klassieke itemparameters als de p-waarde (pi) en de discriminatie (rit) te komen.

Hun methode kent de volgende stappen voor het random construeren van parallelle groepen:
1. Verdeel de items in sampling groepen die dezelfde inhoud hebben.
2. Selecteer random het benodigde aantal items uit elke sampling groep.
3. Bereken de verschillende psychometrische eigenschappen van de toets:

a. *Gemiddelde score* op de toets. Dit is de som van de p-waardes.

$$\mu = \sum_{\text{items}} p_i$$

b. *Standaarddeviatie* van de score op de toets. Dit is te berekenen door de som te nemen van de rit-waardes

*Bij parallelle toetsen is de betrouwbaarheid van de toetsen gelijk*


vermenigvuldigd met de standaarddeviatie van de itemscores over alle items in de test.

$$SD_{toets} = \sum_{items} r_{it}[p_i(1-p_i)]^{\frac{1}{2}}$$

c. *Betrouwbaarheid.* De betrouwbaarheid kan geschat worden met de KR20 benaderingsformule.

$$\text{KR20} = \frac{n}{n-1}\left[\frac{SD^2_{toets} - \sum_{items} p_i(1-p_i)}{SD^2_{toets}}\right]$$

4. Controleer de kwaliteit. Vergelijk daarvoor de gemiddelde score, de standaarddeviatie van de score en de geschatte betrouwbaarheid met de van te voren vastgestelde kwaliteitsstandaard. Voldoet de test er aan dan kan men de test accepteren. Voldoet de test er niet aan, dan gaat men terug naar stap 2 van de methode.

Door deze procedure toe te passen is voorafgaande aan de toetsafname al een inschatting te maken van de psychometrische kwaliteit van de toets. Het is wel belangrijk om te bedenken dat het van de nauwkeurigheid waarmee de klassieke itemparameters geschat zijn afhangt in hoeverre deze inschatting overeenkomt met de werkelijke kwaliteit na de toetsafname. Daar komt bij dat er met random itemselectie vanuit wordt gegaan dat alle items in de bank een zinvolle bijdrage leveren aan de toets. Het zal van de kwaliteit van de items in de bank en van de eisen die bijvoorbeeld aan de betrouwbaarheid worden gesteld, afhangen welk percentage van de random geselecteerde toetsen wordt geaccepteerd. Door bij te houden hoe vaak een item deel uitmaakt van een toets die niet voldoet aan de kwaliteitscriteria, kan men besluiten of dit item nog langer deel uit moet maken van de itembank, of dat het beter is het item te vervangen.

## Tot slot

Alhoewel random toetsconstructie op basis van IRT de voorkeur geniet, bieden Gibson en Weiner met hun methode een alternatief. Daar waar IRT vooraf garanties biedt voor de kwaliteit van de toets, moet men bij de methode van Gibson en Weiner wel afwachten of de kwaliteit na toetsafname overeenkomt met de geschatte kwaliteit. Het grote voordeel is echter dat deze methode ook toegepast kan worden als de itembank niet met IRT gekalibreerd is of het toetssysteem IRT niet ondersteunt.

## Literatuur


Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 397-472.

Gibson, W.J. & Weiner, J.A. (1998). Generating random parallel test forms using CTT in a computer-based environment. *Journal of Educational Measurement*, 35, 297-310.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Van der Linden, W.J. (2005). *Linear models for optimal test design.* New York: Springer.



☐ De heer dr. B.P. Veldkamp is directeur van het Research Center voor Examinering en Certificering en hoofddocent aan de Universiteit Twente. E-mail: b.p.veldkamp@utwente.nl.


**In het volgende nummer van *EXAMENS* komen onder andere de volgende onderwerpen aan bod:**

- Werken met datateams op scholen
- Slecht gelezen of slecht geleerd? – 2
- Feedback bij computergestuurde toetsen
- Een centrale itembank in de financiële sector

***EXAMENS 2, 2012 verschijnt in mei***