

De omvang van een itembank

□ Annie Kempers-Warmerdam en Annemarie de Knecht-van Eekelen

Regelmatig krijgt de Examenkamer de vraag voorgelegd wat de minimale omvang van een itembank moet zijn om die openbaar te kunnen maken. Deze vraag is makkelijker gesteld dan beantwoord. In feite bevat de vraag de volgende twee thema's: 'Hoe groot moet een itembank zijn?' en 'Is het verstandig en wenselijk om een itembank openbaar te maken?'. In dit artikel wordt gezocht naar een aantal praktische, richtinggevende antwoorden op de eerste vraag. Het tweede onderwerp, openbaarheid, wordt in een volgend nummer van *EXAMENS* behandeld.

Context

Met de introductie van het examineren per computer waarbij veelal meerkeuze vragen worden gebruikt, wordt niet meer gewerkt met één of twee unieke examens per jaar waaraan alle kandidaten gelijktijdig deelnemen. In toenemende mate worden itembanken gebruikt waaruit voor iedere kandidaat of een groepje deelnemers een uniek examen gegenereerd wordt. Hiermee heeft het rond de klok examineren op elke gewenste locatie zijn intrede gedaan.

Exameninstellingen investeren veel tijd en geld in het maken van een itembank en met name het actueel houden is kostbaar en vraagt veel expertise en energie. Het is voor een exameninstelling, met name die in de private toetspraktijk, de vraag of het mogelijk is die kosten terug te verdienen. Doorgaans zoekt een exameninstelling naar de minimale omvang van een itembank met juist voldoende items om betrouwbare uitspraken te kunnen

doen over de kennis en vaardigheden van kandidaten. Daarbij is de vraag relevant hoe vaak een bepaald item ingezet kan worden. De 'sleetsheid' of 'effectiviteit' van items is niet gelijk: immers een relatief eenvoudig of herkenbaar item kan minder vaak worden ingezet.

Omvang van een itembank

Dat 70 items geen itembank vormen in de besproken

Een casus

Enkele jaren geleden kreeg een licentie verstrekkende instelling in de industriële sector het volgende verzoek van een examencentrum voorgelegd: Het af te nemen examen bestaat uit 35 items; wij hebben bij de bestaande eind- en toetstermen conform de toetsmatrijs 70 items ontwikkeld. Hiermee kunnen we twee geheel verschillende examens samenstellen: één voor de eerste ronde en één voor de herkansing. Komen wij hiermee als 50ste examencentrum voor een accreditatie aanmerking? Het is dan de vraag welke argumenten gehanteerd kunnen worden bij het al dan niet toekennen van de licentie.

Voor de sector in deze casus is sprake van ongeveer 175.000 kandidaten op jaarbasis, waarbij in de flexibele uitvoeringsvorm wordt gewerkt met een gemiddelde groepsgrootte van 12 kandidaten. Zelfs als de 70 items de toetsmatrijs volgen, is hier geen sprake van een itembank die dekkend is voor de set van eind- en toetstermen. Met deze 70 items zouden honderden examens moeten worden samengesteld. Een aantal van 70 items bij een examen dat bestaat uit 35 vragen is ongeloofwaardig en onaanvaardbaar. De licentie zal in een dergelijke situatie moeten worden geweigerd.



Het actueel houden van de itembank kost veel geld, expertise en energie

casus, mag duidelijk zijn. Maar waar liggen de grenzen, wat is wel of niet aanvaardbaar? Ondanks het feit dat er alom door instellingen en examenexperts met itembanken wordt gewerkt, is het niet eenvoudig een eenduidig antwoord te krijgen op de vraag wat de minimale of optimale omvang van een itembank is. Is er sprake van een vuistregel of misschien van een wet van Meden en Perzen? Een inventarisatie door de Examenkamer van variabelen die een rol spelen bij de bepaling van de omvang van een itembank geeft de volgende lijst:

- * Kenmerken van het examen en items:
 - o Aantal kandidaten dat deelneemt aan de examens
 - o Aantal examens dat uit de itembank moet worden gegenereerd
 - o Frequentie waarmee examens afgenomen worden (herhaling)
 - o Aantal items in het examen
 - o De mogelijke overlap tussen examens en herexamens
 - o Taalniveau
 - o Vraagtype (complexiteit, gebruik beeldmateriaal)
 - o Actualiteit
- * Kenmerken van de toetsmatrijs
 - o Aantal eind-/toetstermen
 - o Samenstelling toetsmatrijs
- * Kenmerken van de exameninstelling
 - o Beschikbare mankracht en expertise voor het ontwikkelen van items
 - o Beschikbare hoeveelheid geld
- * Gedrag van items gedurende het gebruik

Navraag bij betrokkenen en zoeken op internet naar de wijze waarop deze variabelen worden gebruikt, heeft een zeer onvolledig beeld opgeleverd, dat wellicht door de lezers van *EXAMENS* kan worden aangevuld. De auteurs hebben niet van al deze variabelen kunnen achterhalen of en hoe ze worden gebruikt bij het samenstellen van een itembank. De meeste informatie hebben de auteurs gevonden over de vuistregels die worden gehanteerd voor het aantal items in de itembank in relatie tot de lengte van de toets. Een aantal voorbeelden volgt hierna. In geen van deze gevallen wordt duidelijk waarom voor die aantallen wordt gekozen. Een wetenschappelijke onderbouwing lijkt te ontbreken.

Kenmerken van het examen

Uit de literatuur komt een voorbeeld van een richtlijn voor het aantal items in relatie tot het aantal kandidaten. In de praktijk hanteren exameninstellingen in Nederland verschillende richtlijnen voor het gewenste aantal items. Hier volgen enkele voorbeelden:

Aantal kandidaten dat deelneemt aan de examens

Burghof, in 2001 werkzaam aan de School of Education van Flinders University (Adelaide, Australië), schreef in een artikel over het samenstellen van een itembank voor aardrijkskundetoetsen dat het aantal items minimaal 5 en maximaal 10 keer zo groot moet zijn als het aantal deelnemers aan de toets. Hij baseert dat op een artikel van Sands et al. (1997). De marges van 5-10 maken de methode flexibel, maar het aantal deelnemers aan de toets moet wel bekend zijn. Burghof geeft de volgende formule:

Item Bank Ratio = Min (Min x Ns) to Max (Max x Ns): Ns

waarin: Min = minimum aantal benodigde items;

Max = maximum aantal benodigde items;

Ns = aantal deelnemers aan de toets.

Stel dat er 100 deelnemers zijn dan is Min = 5 x 100 = 500 en Max = 10 x 100 = 1000. Dus voor 100 deelnemers is een itembank met een aantal van 500 tot 1000 items nodig. Dit is in de praktijk een zeer ruime marge.

Aantal examens of toetsafnames

Andriessen, een dienstverlener op het gebied van online examineren in Nederland, hanteert de volgende richtlijn: 'Om te kunnen starten met een itembank waaruit willekeurig een examen wordt samengesteld moet u volgens de vuistregel ongeveer 5x zoveel vragen hebben als er in één examen gesteld worden.' (Westeneng, 2009, p. 42).

Det Norske Veritas hanteert als Certificerende Instelling de volgende regel: De examenorganisatie of de schemabeheerder draagt zorg voor een itembank die minimaal drie unieke schriftelijke toetsen bevat per vakbekwaamheidsprofiel en waarvan minimaal 10% van de items jaarlijks dient te worden 'ververst'. Op basis van het aantal toetsafnames, de daadwerkelijke verversing en op aangeven van de examencommissie kan een andere omvang worden voorgeschreven. Bij een examenomvang van 21-30 items voldoet dan de itembank met

Het aantal items hangt niet alleen samen met de aard van het examen



minder dan 100 items en een jaarlijkse vervanging van 8-10 items aan de criteria.

Toetslengte

Xing & Hamilton (2004) analyseerden de invloed van een aantal parameters, waaronder het aantal items, op de kwaliteit van een itembank voor computer adaptieve toetsen. Zij vergeleken een itembank bestaande uit 240 items met een itembank met het dubbel aantal items (480). Bij een toetslengte van 35 items bleek dat zowel de kwaliteit van de afzonderlijke items als een groter aantal items de kwaliteit van de toets positief beïnvloedden.

Cito geeft aan dat voor *high-stakes* testsituaties een initiële omvang van de itembank is aan te bevelen van 8-10 keer de toetslengte. De omvang kan dan tot 20 keer de toetslengte worden uitgebreid in een tijdsbestek van 1-2 jaar.

Bij de Open Universiteit (OU) gaat men er vanuit dat het strikt genomen alleen nodig is te voorkomen dat herkansende studenten één of meer items moeten beantwoorden die zij bij een eerdere tentamen gelegenheid al onder ogen hebben gehad. Dit zou immers niet rechtvaardig zijn jegens studenten die bij de eerste tentamen gelegenheid slagen en het zou de validiteit en betrouwbaarheid van het behaalde resultaat van de herkanser kunnen aantasten. In zijn algemeenheid geldt uiteraard: hoe groter het itembestand, hoe kleiner de kans op (één of meer) identieke items bij een herkansing. Het effect van de OU-regel voor een minimale omvang van de itembank is weergegeven in tabel 1. Hierbij wordt uitgegaan van de vier gangbare tentamenlengten, namelijk 30, 40, 60 en 80 items, een maximaal toegestane overlap tussen tentamens van 25% en een variabele

multiplicatiefactor (= aantal toetsen op basis van de toetslengte). Bij een kortere toetslengte hanteert de OU een hogere multiplicatiefactor. Bovendien geldt bij de OU de regel dat als een item 50 maal is gebruikt het uit de itembank wordt verwijderd.

Relatie tot de toetsmatrijs

Voor de elektronische examens in de vastgoedsector gelden de volgende richtlijnen voor de omvang van de itembank: bij minder dan 500 kandidaten per jaar wordt de omvang van de itembank gesteld op minimaal driemaal de lengte van één toets. Als meer dan 500 kandidaten per jaar deelnemen wordt de omvang van de itembank bepaald in relatie tot de toetsmatrijs. Binnen een categorie items – dat is een groep items op hetzelfde niveau van kennis, inzicht, toepassing of integratie over een afzonderlijk onderwerp uit de toetsmatrijs – zijn minimaal vereist:

- * 12 items bij één eindterm in de toetsmatrijs,
- * 18 items bij twee eindtermen in de toetsmatrijs en
- * 22 items bij drie eindtermen in de toetsmatrijs.

Bij elke extra eindterm in de toetsmatrijs dienen steeds 2 items te worden toegevoegd.

Werk in uitvoering

Voor de industriële examens die in de casus zijn genoemd, is inmiddels een landelijke itembank ontwikkeld met een beduidend grotere omvang (aanvankelijk 800 vragen en inmiddels uitgebouwd tot 1200 vragen) waarmee, gebaseerd op een redelijk verfijnde toetsmatrijs, oneindig veel examens zijn samen te stellen. Voor alle onderwerpen in de eind- en toets termen is bepaald dat er minimaal vijf betrouwbare items op voorraad aanwe-

Toetslengte = aantal items per toets	Maximale overlap: examen versus herexamen	Aantal toetsen op basis van toetslengte = multiplicatiefactor	Minimale omvang itembank
30	25%	8	240
40 (4-keuze)	25%	7	280
60 (3-keuze)	25%	6	360
80 (2-keuze)	25%	6	480

Tabel 1 Variabelen voor het vaststellen van de omvang van de itembank zoals die worden gehanteerd bij de Open Universiteit



Het optimaal houden van de itembank is een dynamisch proces

zig moeten zijn om te voorkomen dat een item te vaak in een examen aan de orde komt. Met behulp van de toetsmatrijs kan er uit de verschillende te toetsen thema's een veelheid aan items geselecteerd worden. Het aantal items, de spreiding over de thema's en onderwerpen en het aantal te genereren examens is in voldoende balans om te garanderen dat kandidaten, al nemen ze vaker deel aan dit examen, niet hetzelfde examen voorgelegd krijgen. En niet in de laatste plaats wordt er bij deze examens via een zorgvuldige toetstechnische analyse op basis van de examenresultaten voortdurend gewaakt over de kwaliteit van de items en wordt de sleetsheid in de gaten gehouden. Door de examencommissie worden dan ook nog de ontwikkelingen in het vakgebied bijgehouden die een regelmatige aanpassing van de itembank vereisen. Zeker als het gaat om wettelijke wijzigingen of elkaar snel opvolgende ontwikkelingen is een tijdige aanpassing een voorwaarde. Het optimaal houden van de kwaliteit en de benodigde omvang van de itembank is een intensief en dynamisch proces.

Kanttekeningen

Bij een examenstelsel waar de kandidaten/studenten door inschrijving voor een langere periode/studie bekend zijn of waarbij het aantal kandidaten vooraf door een numerus fixus wordt vastgesteld, is het relatief eenvoudig om de variabelen in beeld te houden, het toetsstelsel daarop af te stemmen en de itembank daarop te optimaliseren. Ook het toetsstelsel van de OU wordt geoptimaliseerd met als doel te voorkomen dat studenten items krijgen voorgelegd die ze al een keer eerder hebben gehad. Een nieuw systeem onthoudt welke items een student al gehad heeft, zodat deze niet in de herkansingstoets worden opgenomen. De minimale omvang van de itembank kan met dit systeem kleiner zijn.

De richtlijnen van de OU zijn niet zonder meer voor andere soorten examens te gebruiken. Uit tabel 1 blijkt niet of en hoe de OU met alle variabelen rekening houdt. De hogere multiplicatiefactor bij kortere toetsen wordt wellicht ingegeven door het gegeven dat bij een kortere toetslengte de kans om items te onthouden voor kandidaten groter is. Zou het aantal toetsen niet afhankelijk moeten zijn van het aantal studenten dat de betreffende module volgt? Bij open examenstelsels zoals die in de private markt voorkomen, is de situatie zonder

meer anders. Er is veelal sprake van een korte cursusduur en er is minder zicht op de vraagontwikkelingen in de markt. Ook kan er sprake zijn van meerdere aanbieders van eenzelfde examen.

De norm die Det Norske Veritas hanteert maakt wel de minimale vereisten helder, maar er is geen duidelijkheid over andere relevante variabelen zoals het aantal toetsafnames, het aantal eindtermen enzovoort. En ook is niet duidelijk welke argumenten een examencommissie kan/moet hanteren om een andere omvang voor te schrijven. De cruciale vraag is echter of met een dergelijk aantal items een beroepsgroep op een rechtvaardige wijze te beoordelen is. Uit het voorbeeld van de industriële examens blijkt dat het mogelijk is om een itembank te ontwikkelen rekening houdend met zoveel mogelijk variabelen. Het vaststellen van de optimale omvang van een itembank is niet in één kengetal te vangen. Uitgangspunt zal zijn dat er een goede balans wordt gevonden tussen alle aspecten die in technisch en financieel opzicht een rol spelen.

Literatuur

- Burghof, K.L. (2001). Assembling an item-bank for computerised linear and adaptive testing in geography. *International Education Journal* 2(4). Educational Research Conference 2001 Special issue. Te downloaden op <http://ehlt.flinders.edu.au/education/iej/articles/v2n4/BURGHOF/PAPER.PDF>.
- Roode, F.A. de. (2010). Persoonlijke mededeling.
- Sands, W., Waters, B., & McBride, J. (1997). *Computerized adaptive testing: from inquiry to operation*. Washington DC, USA: American Psychological Association.
- Xing, D., & Hambleton, R.K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, 64(1), 5-21.
- Westeneng, P. (2009). *Online toetsen en examineren*. Oisterwijk: Andriessen en Partners. Te downloaden op http://www.andriessen.nl/downloads/online_examinieren_2009_01-16.pdf.

□ Mw. dr. A. Kempers-Warmerdam is directeur van de Stichting Examenkamer en redacteur van EXAMENS. Mw. dr. A. de Knecht-van Eekelen is hoofdredacteur van EXAMENS. E-mail: secretariaat@examenkamer.nl.