HO 1c Toetslengte, antwoordmogelijkheden en afnametijd

Introductie

De toetsmatrijs (zie handout 1b) geeft inzicht in het aantal vragen nodig om de leerdoelen op een valide en betrouwbare manier te toetsen. In deze handout staat hoe bepaald wordt hoeveel vragen worden opgenomen in een toets (toetslengte). Dit hangt natuurlijk sterk af van de omvang van de te dekken leerstof en ook van de raadkans en het aantal antwoordmogelijkheden. Bovendien is er vaak maar een beperkte toetstijd beschikbaar, die al tevoren is ingeroosterd. Ook die afnametijd bepaalt de toetslengte.

- *Hoeveel vragen moet ik opnemen in een toets? En, hoe zit dat als de toets bestaat uit drie-, of vierkeuzevragen?*
- *Hoe bereken ik vooraf hoeveel tijd studenten voor de toets nodig hebben?*

Toetslengte

Vijf factoren zijn bepalend voor de lengte van de toets. De functie van de toets (high stakes of low stakes), de mate van dekking, de verwachte betrouwbaarheid, de moeilijkheidsgraad van de vragen, en de afnametijd.

1. **De functie van de toets:** gaat het om een oefentoets, een formatieve toets of een summatieve toets?

Bij een oefentoets is het doel dat studenten een idee krijgen van de soort vragen die in de summatieve toets zoal worden gesteld, zodat zij zich daarop kunnen voorbereiden.
Criterium voor het aantal vragen in de oefentoets is in dit geval net zoveel vragen als nodig om een representatief beeld te geven van de variatie aan vraagvormen, die in de summatieve toets worden voorgelegd. Over het algemeen zal een set van 10 vragen al voldoende zijn om in de oefenfunctie te voorzien.
Bij een formatieve toets (low stakes test), gaan we net een stap verder dan de oefentoets. Het gaat nu nog niet om slagen of zakken, maar wel om een redelijk betrouwbare en valide inschatting in welke mate en op welke onderdelen de student de leerstof naar inhoud en niveau beheerst, en waar verbetering nodig is. Studenten kunnen zelf conclusies verbinden aan de resultaten op toets(onderdelen), maar gerichte feedback en aanwijzingen van de docent kunnen ook onderdeel uitmaken van de formatieve toets. Het aantal vragen voor een formatieve Mc-toets hangt af van de feedback die je wilt geven op de onderdelen die er toe doen. Daarnaast wekt de formatieve toets ook verwachtingen over de summatieve toets en zal daar dus mee vergelijkbaar moeten zijn. Want dan kunnen studenten zich verbeteren op die onderdelen waar ze uiteindelijk op beoordeeld worden.
Afhankelijk van de filosofie van de opleiding en hoe wordt ingezet op formatieve feedback kan het zijn dat de summatieve toets slechts een steekproef is uit de (vele en omvangrijke) formatieve toetsen die daaraan zijn voorafgegaan. In ons onderwijs is het omgekeerde eerder het geval: de formatieve toets is een representatieve deelverzameling van de uitgebreidere summatieve toets.
Bij de summatieve toets (high stakes test) gaat het om belangrijke zak/slaag-beslissingen die op grond van het toetsresultaat worden genomen. De factoren 2 t/m 5 spelen een rol bij het bepalen van de lengte van de toets.

2. **De validiteit:** de toets dekt de cursusdoelen en is congruent met de studie-activiteiten van de student. De omvang en aard van de doelen zullen leidend zijn voor het aantal op te nemen vragen in de toets en de toetsmatrijs is het instrument om tot een doeldekkende toets te komen. In HandOut 1b is het Wat en het Hoe van de toetsmatrijs uitgelegd.

3. **De betrouwbaarheid** van de toets: de toets bestaat uit zoveel vragen dat een betrouwbaar oordeel kan worden geveld. Wordt de toets voor het eerst afgenomen dan moet achteraf uit de toetsanalyse blijken of de toets ook daadwerkelijk voldoende vragen bevat om de verschillen tussen studenten tot uitdrukking te brengen. De regel is hoe meer vragen, des te betrouwbaarder (consistenter) de toets. De relatie tussen lengte en betrouwbaarheid is in tabel 1 weergegeven. Stel dat de gemeten betrouwbaarheid (α) van een toets 0.60 is en 0.75 de beoogde betrouwbaarheid, dan zal de toets twee keer (2K) zoveel soortgelijke vragen moeten bevatten. Het zal duidelijk zijn dat als daarvoor wordt gekozen ook de afnametijd moet worden aangepast.

Tabel 1: Betrouwbaarheidsschatting (α) bij toetsverlenging/inkorting, bij K-vragen

K	1.5 K	2 K	3 K
α	α'	α''	α'''
0.20	0.27	0.33	0.40
0.40	0.50	0.57	0.60
0.60	0.69	0.75	0.77
0.80	0.86	0.89	0.91

4. **Aantal alternatieven**, de raadkans en doel/moeilijkheid van de vragen,
In de praktijk van het toetsen met MC-vragen wordt bij het vaststellen van de lengte van de toets rekening gehouden met raadkans. Hoe meer alternatieven des te kleiner de raadkans, zo is de logische redenering, en zo zal een toets die uit stellingvragen bestaat meer vragen moeten bevatten dan een toets met vierkeuzevragen.
In tabel 2 is een indicatie gegeven van het aantal op te nemen vragen in een toets die een vergelijkbaar meetbereik (30) hebben bij een verschillend aantal alternatieven (twee-, drie- en vierkeuzevragen).

Tabel 2: Relatie tussen aantal alternatieven, toetslengte en meetbereik bij mc-toetsen
Bron: Milius (2007)

Aantal alternatieven(A)	Toetslengte (N)	Raadkans (R=N:A)	Meetbereik (N-R)
Vier	40	10	30
Drie	45	15	30
Twee	60	30	30

We hebben het hier over de theoretische en gemiddelde raadkans op basis van blind raden. Gelukkig zullen studenten niet blind raden, maar op basis van de kennis die ze hebben en de informatie die de vraag bevat.
Met het toevoegen van een alternatief wordt er ook extra informatie gegeven, op basis waarvan de student juist gemakkelijker tot het goede antwoord kan komen. De kwaliteit (en moeilijkheid) van de vraag zal van invloed zijn op het vaststellen van een beredeneerde raadkans. Hieronder voorbeelden van drie vragen met elk een theoretische raadkans van 50%.

Voorbeeld a ***Stelling vraag*** (raadkans 50%)	***Voorbeeld b*** ***Twee-keuzevraag*** (raadkans 50%)	***Voorbeeld c*** ***Twee-keuzevraag*** (raadkans 50%)
Montevideo is de hoofdstad van Uruguay A. Juist B. Onjuist	Montevideo is de hoofdstad van: A. Uruguay B. Paraguay	Montevideo is de hoofdstad van: A. Uruguay B. Frankrijk

Voorbeeld a en b zijn weliswaar in raadkans gelijk, maar in b is er meer informatie gegeven. Datzelfde geldt voor voorbeeld c, die van de drie vragen het gemakkelijkst is te beantwoorden. Tegelijkertijd is voorbeeld c mogelijk een prima vraag als het zou gaan om het onderwerp 'Landen en hoofdsteden van de wereld' (basisonderwijs). Kortom, de moeilijkheidsgraad is gekoppeld aan het doel en is van invloed op de raadkans. Een beredeneerde raadkans/vraag is direct gerelateerd aan de cesuurbepaling. In HandOut 1b is een eenvoudige werkwijze opgenomen hoe de moeilijkheid-, beheersingsgraad verdisconteerd kan worden in de cesuur.

5. De afnametijd

Is het aantal vragen bekend (punten 1 t/m 4), dan gaat het er vervolgens om een realistische en zo goed mogelijke inschatting te maken van de benodigde afnametijd. Die tijd is bij voorkeur niet te ruim bemeten, omdat daarmee een verkeerd signaal wordt afgegeven aan studenten. Als studenten bijvoorbeeld 3 uur de tijd krijgen voor een MC-toets van 20 vragen, dan kan dat ten koste gaan van de geloofwaardigheid. De tijd is dus goed afgemeten aan de geschatte maaktijd en is meer dan ruimvoldoende.
In ons onderwijs is uitgangspunt dat de studenten voldoende tijd moeten hebben om de toets te kunnen maken. Snelheid van oplossen van problemen mag dus geen rol van betekenis spelen. Vuistregel is dat een 'gemiddelde student' in ⅔ van de tijd de toets kan afleggen. Is dat bijvoorbeeld 60 minuten dan is 90 minuten een redelijke tijdsduur voor alle studenten.
Los van deze overwegingen is er een indicatie te geven voor de tijd die gemoeid is met het beantwoorden van een 2-, 3-, 4-, of 5-keuzevraag.
Vuistregel voor de benodigde tijd en het aantal alternatieven.
Studenten zijn meer tijd kwijt naarmate het aantal alternatieven toeneemt. Zo is de student gemiddeld 50 seconden kwijt met het beantwoorden van een tweekeuzevraag, 60 seconden met een driekeuzevraag en 75 seconden met een vier- of vijf keuzevraag (Van Berkel & Bax, 2013). Dit zijn gemiddelden, en het zal duidelijk zijn dat deze tijdsindicatie niet opgaat voor vragen waar de studenten een berekening moeten uitvoeren, of een casus met ruwe gegevens moet lezen en interpreteren.