HO 4a Toetsanalyse

Introductie

Na het afnemen van de toets en het analyseren van de scores komt het vaststellen van de uitslag. Dat gaat eerst globaal, daarna wordt een diepgaandere analyse uitgevoerd waarbij de kans bestaat dat er nog wijzigingen in de cesuur worden aangebracht. Dit is de laatste stap, voordat de resultaten naar de studenten kunnen worden gecommuniceerd. Centrale vragen in deze fase zijn:

- *Hoe weet ik of de multiple choice toets goed was?*
- *Differentieert de toets voldoende?*
- *Is er voldoende variatie in moeilijkheidsgraad aangebracht?*
- *Hoe is het gesteld met de kwaliteit van de afzonderlijke vragen?*
- *Wat te doen als de resultaten tegenvallen/de kwaliteit van de toets(vragen) onder de maat is?*

Belang van de toetsanalyse

Het doel van de toetsanalyse is allereerst evidentie te leveren voor de betrouwbaarheid van de oordelen over de geleverde prestaties van de studenten. Daarnaast geeft de analyse een indicatie van de kwaliteit toets als geheel en van de afzonderlijke vragen. Voor de toets zelf kunnen er nog wijzigingen worden aangebracht door bijvoorbeeld meerdere alternatieven goed te rekenen, een vraag te verwijderen uit de toets en/of de voldoende-onvoldoende grens (cesuur) aan te passen. De analyse geeft informatie over de vraagkwaliteit zoals moeilijkheidsgraad en het onderscheidend vermogen. Die informatie is van belang om de vraag geschikt te bevinden voor de vragenbank of dat er nog aan geschaafd moet worden.
De analyse geeft antwoord op de volgende vragen: is de meting voldoende consistent, differentieert de toets wel in voldoende mate, is er voldoende variatie in moeilijkheidsgraad aangebracht, is het aandeel inconsistente beslissingen verdedigbaar?

Een werkwijze voor screening van de toetskwaliteit: van eerste indruk tot een screening van de vraagkwaliteit in detail

Vragen vooraf

1) **Gaat het om een eerste afname of een aanvullende toets? En is het aantal deelnemende studenten > 60?**
   De richtlijnen die verderop worden gegeven voor een goede en onderscheidende toets of een goede toetsvraag zijn direct afhankelijk van de samenstelling van de groep waar de toets bij is afgenomen. Dat kan nogal verschillen. We gaan er hier van uit dat de toets een selectieve functie heeft en dat het doel is de verschillen tussen studenten in beheersing van de leerstof zo nauwkeurig mogelijk in scores weer te geven. Maar, als er minder verschil is tussen de studenten (stofbeheersing) dan zal het onderscheid dat in de resultaten van de toetsanalyse wordt uitgedrukt in de diverse prameters naar verwachting ook niet groot zijn, de betrouwbaarheid en kwaliteit valt dan tegen. Wanneer je dus een kleine, of homogenere studentenpopulatie hebt dan zal naar verwachting de kwaliteit van de vragen lager zijn dan gedacht en ook de betrouwbaarheid overall lager uitvallen dan gehoopt. Vragen die vooraf van belang zijn om de analyseresultaten op waarde te kunnen schatten zijn dan van belang.
2) **Is bij studenten al iets bekend over de norm?**
   Zijn er al verwachtingen gewekt over de norm? Met het gebruiken van analyseresultaten om achteraf aanpassingen aan te brengen in de toets of cesuur is het van belang dat de aanpassing strookt met de informatie die vooraf aan studenten is beloofd, dan wel dat de aanpassing een gelijke of verbetering van de individuele studentprestatie tot gevolg heeft.
3) **Is dit de enige toets waarop het eindcijfer wordt gebaseerd, of zijn er meer toetsen/opdrachten die meetellen?**
   Bij de uitleg van de betekenis van de normen voor een optimale betrouwbaarheid, of vraagkwaliteit is ervan uitgegaan dat de beslissing over zakken en slagen geheel of voornamelijk op de uitslag van deze toets is gebaseerd. Meestal zijn er meer prestaties geleverd door de studenten op een gevarieerde mix aan toetsvormen. In die gevallen kunnen de gehanteerde normen naar rato worden bijgesteld.

De uitleg hieronder is gebaseerd op de toets- en itemanalyse en zoals gegenereerd door Remindo.

Stappenplan gebruik van de analyseresultaten

Stap 1: De eerste, algemene indruk
Is het slaagpercentage conform verwachtingen? En, is de betrouwbaarheid van de toets voldoende? NB: Is het de enige toets waarop het cursusoordeel is gebaseerd?
Zo ja, dan volstaat een Quick Scan (1A)! Zo niet, ga verder met Stap 2.
Stap 1A) Quick-scan*: Kwaliteit van de vragen (aanpassing voor opname in de itembank)*
Gebruik de item-analyse om mogelijk dubieuze vragen te markeren (P" en Rir < 0.1). Kijk naar die afleiders die aantrekkelijk blijken voor veel studenten (a-waarde) en de goed presterende studenten (z-waarde of rar-waarde). Zoek naar een verklaring voor het afwijkend scorepatroon door naar de inhoud van de vraag te kijken. Is de formulering van de vraag en de alternatieven te verbeteren?

Stap 2: Bij een tegenvallend slaagpercentage
Voor je meteen de diepte ingaat (inspectie vraagkwaliteit) is het verstandig om te kijken of een marginale aanpassing al zou kunnen leiden tot een grote stijging van het percentage geslaagden. Is dat het geval dan haalt dat veel druk van de ketel. Daarvoor zou je de volgende vragen kunnen stellen: Bij welke cesuur zou het slaagpercentage wel volgens verwachting (aanvaardbaar) zijn? Hoeveel scheelt dit met de gehanteerde cesuur, is met een kleine wijziging een groot effect te bewerkstelligen?

Stap 3: Deep Scan*: Maak gebruik van de uitleg, de interpretatie van analyseresultaten*
- Bij een laag slaagpercentage vooral gericht op (te) moeilijke vragen: P"-waarde < 0.15
- Bij een laag slaagpercentage en een lage betrouwbaarheid: gericht op vragen met een (te) lage Rir-waarde < 0.10 èn P"-waarde die laag is (< 0.10).

Zoek naar een afwijkend scorepatroon (altijd op basis van inhoudelijke argumenten!):
- Bij negatieve Rir-waarde: zoek naar het alternatief met een positieve Z of Rar-waarde;
- Bij een lage P"-waarde: zoek naar de aantrekkelijke afleider (veel gekozen en dus een relatief hoge a-waarde).

In beide gevallen is de vraag of er niet iets te zeggen is voor het eveneens goed rekenen van dit alternatief, of zelfs alle alternatieven?
Bij het beoordelen van de kwaliteit van de vragen is het goed om de betekenis van beide parameters juist te interpreteren. Het doel is zo tot eventueel verantwoorde maatregelen te komen. In de Uitleg en interpretatie van de analyseresultaten is in meer detail aangegeven wat de betekenis is van de parameters die in de analyserapportage staan.

Uitleg en Interpretatie van analyseresultaten

Zo nodig aanpassen van de toets, de vragen en de cesuur.

Betrouwbaarheid: Coëfficiënt alfa (α)

- Max: 1; Min: 0
- Streefwaarde: 0.6-0.8

De coëfficiënt α is een maat voor de betrouwbaarheid van de toets. Hoe betrouwbaarder de toets des te nauwkeuriger de scores geïnterpreteerd kunnen worden. De coëfficiënt α kan maximaal 1 aannemen (volledig betrouwbaar) en minimaal de waarde 0 (volkomen onbetrouwbaar: scores zijn toevallig tot stand gekomen). Voor een toets wordt een α nagestreefd van 0.8 om toch in redelijke mate een uitspraak te kunnen doen over het kennisniveau van de student (*high stake* test). Is het de enige toets waarop het oordeel is gebaseerd, dan is de 0.8 norm het streven. Is de beslissing (geslaagd-gezakt voor de cursus) mede gebaseerd op het resultaat van andere toetsen (open-vragen, tussentoetsen dan is een lagere betrouwbaarheid dan de gewenste 0.8, verdedigbaar. Regel is dat hoe langer de toets, des te beter de differentiatiegraad en des te hoger de betrouwbaarheid.
De analyse van een herkansing is een geval apart en ligt het in de rede dat de betrouwbaarheid lager uitvalt (0.4-0.6) omdat de onderlinge verschillen in de populatie geringer zijn dan bij een eerste afname.

Tabel 1: Percentages niet-consistente beslissingen als functie van afwijzingspercentage en toetsbetrouwbaarheid (α).
Bron: Dousma, Horsten, Brants, Tentamineren (1997).

Afwijzings% *(gezakt)*	Betrouwbaarheid (α)						
	0,50	**0,60**	**0,70**	**0,80**	**0,90**	**0,95**	**1,00**
5	8	7	6	5	4	3	0
10	14	12	11	9	6	4	0
15	18	17	14	12	8	6	0
20	23	20	17	14	10	7	0
25	26	23	20	16	11	8	0
30	29	25	22	18	12	9	0
35	31	27	23	19	13	9	0
40	32	29	24	20	14	10	0

45	33	29	25	20	14	10	0
50	33	30	25	20	14	10	0

Nauwkeurigheidsmarge : de Standaardmeetfout (S_m)

- Streefwaarde: Lager dan 10% van de maximumscore.

De standaardmeetfout is een nauwkeurigheidsmaat en geeft aan wat de waarschijnlijkheid is dat gemeten scores overeenkomen met 'feitelijke' kennis bij studenten. De standaardmeetfout (S_m) is direct afhankelijk van de spreiding in toetsscores (S_A) en de betrouwbaarheid (α): $Sm = Sa\sqrt{1-\alpha}$.
Is de toets onbetrouwbaar dan kan aan de gemeten scores geen betekenis worden gehecht. Is de standaardmeetfout bijv. 2, dan betekent dit voor een student met een gemeten score van 13, dat hij/zij met 67 % waarschijnlijkheid kennis heeft overeenkomend met een score van 13 ± 2: tussen de 11 en 15.

Moeilijkheidsgraad: de Proportie goed (P-, en P"-)waarde

- P: Max 1; Min 0
- P" : Max: 1; Min: -1
- Streefwaarde: hoger dan 0.10

Zowel de P- als de P'-waarde geeft de moeilijkheidsgraad aan van een vraag. De p-waarde is een relatieve maat, de P'-waarde een absolute maat. Bijvoorbeeld: een vraag heeft een p-waarde van 0.5. Dat wil zeggen dat 50 % van de studenten het juiste alternatief heeft gekozen. Op zich zegt dit nog niets, want hoe dit getal te interpreteren is afhankelijk van het aantal alternatieven. Bij een 4-keuzevraag zal de vraag scheidend zijn. Maar bij een 2-keuzevraag komt het percentage overeen met de raadkans! De p'-waarde is een gecorrigeerde p-waarde voor raden en is een preciezere indicatie voor het deel van de studenten die het antwoord werkelijk wisten i.p.v. het goed hadden op de gok. De P" wordt als volgt berekend:
P"= P_g-($\frac{\text{Pf}}{\text{afleiders}}$) = P -($\frac{1-P}{a-1}$) . Waarbij P staat voor Proportie*goed* (P_g): het aandeel van de studenten die het goede antwoord hadden; 1-P staat voor de Proportie*fout* (P_f) het aandeel van de studenten die voor een afleider kozen; en a-1 staat voor het aantal afleiders (aantal alternatieven – 1).
De redenering is vervolgens dat de p-waarde gecorrigeerd moet worden met de proportie raders voor het goede alternatief, waarvoor de proportie*fout* per afleider de beste graadmeter is.

Voorbeeld: een vraag heeft een p-waarde = 0.60
P" bij een 2 keuze vraag = 0.20. Want: 0.60 – (0.40:1)
P" bij een 3-keuze vraag = 0.40. Want: 0.60 – (0.40:2), en
P" bij een 4-keuze vraag = 0.47. Want: 0.60 – (0.40:3).

Heeft de p' een waarde van bijna 1, dan is de conclusie dat het een zeer gemakkelijke vraag was, en dus naar waarschijnlijkheid ook een weinig scheidende vraag is (scheidt alleen de 2-en van de enen). Eveneens kun je dit zeggen bij extreem lage p'-waarden: de vraag is zeer moeilijk en scheidt slechts de 1O-en van de negens. Het gaat er natuurlijk niet om de enen van de tweeën en de tienen van de negens te scheiden. Maar juist om de 5-en van de zessen te kunnen onderscheiden. Dit vraagt om een gevarieerd beeld aan p-waarden, waarvan alle waarden tussen 0.1 en 0.9 vertegenwoordigd zijn met een zwaartepunt bij vragen met een p'-waarde van 0.5-0.6.

Aantrekkelijkheid van de afleiders: de a-waarde

- Streefwaarde: de som van de a-waarden (proportie fout), is lager dan de p-waarde (proportie goed) van de vraag.

De a-waarde geeft de proportie (van 0 tot max. 1.0) weer van de studenten die voor de betreffende afleider heeft gekozen. De a-waarde is te vergelijken met de p-waarde voor het goede antwoord, maar dan voor de afleiders. Is de a-waarde vergelijkbaar of hoger dan de p-waarde, dan is deze afleider zeer aantrekkelijk geweest (misschien een instinkerdje?). Is de a-waarde laag, dan trekt de afleider weinig studenten, is de afleider niet effectief gebleken. In zo'n geval (bijvoorbeeld geen of slechts één enkele student heeft de afleider gekozen) is er iets voor te zeggen de afleider voor hergebruik te schrappen (4-keuzevraag is een 3-keuzevraag), òf een betere afleider te formuleren.

Onderscheidend vermogen: de Rir-waarde

- Max: 1; Min: -1

- Streefwaarde: positief, hoger dan 0.10

De p-waarde geeft het aandeel weer van de studenten die de vraag goed hadden. De Rir-waarde geeft aan in hoeverre de vraag de goede van de slechte studenten heeft gescheiden. Is de Rir hoog (0.3-0.5) dan heeft de vraag zijn werk gedaan: de goede studenten hebben de vraag goed, en de slechte studenten kiezen voor een afleider. Wordt de Rir negatief dan kan er iets aan de hand zijn. Juist de goede studenten kiezen voor een afleider, terwijl de slechtere studenten voor het juiste alternatief hebben gekozen. Als de Rir negatief is betekent dit per definitie dat één van de afleiders positief correleert (een positieve z-waarde). Gecontroleerd moet worden of er niet iets te zeggen valt voor die afleider, immers de best presterende studenten kiezen daar tenslotte voor.

Tabel 2: Interpretatie van mogelijke combinaties P' en Rir-waarden

	Rir < dan 0.1	***Rir > dan 0.1***
P' < dan 0.1	Sleutel correct? Detail? Vraagformulering eenduidig? Ander alternatief ook plausibel?	Instinkerdje? Te moelijk / te complex?
P' tussen 0.1 en 0.8	Ander alternatief ook waarschijnlijk (meerdere alt. goed rekenen)?	In orde
P' > dan 0.8	Weggever (op te lossen met boerenverstand)?	Behoeft geen actie

Discriminerend vermogen van alternatieven: de z-waarde / rar-waarde?

- Max: zelden hoger dan 3; Min: zelden lager dan -3.
- Streefwaarde: positief voor juiste alternatief, negatief voor afleiders.

De z-waarde geeft vergelijkbare informatie als de Rir, en zegt iets over het onderscheidend vermogen van een alternatief. De z-waarde is uitgedrukt in standaardmeeteenheden. Naarmate het verschil tussen de z-waarden op twee alternatieven groter is, des te waarschijnlijker dat deze alternatieven bijdragen aan een gemeten kennisverschil (verschil = 1, waarschijnlijkheid 67%; verschil 2, waarschijnlijkheid 95%).