UNIVERSITEIT VAN AMSTERDAM

Toetsanalyse

Leidraad

Digitaal Toetsen FGw

Inhoudsopgave

1 Toetsanalyse

Bij de analyse van een digitale toets kan gebruik worden gemaakt van verschillende informatiebronnen. Hieronder staan de meest voorkomende uitgelegd.

Let op! Voor een betrouwbare analyse van een toets heb je minstens 100 afnames nodig, maar vanaf 25 afnames krijg je al wel een indruk van de moeilijkheid van de toets.

1.1 *P*-waarde

De *p*-waarde van een toets is de gemiddelde moeilijkheid van een toets, dat wil zeggen de proportie voldoendes op de toets als geheel. Hoe hoger de waarde, hoe makkelijker de toets. Een voorbeeld: als een toets een *p*-waarde heeft van 0.60, heeft 60% van de studenten een voldoende behaald.

In principe zouden studenten hun toetsen moeten kunnen halen. Als dit niet lukt, kunnen de oorzaken bijvoorbeeld worden gezocht in tekortkomingen in de toets of in het onderwijs en / of door onvoldoende inspanning van de student.

In de propedeuse wijst een tentamen met meer dan 30% onvoldoendes op niet-student-gerelateerde tekortkomingen in de toets of in het voorgaande onderwijs. In de hoofdfase zou 90% van de studenten voor een toets moeten slagen.

Dit betekent dat de *p*-waarde van een digitale toets in de propedeuse idealiter 0.70 bedraagt en die van een toets in de hoofdfase 0.90 telt. In de gevallen van de *p*-waarde beduidend lager / hoger uitvalt, is het goed om de oorzaak te zoeken in tekortkomingen in de toets of in het onderwijs. Als de *p*-waarde lager is, kan dit bijvoorbeeld betekenen dat de toets misschien te moeilijk was ten opzichte van de lesstof. Als de *p*-waarde beduidend hoger uitvalt, kan het zijn dat de toets wellicht te makkelijk is geweest.

1.2 Betrouwbaarheid Alpha/KR-20

De waarde van de betrouwbaarheid van een toets ligt altijd tussen 0 en 1. Hoe dichter deze waarde bij de 1 ligt, hoe betrouwbaarder de toets is. Betrouwbaarheid kan worden gedefinieerd als de mate waarin de toetsscores consistent, nauwkeurig en reproduceerbaar zijn, ofwel vrij van meetfouten.

De norm voor de betrouwbaarheid van een digitale toets is afhankelijk van het doel van de toets. Als het tentamen bedoeld is om de geschiktheid van een student te bepalen, is een betrouwbaarheid van 0.80 of hoger gewenst. Dit is van toepassing als alle toetsen met een voldoende moeten worden afgesloten en compenseren niet mogelijk is. Als het gaat om formatieve toetsen die het leren slechts ondersteunen, is een betrouwbaarheid van 0.60 voldoende. Dit is ook het geval bij een compensatorische examenregeling.

Er zijn een aantal kanttekeningen te plaatsen bij de betrouwbaarheid als kengetal voor de kwaliteit van een toets:

- De betrouwbaarheid is lager naarmate de toets heterogener is, dat wil zeggen verschillende soorten kennis en vaardigheden meet.
- De betrouwbaarheid wordt lager naarmate de groep studenten homogener is, dat wil zeggen als de verschillen in niveau van studenten klein zijn. Dat is vast te stellen door het verschil in scores tussen de 5% beste en 5% slechtste studenten.

- Het kengetal voor de betrouwbaarheid is de laagste ondergrens. In werkelijkheid kan de betrouwbaarheid hoger zijn.
- De betrouwbaarheid wordt groter als het tentamen meer items bevat.

2 Itemanalyse

Bij alle toetsvormen wordt informatie over de kwaliteit van toetsvragen verzameld. Hieronder wordt ingegaan op de *p*-waarde, *a*-waarde en de *Rit*- en *Rir*-waarde.

Let op: Statistische maten met betrekking tot items zijn alleen betekenisvol bij voldoende grote aantallen studenten. Meestal wordt als norm meer dan 50 aanbevolen.

2.1 *P*-waarde

Bij het beoordelen van de kwaliteit van een item moet worden bekeken wat de *p*-waarde is. De *p*-waarde is de proportie correcte antwoorden op een individueel item. De *p*-waarde varieert van 0 (iedereen fout) en 1 (iedereen goed).

Normen voor p-waarden bij summatieve toetsen

	Optimale *p*-waarde	**Ondergrens**	**Bovengrens**
bij tweekeuzevragen	0.75	0.61	0.90
bij driekeuzevragen	0.67	0.50	0.90
bij vierkeuzevragen	0.62	0.44	0.90
bij open vragen	0.50	0.25	0.90

Bron: Berkel van, H., Bax, A. & Joosten-ten Brinke, D. (2014). *Toetsen in het hoger onderwijs.* Houten: Bohn Stafleu van Loghum.

2.2 *A*-waarde

Bij meerkeuzevragen kan worden gekeken naar de afleiders (*a*-waarden). De *a*-waarden geven aan hoe vaak een afleider is gekozen. Als een afleider niet of nauwelijks wordt gekozen, dan is de kwaliteit ervan waarschijnlijk onvoldoende en wordt aangeraden de afleiders onder de loep te nemen.

2.3 *Rit* en *Rir*-waarde

De *Rit*-waarde geeft het onderscheidend vermogen weer van een item en staat voor de correlatie tussen het item en de totaalscore op de toets.

De totaalscore op de toets bevat ook de score op het item waarmee je wilt correleren. Dat vertekent de correlatie op een rooskleurige manier. de *Rir*-waarde geeft de correlatie (R) weer tussen het item en de totaalscore minus de score van de betreffende vraag (restwaarde). Op deze manier wordt een eerlijkere weergave gegeven van het onderscheidend vermogen van een item.

Een vraag heeft een hoog onderscheidend vermogen (*Rir* meer dan 0.25) als studenten met een hoog toetsresultaat de vraag goed maken en de studenten met een laag resultaat de vraag fout maken. Het item differentieert dan tussen de goed presterende en minder goed presterende studenten.

Als studenten van alle niveaus hetzelfde scoren op een vraag, dan is er geen onderscheidend vermogen (*Rir* is 0).

Als studenten met een lage toetsscore de vraag goed hebben en studenten met een hoog resultaat hebben de vraag fout (*Rir* is negatief), dan is er waarschijnlijk iets aan de hand, zoals een foute antwoordsleutel of een onduidelijke formulering van de stam van de vraag.