UNIVERSITEIT VAN AMSTERDAM

# Leidraad Tentamenanalyse

Digitaal Toetsen & Toetsdeskundigen FGw

**Contact:** goedtoetsen-fgw@uva.nl

# 1 Quickscan

De kwaliteit van een toets kan worden bepaald nadat deze is afgenomen. Zo'n 'psychometrische analyse' kan helpen de beoordeling van de resultaten scherper te krijgen en de juiste zak/slaaggrens te vinden.

De tabel hieronder vat de belangrijkste indicatoren van zo'n analyse samen. In de rest van het document worden de indicatoren verder uitgelegd en vind je mogelijkheden de toets te verbeteren na analyse.

| **Indicator** | **Uitleg** | **Optimale waarden** |
|---|---|---|
| p-waarde toets | Deze waarde tussen 0 en 1 drukt de moeilijkheid van de toets als geheel uit. Hoe hoger het cijfer, hoe makkelijker de toets. | Propedeuse: tussen 0,65 en 0,75<br>Postpropedeuse: tussen 0,85 en 0,95 |
| Betrouwbaarheid Alpha (KR-20) | Deze waarde tussen 0 en 1 drukt de betrouwbaarheid van de toets als geheel uit. Hoe hoger het cijfer, hoe betrouwbaarder de toets. | 0,8 of hoger is goed<br>0,7 of hoger is voldoende<br>0,6 of lager is onvoldoende |
| p-waarde per vraag (item) | Deze waarde tussen 0 en 1 drukt uit hoe moeilijk studenten de vraag vonden. Lager is moeilijker. | Varieert per vraagsoort, zie hieronder. |
| $R_{it}$- en $R_{ir}$-waarden | Deze waarden tussen -1 en 1 drukt het onderscheidend vermogen van een vraag uit. Hoe hoger het cijfer, hoe beter de vraag kan onderscheiden tussen sterke en zwakke studenten. | 0,35 of hoger is zeer goed<br>0,25-0,35 is goed<br>0,15-0,25 is voldoende<br>0,15 of lager is onvoldoende |

# 2 Kwaliteit van de toets als geheel

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 kwaliteit van de toets.

**Betrouwbaarheid Alpha/KR-20**

De belangrijkste indicator voor de betrouwbaarheid van de toets als geheel is de zogenaamde Alpha of KR-20. Betrouwbaarheid kan worden gedefinieerd als de mate waarin de toetsscores consistent, nauwkeurig en reproduceerbaar zijn, ofwel vrij van meetfouten. Alpha wordt uitgedrukt in een cijfer tussen 0 en 1: hoe hoger de waarde, hoe beter.

- Voor een summatieve toets is een waarde van 0,8 of hoger gewenst; minder dan 0,7 is matig en minder dan 0,6 is onacceptabel.
- Bij formatieve toetsen, of als toetsen onderling kunnen worden gecompenseerd, is een betrouwbaarheid van 0,6 of hoger, matig maar acceptabel.

Als de betrouwbaarheid te laag is (kleiner dan 0,7 of 0,6) kun je zoeken naar oorzaken. Er kan bijvoorbeeld een fout in het antwoordmodel zitten, of een vraag blijkt achteraf niet goed gesteld (voor meer informatie hierover zie '3. Vraagkwaliteit'). Na correctie van deze fouten, zie je de Alpha vaak weer stijgen. Bij toetsen met weinig vragen is de Alpha vaak laag: hoe meer items je toets bevat, hoe groter de kans is dat je een goede Alpha hebt.

Bij de Alpha past wel een relativering. Hij kan wat lager uitvallen als een toets veel verschillende soorten kennis en vaardigheden meet. En bij een herkansing zegt hij niet veel omdat dan de studentengroep niet representatief en vaak te klein is.

**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, kan dit wijzen op tekortkomingen in de toets of in het onderwijs en / of door onvoldoende inspanning van de student.

Een vaak genoemde vuistregel voor de propedeuse is, dat een tentamen met meer dan 30% onvoldoendes wijst op tekortkomingen in de toets of in het onderwijs die niet student-gerelateerd zijn. In de hoofdfase is een slagingspercentage van 90% normaal.

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.

| **P-waarde propedeuse** | **P-waarde postpropedeuse** | **Uitleg** |
|---|---|---|
| < 0,65 | p < 0,85 | toets te moeilijk? evt. cesuur aanpassen |
| 0,65 < p < 0,75 | 0,85 < p < 0,95 | normaal, wijst niet op afwijkingen |
| p > 0,75 | p > 0,95 | toets (te) makkelijk? evt. cesuur aanpassen |

# 3 Vraagkwaliteit

Als je een digitaal tentamen afneemt wordt er informatie over de kwaliteit van toetsvragen (items) verzameld. Hieronder wordt ingegaan op de p-waarde, a-waarde en de $R_{it}$- en $R_{ir}$-waarde.

Let op! Dergelijke statistische maten zijn vooral betekenisvol bij voldoende grote aantallen studenten (>100). Vanaf 25 afnames krijg je echter al wel een indruk van de kwaliteit van de vragen.

**P-waarde**

Bij het beoordelen van de kwaliteit van een item bekijk je eerst de p-waarde: de proportie correcte antwoorden op een individueel item. De p-waarde varieert van 0 (iedereen fout) en 1 (iedereen goed).

Niet alle vragen in een toets zijn even makkelijk of moeilijk, en dat is niet erg. Zo maak je een onderscheid tussen sterke, zwakke en gemiddelde studenten. In combinatie met de $R_{it}$-waarde (zie hieronder) kun je bijvoorbeeld zien of een moeilijke vraag inderdaad ook goed is beantwoord door sterk scorende studenten.

***Normen voor p-waarden bij summatieve toetsen***

Bij meerkeuzevragen moet de p-waarde hoger zijn dan de gokkans, anders geeft de vraag geen zinvolle informatie over de kennis van de studenten. De optimale p-waarde ligt in het midden tussen de maximale p-waarde (1,0) en de gokkans uitgedrukt als decimaal.

| | **Gokkans** | **Optimale p-waarde** | **Ondergrens** | **Bovengrens** |
|---|---|---|---|---|
| *bij tweekeuzevragen* | 50% (0,5) | 0,75 | 0,61 | 0,90 |
| *bij driekeuzevragen* | 33% (0,33) | 0,67 | 0,50 | 0,90 |
| *bij vierkeuzevragen* | 25% (0,25) | 0,62 | 0,44 | 0,90 |
| *bij open vragen* | | 0,50 | 0,25 | 0,90 |

Ontleend aan: Berkel van, H., Bax, A. & Joosten-ten Brinke, D. (2017). *Toetsen in het hoger onderwijs*. Houten: Bohn Stafleu van Loghum.

**A-waarde**

Bij meerkeuzevragen kun je ook kijken naar hoe de foute antwoorden (de ‘afleiders’) scoorden. 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, of geen realistische antwoordoptie voor studenten. Als de vraag wordt hergebruikt, is het beter om die afleider aan te passen of niet meer te gebruiken.

**$R_{it}$ en $R_{ir}$-waarde**

De $R_{it}$-waarde geeft het onderscheidend vermogen weer van een item en staat voor de correlatie tussen het item en de totaalscore op de toets.

Een vraag heeft een hoog onderscheidend vermogen ($R_{it}$ 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 ($R_{it}$ is 0).

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

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 $R_{ir}$-waarde geeft daarom 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. Voor toetsen met meer dan 25 vragen is het verschil tussen de $R_{it}$- en de $R_{ir}$-waarde verwaarloosbaar.

***Normen voor de $R_{it}$-waarde***

| **Waarde** | **Kwalificatie** |
|---|---|
| 0,35 en hoger | zeer goed |
| 0,25-0,35 | goed |
| 0,15-0,25 | voldoende |
| minder dan 0,15 (ook negatieve waarden mogelijk) | onvoldoende |

Ontleend aan: Berkel van, H., Bax, A. & Joosten-ten Brinke, D. (2017). *Toetsen in het hoger onderwijs.* Houten: Bohn Stafleu van Loghum.

**Reparatiemogelijkheden op basis van toetsanalyse**

Als uit de psychometrische analyse blijkt dat sommige waarden niet optimaal zijn, dan zijn er beperkte mogelijkheden om de kwaliteit van de toets alsnog te verbeteren. Hieronder een samenvatting van de mogelijkheden. Let op! Vragen uit de toets verwijderen is vaak een optie voor verbetering, maar zorg er voor dat de inhoud van de toets zoveel als mogelijk representatief blijft. Daarom is het belangrijk om meerdere vragen te stellen over elk onderwerp.

| **Probleem** | **Reparatie** |
|---|---|
| p-waarde is lager of gelijk aan raadkans | Is het antwoord correct gesleuteld?<br>- zo nee, sleutel wijzigen<br>- zo ja, vraag uit toets verwijderen |
| p-waarde is hoger dan de raadkans, maar beduidend lager dan de gewenste waarde | - indien $R_{it}$ positief is, vraag handhaven<br>- indien $R_{it}$ negatief is, vraag verwijderen |
| p-waarde is (bijna) 1 | vraag handhaven |
| $R_{it}$ is negatief of 0 | Is het antwoord correct gesleuteld? |

| | |
|---|---|
| | - zo nee, sleutel wijzigen<br>- zo ja, vraag uit toets verwijderen |
| $R_{it}$ is positief, maar lager dan 0,15 | Indien de p-waarde ook laag is, vraag uit toets verwijderen |
| Alfa is lager dan 0,6 | Te veel onjuiste zak/slaagbeslissingen. Toets als formatief beschouwen |
| Alfa is hoger dan 0,6 maar lager dan 0,8 | - Indien de toets kan worden gecompenseerd met andere toetsen is dit nog wel acceptabel, maar is het dringend noodzakelijk naar de kwaliteit van de vragen te kijken en aanpassingen te doen.<br>- In andere gevallen de vragen met een lage of negatieve $R_{it}$ uit de toets verwijderen, en de analyse opnieuw uitvoeren |

Ontleend aan: Berkel van, H., Bax, A. & Joosten-ten Brinke, D. (2017). *Toetsen in het hoger onderwijs*. Houten: Bohn Stafleu van Loghum.

**Heb je vragen over dit document, of wil je advies over jouw analyseresultaten? Neem contact op met de toetsdeskundigen van de faculteit: goedtoetsen-fgw@uva.nl.**