Richtlijnen voor toetsanalyse en cesuurbepaling bij bloktoetsen

interpretatie itemindices & betrouwbaarheid & cesuurmethoden

Randvoorwaarden interpretatie item-indices:

Bij een relatief kleine groep toetskandidaten (N < 100) en een relatief klein aantal items (< 80 vragen) moeten de waarden op de item-indices met de nodige *voorzichtigheid geïnterpreteerd* worden!! De betrouwbaarheid van de waarden is dan lager.

De item-indices zijn zowel voor toetsen met gesloten vragen als toetsen met open vragen toepasbaar; informatie over p' en f-waarde en z-waarde is alleen van toepassing op meerkeuzevragen.

p-waarde

p-waarde = moeilijkheidsgraad (proportie studenten die het item goed heeft)

p'-waarde

p'-waarde = voor aantal alternatieven gecorrigeerde moeilijkheidsgraad (proportie studenten die het item kent).

formule $p' = p - \frac{p_f}{a-1}$

p_f = *proportie studenten die een fout antwoord geven*
a = aantal antwoordmogelijkheden

(Bij gedwongen raden –d.w.z. geen vragen openlaten en fout antwoord=0 punten i.p.v. puntenaftrek- is p_f *gelijk aan 1-p.)*

theoretisch is de ideale p'-waarde: 0.50

--> hoogste selectieve bijdrage aan de rangorde van toetsscores van de toetskandidaten

extreme p'-waarden: p' < 0.10 en p' > 0.90

--> respectievelijk zeer moeilijke of zeer makkelijke vraag

--> nauwelijks selectieve bijdrage aan de rangorde van toetsscores

Rir-waarde

Rir-waarde (item-rest-correlatie): samenhang (correlatie) tussen itemscores en totaalscores met weglating van het item zelf

→ onderscheidend vermogen van een item: scoren de 'goede' studenten goed op het item en de 'slechte' studenten slecht?

streefwaarde: Rir > 0.10

extreme waarden: Rir < 0.10

→ item discrimineert niet

f-waarde

f-waarde = aantal studenten dat het betreffende alternatief heeft ingevuld dan wel de vraag open laat

→ streven is dat de afleiders (de onjuiste alternatieven) gelijkwaardige opties zijn en dat de frequentie van het juiste alternatief groter is dan de frequentie van de afzonderlijke afleiders en het aantal open.

z-waarde

z-waarde = gestandaardiseerde gemiddelde toetsscore van de studenten die voor een bepaald antwoordalternatief kiezen bij meerkeuzevragen.

→ deze index geeft informatie over het onderscheidend vermogen van een alternatief (voorbeeld: z-waarde van alternatief 3 = +0.50 → gemiddelde toetsscore van groep die alternatief 3 kiest wijkt een halve standaarddeviatie (positief) af van de gemiddelde toetsscore exclusief die vraag) → streven is dat de z-waarde van de sleutel (juiste alternatief) positief is en groter is dan z-waarden van de afleiders (onjuiste alternatieven) → de goede studenten kiezen voor het juiste alternatief, terwijl de slechtere studenten voor een afleider kiezen.

Tabel voor de interpretatie van de combinatie van p'-waarde en Rir-waarde:

	Rir lager dan 0.1	***Rir hoger dan 0.1***
P' lager dan 0.1	• Sleutel correct? • Detailvraag? • Vraagformulering eenduidig? • Ander alternatief ook plausibel? • Niet aan orde in onderwijs?	• Instinkertje? • Te moeilijk/complex? • Niet aan orde in onderwijs?
P' tussen 0.1 en 0.8	• Sleutelfout / ander alternatief ook plausibel?	• In orde • Mogelijk aandacht niet-functionerende afleiders (toekomst)
P' hoger dan 0.9	• Weggever (op te lossen met boerenverstand)? • Te gemakkelijke vraag / alternatieven geen goede afleiders / zeer intelligente groep studenten / onderwijs effectief	• Te gemakkelijke vraag / geen goede afleiders / zeer intelligente groep studenten / onderwijs effectief

Een item dient áltijd beoordeeld te worden aan de hand van: de **p'** en de **Rir** *(een p' kan bijvoorbeeld een ideale waarde hebben, maar een lage Rir; dan is er iets vreemds aan de hand met dat item)* **EN** aan de hand van het **studentcommentaar** en **door de vraag inhoudelijk en/of redactioneel te bekijken.**

Opmerkingen bij laten vervallen van items:

- met name items die in het *grijze gebied* van bovenstaande tabel vallen
- wanneer twee alternatieven beide juist blijken te zijn: dan is de suggestie om dit item te laten vervallen
- *let op!* het laten vervallen van items kan de (inhouds-)validiteit (=*representativiteit v/d items ten opzichte van de onderwijsstof naar vakinhoud en vraagniveau*) van de toets aantasten
- (mogelijke) voordeel: hogere betrouwbaarheid

Advies: bekijk de waarden op items en alternatieven ook ten behoeve van toekomstig gebruik van het item (in aangepaste vorm) c.q. de alternatieven. Wanneer bijvoorbeeld blijkt dat bij een 4-keuze item één van de alternatieven niet goed afleidt, dan kan dit item wellicht een volgende keer als 3-keuze item worden opgenomen. Of vervang een slecht functionerend item als geheel door een andere vraag of pas de redactie van de vraag aan.

Coëfficiënt alpha (α) – betrouwbaarheidsschatting toets

Hoe betrouwbaarder de toets des te nauwkeuriger de scores geïnterpreteerd kunnen worden. De alpha geeft de onderlinge samenhang van de items weer, de interne consistentie. Het geeft vooral aan in hoeverre van item tot item hetzelfde wordt gemeten. Voor de theoretische betrouwbaarheidscoëfficiënt geldt $0 \leq$ alpha ≤ 1. Voor een toets wordt een α nagestreefd van 0.80, om in redelijke mate een uitspraak te kunnen doen over het kennisniveau van de student. Regel is dat hoe langer de toets, des te beter de differentiatiegraad, en des te hoger de interne consistentie c.q. betrouwbaarheid.
Bij een herkansing is de verwachting dat de coëfficiënt alpha wat lager uitvalt omdat de onderlinge verschillen in deze groep geringer zijn dan bij een eerste toetsafname.

Cesuurbepaling – zak-/slaaggrens

Er zijn grofweg drie methoden om de grens tussen onvoldoende en voldoende (cijfer 5,5) te bepalen: relatieve cesuur, absolute cesuur en een combinatie van een absolute cesuur op een relatief referentiepunt.

- relatieve cesuur: cesuur geheel o.b.v. de behaalde scores (b.v. gemiddelde – 1x SD) of percentage geslaagden (vooraf) vastgesteld. Deze methode wordt voor de bloktoetsen binnen ons onderwijsinstituut niet gehanteerd.

- absolute cesuur o.b.v. maximaal mogelijke score: percentage goede antwoorden (kennisniveau) vooraf vastgesteld, bij een bloktoets ligt dit percentage tussen 50 en 60%.
Bijvoorbeeld: bij een meerkeuzetoets met 80 vragen en een vereist kennispercentage van 50% ligt de cesuur op een toetsscore van 40.
Bij gedwongen raden wordt deze toetsscore veelal nog vermeerderd met de kansscore; bij correctie voor raden is al gecorrigeerd voor de gokkans in de scoring van de afzonderlijke vragen.

- absolute cesuur o.b.v. relatief referentiepunt: **de cesuurmethode van Cohen-Schotanus**. *Deze is gebaseerd op het principe dat een absolute cesuur van b.v. 60% kennisniveau wordt toegepast ten opzichte van een relatief referentiepunt, namelijk de score van beste student of het 95e-percentiel. De methode van Cohen-Schotanus houdt dus rekening met de moeilijkheidsgraad van de toets en de ruis aan de onderwijskant.*
Gebruikelijk is om bij scoring met correctie voor raden de cesuur te berekenen door 60% van de 95^{e} percentielscore te nemen. Bijvoorbeeld: bij een meerkeuzetoets van 80 vragen en een 95^{e} percentielscore van 60 (de ondergrens van de 5% beste scoorders) ligt de cesuurgrens bij een toetsscore van (60% van 60 =) 36.

NB De cesuurmethode Cohen-Schotanus is niet toepasbaar bij herhaaltoetsen!
De beste scoorder(s) van de groep herhaalstudenten is niet representatief om te corrigeren voor de ruis aan de onderwijskant. Ook voor kleine groepen studenten (< ± 70) is deze methode ongeschikt.

→ Neem contact op met toetsservice (via Onderwijscoördinatie) voor advies bij toets-/itemanalyse, cesuurbepaling en score→cijfertransformatie.