

Klopt dit wel?



Als de toets een lage alfa heeft, dan is de toets slecht

■ Silvester Draaijer

Silvester Draaijer is onderwijsadviseur en adviseur ICT bij de Vrije Universiteit Amsterdam.

Email: s.draaijer@vu.nl

■ Jackelien ter Burg

Jackelien ter Burg is werkzaam als freelance toetsdeskundige en redactielid van het tijdschrift Examens.

E-mail jackelien@toetsrealisatie.nl

In het kader van de toetsanalyse bij een toets met gesloten vragen wordt vaak Cronbach's alfa berekend om een uitspraak te doen over de betrouwbaarheid van de toets. Daarbij wordt soms zondermeer aangenomen dat Cronbach's alfa hoger moet zijn dan een bepaalde waarde om van een betrouwbare toets te mogen spreken. Dat is echter niet zo eenduidig aan te geven als het lijkt. Hoe zit dat?

Wat is Cronbach's alfa?

Cronbach's alfa - ook wel coëfficiënt alpha genoemd (verder: alfa) - is een van de mogelijke uitkomsten van een statistische toetsanalyse.

Veel toetsservicesystemen bevatten een mogelijkheid om deze coëfficiënt te berekenen. Alfa

wordt berekend op basis van het aantal vragen in de toets, de gemiddelde covariantie tussen de scores op de toetsvragen en de variantie van de totale score op de toets. Je kunt alfa gebruiken als een maat voor de meetnauwkeurigheid van de toets bij een enkele toetsafname. Alfa wordt weergegeven als een getal tussen de 0 en 1 en in het algemeen geldt: hoe hoger alfa is, hoe minder ruis, hoe beter.

In de literatuur worden voor alfa verschillende normwaarden genoemd, zie bijvoorbeeld de tabel.

Stel nu dat de toetsanalyse van een bepaalde toets een alfa geeft van 0.5, is de toets dan slecht? Het antwoord daarop is niet zondermeer 'ja'!

Waarom klopt dit niet?

Eerst moet je je afvragen of alfa wel een juiste

Tabel. Normen voor alfa van een summatieve toets (Bron: Berkel, 2014)

Gevonden waarden	Kwalificatie
0,90 en hoger	goed/zeer goed
0,80 tot 0,90	voldoende/goed
0,70 en 0,80	middelmatig/voldoende
Minder dan 0,70	slecht/middelmatig

maat is om de betrouwbaarheid van de toets aan te tonen. Een lage alfa kan namelijk verschillende (samenhangende) oorzaken hebben. Eén van de belangrijkste eigenschappen van alfa is dat de waarde hoger wordt, naarmate er meer spreiding in de toetsscores is. Hoe lager de spreiding, hoe lager alfa. Een lage spreiding treedt bijvoorbeeld op als de kandidaten allemaal ongeveer even vaardig zijn. De toets maakt dan geen onderscheid tussen de kandidaten. Dit kan aan de orde zijn bij een toets waarbij de meeste vragen incorrect beantwoord worden: de gehele groep scoort laag. Een oorzaak hiervan kan zijn dat de toets op zich past bij de inhoud van het vak of bij het construct dat je wilt meten, maar mogelijk paste het onderwijs niet bij de doelgroep of was het onderwijs zelf onder de maat. De lage waarde van alfa zegt dan niets over de betrouwbaarheid van de toets maar iets over de meting die met die toets is gedaan. Een voorbeeld hiervan zijn herkansingstoetsen. Zij hebben vaak een lagere alfa omdat de groep kandidaten die de herkansing maakt minder spreiding in behaalde scores laat zien dan de groep bij de eerste kans. Immers, de herkansing wordt meestal gemaakt door een groep bestaande uit de minder vaardige kandidaten.

De toets maakt ook geen onderscheid als de kandidaten de meeste toetsvragen correct beantwoorden zoals het geval kan zijn bij beheersingstoetsen (mastery tests) die onder andere in de zorgsector gebruikt worden. Bij een beheersingstoets moet een kandidaat laten zien dat hij/zij de stof nog voldoende beheerst. Ook bij deze toetsen is de spreiding laag en zal de berekende waarde van alfa om die reden laag zijn. Als je in deze gevallen de meetnauwkeurigheid

wilt bepalen rond de zak- en slaaggrens dan kan je beter beslissingsprecisiematen nemen zoals Livingston's k_2 of Sc gebruiken. Deze maten gaan uit van de proportie misclassificaties voor zakken en slagen op een toets rondom de zak- en slaaggrens in plaats van de gemiddelde score. Een andere belangrijke eigenschap van alfa is dat deze waarde hoger is naarmate er meer vragen in de toets zitten en die vragen goed onderscheid maken in de mate waarin de kandidaten de stof beheersen. Een toets die uit weinig vragen bestaat, resulteert daarom als vuistregel niet in een hoge alfa. De toets kan eenvoudig verbeterd worden door meer toetsvragen in de toets op te nemen.

Tot slot is het zo dat alfa een onderschatter van de betrouwbaarheid is. Minder bekende betrouwbaarheidsmaten die een nauwkeuriger betrouwbaarheidsschatting geven, en dus minder onderschatten, zijn bijvoorbeeld Guttman's λ_2 of Revelle's β . Je kunt ook overwegen om Item Response Theorie (IRT) te gebruiken om nauwkeurigheid van de meting te optimaliseren. In combinatie met adaptieve toetsing kan met een minimaal aantal toetsvragen, bij een gewenst niveau van nauwkeurigheid, bepaald worden welke vaardigheid een kandidaat heeft. Toepassing van IRT vergt echter gespecialiseerde kennis, software en uitgebreide mogelijkheid tot het verzamelen en analyseren van toetsdata voorafgaand aan de echte toetsing.

Conclusie

Ondanks een lage alfa hoeft een toets op zich geen onbetrouwbaar meetinstrument te zijn. Ten eerste moet nagegaan worden of alfa eigenlijk

wel een geschikte maat is, gegeven de aard van de toets. Een lage alfa is een indicatie dat je nader onderzoek moet doen naar de toets zelf, maar ook naar de groep kandidaten die de toets maakt, het gegeven onderwijs, of de toets een eerste-kans of herkansing betreft en mogelijk zelfs andere omstandigheden. Je moet ingeval van een lage alfa natuurlijk wel op een ander manier kunnen aantonen dat de toets, gegeven de situatie waarin de toets gebruikt wordt, voldoende nauwkeurig meet. ■

Meer lezen?

- Berkel, H.J.M. van, Bax, A.E. & Brinke, Joosten-ten Brinke, D. (2014). Toetsen in het hoger onderwijs. Houten: Bohn Stafleu van Loghum.
- Sijsma, K. (2009). Over misverstanden rond

Cronbachs alfa en de wenselijkheid van alternatieven. <http://foswiki.cs.uu.nl/foswiki/pub/Toetsing/ToetsAdviesCommissie/MisverstandenRondCronbachAlpha.pdf>

- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijsma. *Psychometrika*, 74(1), 145–154.
- Gruijter, D. N. M. de (2008). Toetsing en toetsanalyse <http://media.leidenuniv.nl/legacy/toetsing-en-toetsanalyse.pdf>
- Coscarelli, W., & Shrock, S. (2002). The two most useful approaches to estimating criterion-referenced test reliability in a single test administration. *Performance Improvement Quarterly*, 15(4), 74–85. <http://doi.org/10.1111/j.1937-8327.2002.tb00266.x>

Gesignaleerd

De lichtung van 2016: de examens zijn afgelopen.

Wat zou het einde van de eindexamens zijn zonder examenreis?

Weken hebben de leerlingen van de Lichting van 2016 met hun neus in de boeken gezeten, nu mogen ze eindelijk het vliegtuig in stappen richting de zon. Die examenuitslag kan nog wel even wachten.

Jelmer, Rafael en Chen reizen alle drie (maar niet samen) af naar Lloret de Mar. Friezin Chen doet dat met een vriendin, de twee heren met hun vriendengroepen. Wat de voor-malige eindexamenkandidaten daar gaan doen, laat zich raden. Niets laat de examenstress sneller verdwijnen dan een flinke dosis zon, strand en dance-muziek.

(de Volkskrant 27 mei 2016)