openjournals.nl

Tijdschrift voor Hoger Onderwijs
https://www.tvho.nl/ | E-ISSN 2468-9483

Gangbare vuistregels voor indicatoren van toetskwaliteit zijn te strikt

Niels Smits en Sharon Klinkenberg

Samenvatting In dit artikel wordt betoogd dat hoewel het gebruik van statistische indicatoren van de kwaliteit van studietoetsen in het Hoger Onderwijs raadzaam is, de daarbij gehanteerde vuistregels in veel gevallen te strikt zijn. Ze zijn oorspronkelijk vastgelegd voor de selectie van vragen voor een gestandaardiseerd instrument in een normgeoriënteerde toetsituatie waarbij er sprake is van veel verschil in vaardigheid tussen studenten, vragen die alle gemiddeld moeilijk zijn en afnames bij grote groepen studenten, waardoor ze niet algemeen bruikbaar zijn. Er worden vier vragen geïntroduceerd die toetsconstructeurs kunnen stellen om te bepalen hoeveel gewicht er aan deze vuistregels gegeven dient te worden: (i) Wat is de verklaring? (ii) Wat is het toetsdoel? (iii) Wat is het type instrument? en (iv) Hoe zit het met de spreiding? Er wordt afgesloten met suggesties voor het bijstellen van de vuistregels.

Trefwoorden studietoetsen, toetsconstructie, toetskwaliteit, betrouwbaarheid

Artikelgeschiedenis
Ontvangen: 17 januari 2025
Geaccepteerd: 15 december 2025
Online: 6 maart 2026

Contactpersoon
Niels Smits, n.smits@uva.nl

Over de auteur(s)
Niels Smits is als universitair hoofddocent werkzaam bij het Research Institute of Child Development and Education, Faculteit der Maatschappij- en Gedragswetenschappen, Universiteit van Amsterdam; Sharon Klinkenberg is als docent werkzaam bij Communicatiewetenschap en is co-directeur van het Teaching & Learning Center van de Faculteit der Maatschappij- en Gedragswetenschappen, Universiteit van Amsterdam.

Inleiding

Toetsen spelen een centrale rol in het hoger onderwijs. Ze worden niet alleen gebruikt om studenten te beoordelen op hun voortgang, maar ook om hun leergedrag te beïnvloeden (Biggs & Tang, 2011). Er is een toename in de belangstelling voor de *kwaliteit* van toetsing die mogelijk het gevolg is van verschillende maatschappelijke ontwikkelingen, zoals de toegenomen behoefte aan meten (Van der Kolk, 2021), het groeiende (toegekend) belang van een diploma (Elffers, 2020) en de juridisering van de samenleving (Van Berkel, 2017).

Deze veranderingen komen op Nederlandse universiteiten en hoge scholen tot uitdrukking in centraal geformuleerd toetsbeleid (Martens & Moerkerke, 2017) en, gezien hun wettelijke verantwoordelijkheid voor de borging van kwaliteit van toetsing en examens, in een toename in het werk van examencommissies (Schaap, 2023). Daar waar het *beleid* omtrent toetsing het werk is van het bestuur van een opleiding, blijft de *constructie* van toetsen doorgaans de verantwoordelijkheid van de vakdocent. Docenten leren tijdens hun opleiding Basiskwalificatie Onderwijs basale concepten als validiteit,

Citeren Smits, N. en Klinkenberg, S. (2026). Gangbare vuistregels voor indicatoren van toetskwaliteit zijn te strikt, *Tijdschrift voor Hoger Onderwijs 44*, 1–16, DOI: 10.59532/hsnnhe76

betrouwbaarheid, toetsmatrijs en normering (Maassen et al., 2015; Van Berkel, 2011). Echter, de meeste docenten hebben slechts in beperkte mate kennis van de testleer die zich bezighoudt met het *modelleren* van testgegevens (Van den Brink & Mellenbergh, 1998), en statistische indicatoren heeft voortgebracht waarmee de kwaliteit van toetsen kan worden bepaald. Wel zijn er vuistregels beschikbaar (bijv. Rudolph et al., 2019; Van Berkel & Bax, 2023) die docenten kunnen gebruiken bij het analyseren van hun toets. Hoewel deze nuttig kunnen zijn zien de auteurs vanuit hun werkzaamheden in de toets/examencommissie een aantal problemen in de *verwoording* van deze vuistregels. Naar hun mening leiden ze te snel tot het verwijderen van toetsvragen die eigenlijk in orde zijn. Dit artikel heeft de volgende structuur. Eerst worden de meest populaire vuistregels besproken en vervolgens wordt een reeks vragen geïntroduceerd die docenten kunnen stellen om te bepalen hoeveel gewicht ze aan de vuistregels dienen te geven bij de evaluatie van hún toets. In de Discussie wordt beschreven hoe de vuistregels beter kunnen worden verwoord.

Indicatoren van toetskwaliteit

In deze bijdrage staan studietoetsen centraal. Met een studietoets wordt hier bedoeld een toetsingsinstrument met een ingebouwde puntentelling waarmee middels een score de mate van het bereiken van eenvoudige leerdoelen voor kennis, inzicht en vaardigheid op een gegeven vakgebied wordt uitgedrukt (c.f. De Groot & Van Naerssen, 1969; Van Berkel & Bax, 2023). Volgens deze definitie valt een schriftelijk tentamen met meerkeuzevragen en/of korte open vragen wel, maar een portfolio of masterscriptie niet onder de noemer studietoets (omdat het in de laatste twee gevallen complexe leerdoelen betreft). Het is gebruikelijk om na de afname van een studietoets de kwaliteit van de gebruikte vragen te evalueren en eventueel reparaties uit te voeren om de eerlijkheid van de toetsuitslag te verbeteren. Daartoe kunnen indicatoren van de kwaliteit van studietoetsen gebruikt worden zoals:

(a) de moeilijkheid. De gemiddelde score op een vraag, in het geval van dichotoom gescoorde antwoorden is dit de proportie correcte antwoorden, vaak aangeduid met 'p-waarde';
(b) het onderscheidingsvermogen van toetsvragen. De mate waarin een vraag meet wat de totale toets meet, gekwantificeerd met de itemtestcorrelatie, r_{it}, of itemrestcorrelatie, r_{ir}; en
(c) de betrouwbaarheid van de toets. De mate waarin de toetsscore vrij is van meetfouten, vaak geschat met Cronbach's α.

Voor elk van deze indicatoren bestaan suggesties omtrent het afkappunt voor minimale vereisten; bijvoorbeeld, Van Naerssen (1969, p. 252) achtte een onderscheidingsvermogen van minimaal 0.20 wenselijk. Vaak worden naast signaalwaarden ook kwalificaties gege-

Tabel 1 Een selectie van de vuistregels voor statistische indicatoren volgens Van Berkel en Bax (2017, 2023)

p-waarde				Onderscheidingsvermogen		Cronbach's alfa	
Soort vraag	onder	ideaal	boven				
open	0.25	0.50	0.90	goed/zeer goed	0.35 en hoger	goed/zeer goed	0.90 en hoger
driekeuze	0.50	0.67	0.90	voldoende/goed	0.25–0.35	voldoende/goed	0.80–0.90
				middelmatig/voldoende	0.15–0.25	middelmatig/voldoende	0.70–0.80
				slecht/middelmatig	minder dan 0.15	slecht/middelmatig	minder dan 0.70

ven om goede, matige en slechte vragen van elkaar te onderscheiden. In toetssoftware worden soortgelijke aanbevelingen verwerkt (zie Ans, 2025) waarbij individuele vragen soms van een kleur worden voorzien (zoals: groen, oranje en rood in *TestVision*, 2024). In Tabel 1 staan de vuistregels zoals door Van Berkel en Bax (2017, 2023) gebruikt en in de onderhavige bijdrage wordt deze tabel centraal gesteld, omdat het staat in het meest populaire boek over toetsen in het Hoger Onderwijs in het Nederlandse taalgebied, en omdat het door de uitgever wordt aangeprezen als een belangrijke bron bij het behalen van de certificaten Basiskwalificatie- en Seniorkwalificatie Examinering. Bovendien lijkt de genoemde toetssoftware zich ook op deze bron te baseren. Van Berkel en Bax geven overigens voor slechts één indicator een bronvermelding. Voor het onderscheidingsvermogen wordt verwezen naar het boek *Essentials of Educational Measurement* van Ebel en Frisbie (1991, niet langer in druk) die zelf als volgt de oorsprong van hun vuistregels beschreven: "Experience with a wide variety of classroom tests suggests that the indices of item discrimination for most of them can be evaluated in these terms" [Ervaring met een brede verscheidenheid aan toetsen suggereert dat de indices voor onderscheidingsvermogen voor de meeste ervan in deze termen kunnen worden beoordeeld] (p. 232). Voor p-waarden wordt weliswaar geen expliciete verwijzing gegeven, maar zowel inhoud als presentatie doen vermoeden dat ook deze vuistregels uit Ebel en Frisbie (1991, pp. 130–131 en p. 223) zijn overgenomen; ook voor de vuistregels omtrent de betrouwbaarheid ontbreken referenties.

Vier vragen voor vuistregels

Statistische vuistregels kunnen veel houvast bieden (van Belle, 2008), maar de voorwaarde is wel dat ze in de juiste context worden toegepast (Gigerenzer et al., 2000). Wanneer ze verkeerd worden gebruikt, kan het leiden tot het ten onrechte diskwalificeren van goede toetsvragen waardoor de betrouwbaarheid en de dekking van het kennisdomein in het geding komen. Het is daarom aan te bevelen dat toetsconstructeurs

 DOI: 10.59532/hsnnhe76

de geschiktheid van deze vuistregels voor hun situatie bepalen alvorens ze besluiten ze toe te passen. We introduceren vier vragen voor toetsconstructeurs die daarbij kunnen helpen:

1. Wat is de verklaring?
2. Wat is het toetsdoel?
3. Wat is het type instrument?
4. Hoe zit het met de spreiding?

Deze vragen zijn geen gangbare, in de literatuur reeds bestaande, set vragen, maar ze volgen direct uit basale principes in de testleer. Hieronder wordt voor elke vraag het belang beschreven. Daar waar de eerste drie vragen betrekking hebben op logische principes, richt de vierde zich op statistische eigenschappen van toetsgegevens die het beste kunnen worden besproken middels een simulatiestudie. De betreffende sectie bevat daarom een eigen Methode-, Resultaten- en Conclusieparagraaf.

Wat is de verklaring?

Hoewel Van Berkel en Bax (2017, 2023) waarschuwen tegen een mogelijk verlies van representativiteit is de algemene teneur dat lage indicatorwaarden toch echt vragen om aanpassingen; Het is van belang dit te nuanceren. Een indicator die een extreme waarde geeft, is op zichzelf namelijk geen bewijs dat er iets aan de hand is (Di Bello, 2021). Voordat de toets wordt aangepast, is er een *verklaring* nodig voor de afwijking, en dat vraagt om het verzamelen van extra informatie. Toetsanalyses zijn een vorm van statistische kwaliteitscontrole, en in die context is het vereist om na de detectie van afwijkingen de *oorzaak* ervan te bepalen (Montgomery, 2019). Er zijn twee mogelijke oorzaken: een toevalsoorzaak en een aanwijsbare oorzaak. In het eerste geval is een extreme waarde inherent aan het gebruik van kwaliteitsindices, en zal deze dus af en toe voorkomen. In het laatste geval is deze het gevolg van een verstoring in het toetsproces. Om te bepalen welke van de twee situaties het betreft dient, zoals Van Naerssen (1969) dat noemde, een 'verbaal-logische analyse' van de vragen te worden uitgevoerd. Het kan blijken dat er een fout in de nakijksleutel stond, meerdere opties goed waren of dat de stam van de vraag dubbelzinnigheden bevatte. Ook kan het bij nader inzien blijken dat de stof in kwestie onvoldoende of verkeerd is behandeld, zodat vaardigere studenten systematisch een fout antwoord gaven. In dat geval is een aanpassing inderdaad noodzakelijk. Als deze analyse echter geen aanwijsbare oorzaak oplevert, dan moet men concluderen dat het om een toevalsoorzaak gaat en het daarbij laten.

Wat is het toetsdoel?

Studietoetsen kunnen op twee manieren gebruikt worden: normgeoriënteerd en criteriumgeoriënteerd (Van den Brink & Mellenbergh, 1998, pp. 19–20). Bij het normgeoriënteerde gebruik is het doel om studenten onderling te vergelijken, en ze te rangordenen van de laagste tot de hoogste prestatie. Bij het criteriumgeoriënteerde gebruik is het doel

een vergelijking van de prestatie van een student met een onderwijskundige doelstelling (bijvoorbeeld tenminste 55% kennis van de studiestof), en zijn de scores van andere studenten niet van belang. De vuistregels uit Tabel 1 zijn volgens de betreffende auteurs (Van Berkel & Bax, 2017, 2023) gebaseerd op Ebel en Frisbie (1991, H. 13). Inspectie van deze bron laat zien dat ze *expliciet* voor normgeoriënteerde situaties geformuleerd zijn. Bij normgeoriënteerd toetsen zijn niet extreme p-waarden en hoge waarden van r_{it}, r_{ir} of α wenselijk omdat ze bijdragen aan het in kaart brengen van verschillen tussen studenten. Bij criteriumgeoriënteerd toetsen ligt de nadruk op de precisie van zak/slaagbeslissingen, en zijn ze minder belangrijk. Anderssoortige indicatoren zijn veel belangrijker (Crocker & Algina, 1986, H. 9; Standards, 2014, Standard 2.16). Bijvoorbeeld, bij een inhoudelijk hoogwaardige toets waarvoor een lichting studenten met hoge cijfers slaagt, zullen door de slechts kleine onderlinge verschillen het onderscheidend vermogen en de interne consistentie te laag zijn en de vragen te makkelijk. Maar de docent mag tevreden zijn, omdat voor iedereen het betrouwbaarheidsinterval voor de ware score boven de cesuurscore lag, en er dus was voldaan aan de leerdoelen. Het is overigens niet gezegd dat de indicatoren nooit bruikbaar zijn bij criteriumgeoriënteerd toetsen. Indien er voldoende natuurlijke spreiding is in studentvaardigheid (zie ook de vierde vraag), bieden ze mogelijk relevante informatie, maar dan moeten er minder strikte grenswaarden worden gehanteerd dan die in de tabel worden genoemd (Ebel & Frisbie, 1991, p. 224).

In het hoger onderwijs hangt het van de docent en toetssituatie af welk doel wordt beoogd. Het is aannemelijk dat docenten vaak beide doelen nastreven: zowel een precieze scheiding van goede en slechte studenten als een deugdelijke ordening in vaardigheden. Als de nadruk enkel ligt op normgeoriënteerd toetsen, dan zijn de suggesties uit Tabel 1 op zijn plaats. Maar als men criteriumgeoriënteerd wil toetsen of als men beide doelen nastreeft, hoeft men minder gewicht toe te kennen aan de vuistregels (en is de inspectie van anderssoortige indicatoren evenwel nodig).

Wat is het type instrument?

Studietoetsen kunnen grofweg worden ingedeeld in gestandaardiseerde (of 'geijkte') en ongestandaardiseerde instrumenten (De Groot & Van Naerssen, 1969, H. 11; Ebel & Frisbie, 1991, H. 16). In de VS hebben drie wetenschappelijke verenigingen (AERA, APA en NCME, 2014) samen maatstaven ('*Standards*') opgesteld waaraan gestandaardiseerde onderwijskundige en psychologische meetinstrumenten moeten voldoen. Een belangrijke eis is dat alvorens vragen in de toetspraktijk worden ingezet, hun geschiktheid wordt verzekerd door ze uit te proberen in de beoogde studentpopulatie. In de toetspraktijk van het hoger onderwijs kan hier echter niet aan worden voldaan (zie de '*Standards*', p. 183). Het uitproberen zou ertoe leiden dat vragen bekend ('gelekt') raken, wat de validiteit van de toets compromitteert (Molkenboer & Kempers, 2017). Men kan overwegen vragen uit voorgaande jaren te hergebruiken, en de indicatoren aan te wenden om een goede toets samen te stellen. Echter, examencommissies eisen doorgaans dat toetsen slechts voor een klein deel uit oude vragen bestaan (Van Loosbroek, 2019), en de kwaliteit van de

 DOI: 10.59532/hsnnhe76

noodzakelijk nieuwe vragen blijft dan evengoed ongewis. Bovendien veranderen vakken voortdurend qua opzet en inhoud, en zullen er voor de meeste toetsen nieuwe vragen nodig zijn. Als we opnieuw naar de bron van de vuistregels uit Tabel 1 kijken dan blijkt dat deze niet zijn opgesteld voor aanpassingen na toetsafnames, maar voor de *selectie* van vragen voor een verbeterde versie van een toets (Ebel & Frisbie, 1991, p. 232).

Dus alleen in het geval docenten de beschikking hebben over een geijkte vraagcollectie kunnen ze Tabel 1 als leidraad gebruiken om een toets samen te stellen. Bij afnames met (gedeeltelijk) nieuwe vragen betreft het geen toetsconstructie, maar slechts kwaliteitscontrole, waarbij kwalificaties als 'goed' of 'zeer goed' overbodig zijn. In dat geval hoeft men de tabel alleen te gebruiken om te bepalen of vragen aan minimale vereisten voldoen.

Hoe zit het met de spreiding?

De vuistregels uit Tabel 1 houden weinig rekening met factoren waarvan in de testleer bekend is dat ze van invloed zijn. Deze factoren hebben alle te maken met spreiding: de mate waarin eigenschappen variëren, doorgaans gekwantificeerd met standaardafwijking en variantie. Om inzicht te geven in de rol die spreiding kan spelen bij de grootte van indicatoren werd een kleine simulatiestudie uitgevoerd. In simulatiestudies worden gefingeerde data gegenereerd met een structuur die onder controle staat van de onderzoeker, waardoor de impact van specifieke factoren op relevante uitkomsten op eenduidige wijze kunnen worden geïllustreerd. We gaan bij de methode, de resultaten en de conclusies vooral in op de relevantie van de simulatiestudie voor toetsconstructeurs.

Methode

Er wordt achtereenvolgens ingegaan op het gebruikte testtheoretische model, de gebruikte vijf factoren, de centraal gestelde vier uitkomsten, en de analyse.

Itemresponstheorie en model

Er werd voor het genereren van toetsdata gebruik gemaakt van itemresponstheorie (IRT), een onderdeel van wat vaak moderne testtheorie wordt genoemd. IRT-modellen worden bij instituten als het Cito gebruikt om examengegevens te analyseren. De in het artikel centraal staande indicatoren zijn juist afkomstig uit de klassieke testtheorie. De reden waarom die in de toetspraktijk nog altijd veel worden gebruikt is dat ze (a) voor gebruikers eenvoudiger te begrijpen zijn, en belangrijker (b) deze veel minder grote groepen studenten nodig hebben om ze precies te schatten dan de moderne. Hoewel moderne testtheorie in de onderwijspraktijk dus vaak onbruikbaar is, bieden IRT-modellen bij simulatiestudies als deze wel de voorkeur, omdat ze een fijnmazigere controle over te genereren toetsgegevens bieden.

 DOI: 10.59532/hsnnhe76

Bij het genereren van de toetsgegeven stond het volgende IRT-model centraal, waarbij de kans (P) dat de vraag goed wordt beantwoord ($X = 1$) wordt gegeven door:

$$P(X = 1) = \frac{\exp[\alpha(\theta - \beta)]}{1 + \exp[\alpha(\theta - \beta)]}.$$

In deze logistische functie is θ de vaardigheid van de student, en zijn α en β, respectievelijk, de discriminatie- en moeilijkheidsparameter. (Om het logistische model in overeenstemming te brengen met de gebruikelijke interpretatie van de cumulatieve standaardnormale verdeling werd de logit met een constante van 1.702 vermenigvuldigd.) De moeilijkheid geeft de locatie op de vaardigheidsschaal waar de kans op een goed antwoord fifty-fifty is; als β hoger (lager) is dan θ dan is de kans op een goed antwoord kleiner (groter) dan een half. Het discriminatievermogen geeft de mate aan waarin een vraag verschillen in vaardigheid tussen personen kan onderscheiden; hoe hoger α hoe beter het item. Dit model kan in het geval van meerkeuzevragen, waarbij men het juiste antwoord kan gokken, nog worden uitgebreid met een raadparameter. Deze verfijning wordt omwille van de leesbaarheid buiten beschouwing gelaten. Voor een uitgebreidere beschrijving van IRT-modellen, zie bijvoorbeeld Eggen & Sanders (1993).

Vijf factoren

Er werden vijf factoren in de simulatiestudie opgenomen: aantal studenten, soort vragen, toetslengte, spreiding van de vraagmoeilijkheid in de toets en spreiding van de vaardigheid van studenten. Een toelichting op deze vijf factoren:

Ten eerste werden er twee aantallen studenten (N) gebruikt: 50 en 200; het eerste aantal is gekozen omdat het doorgaans de praktische bovengrens vormt voor het gebruik van open vragen (Van Berkel & Bax, 2023, p. 195) en de tweede waarde representeert in de onderwijspraktijk vaak een grote groep.

Ten tweede werden er twee *soorten vragen* (Vraagtype) gebruikt: open vragen en driekeuzevragen, omdat dit de meest gebruikte vraagvormen zijn; de meerkeuzevragen hadden drie opties omdat dit vaak als optimaal aantal wordt aanbevolen (bijv. Van Berkel & Bax, 2023, p. 197). Bij de conditie met driekeuzevragen werd het IRT-model uitgebreid met een raadparameter waardoor de gokkans steeds een derde was.

Ten derde werden er twee *toetslengtes* (K) gebruikt: 30 en 60 vragen. De waarde van 60 wordt door Van Berkel & Bax (2023, p. 200) als minimum voor een acceptabele betrouwbaarheid bij een toets met driekeuzevragen gesuggereerd, en de waarde van 30 komt in de buurt van het aantal dat in de praktijk vaak wordt gebruikt.

Ten vierde werden er twee soorten verdelingen van itemmoeilijkheid (Verdeling β) gebruikt: een spitse en een platte. In beide condities was de moeilijkheid gemiddeld gelijk aan −0.75 (vergelijk Thompson, 2009), een waarde waarvoor het in een standaard normaalverdeelde populatie betekent dat 77 % van de studenten een vaardigheid heeft die erboven valt (en waarvoor de kans op een goed antwoord voor de meesten groter is dan de kans op een fout antwoord). In de conditie met een spitse verdeling hadden

alle vragen in de toets een moeilijkheidsparameter van –0.75. In de conditie met een platte verdeling kwamen de itemmoeilijkheden uit een interval en werd dit interval, afhankelijke van het aantal items, in 30 of 60 oplopende waarden met gelijke afstand verdeeld. Het interval liep van –2.75 tot 1.25 waardoor de gemiddelde moeilijkheid –0.75 was en de asymmetrie ervoor zorgde dat vragen niet extreem moeilijk waren (vergelijk Woods, 2008). Het discriminatievermogen werd constant gehouden: Alle items hadden een discriminatieparameter van 1.

Ten vijfde werden er twee studentpopulaties gebruikt: een homogene populatie en een heterogene populatie. In beide condities werden studentvaardigheden getrokken uit een normale verdeling met een gemiddelde van 0, maar werden er verschillende standaarddeviaties (SD's) gebruikt: bij de heterogene populatie was de SD gelijk aan 1 en bij de homogene populatie aan 0.5.

Merk dus op dat alle toetsen in werkelijkheid vragen bevatten die even onderscheidend waren en (gemiddeld) even moeilijk, en dat de centrale vraag was hoe de indicatoren van toetskwaliteit variëren als functie van de vijf factoren.

In totaal waren er in de simulatie $2 \times 2 \times 2 \times 2 \times 2 = 32$ combinaties van factoren en werden er in elke cel 100 databestanden gegenereerd.

Vier uitkomsten

Er stonden vier uitkomsten centraal:

1. Het percentage vragen binnen een toets waarvoor de in Tabel 1 onder 'p-waarde' genoemde ondergrenzen werden overschreden; voor driekeuzevragen was de ondergrens 0.50 en voor open vragen 0.25.
2. Het percentage vragen binnen een toets met een r_{it} waarde tussen 0 en 0.15, de waarde die Tabel 1 als ondergrens van het onderscheidingsvermogen geeft.
3. Het percentage vragen dat in een toets een negatieve r_{it}-waarde had; een negatieve waarde ligt weliswaar ook beneden de ondergrens, maar omdat de interpretatie ervan ('vaardigere studenten scoren juist slechter') anders is, zijn de resultaten opgesplitst.
4. De geschatte Cronbach's α van de itemscores.

De eenheid van analyse was de toets (i.e., databestand) en voor alle vier de uitkomsten werd binnen elke cel over de 100 replicaties gemiddeld. De simulaties werden in R uitgevoerd met de *mirt* (Chalmers, 2012) en *SimDesign* (Sigal & Chalmers, 2016) pakketten. Het bestand met R-syntax is beschikbaar bij de eerste auteur.

Analyse

De analyse bestond uit (a) het bekijken van patronen in de gemiddelden, en (b) het vaststellen van de relatieve bijdrage van elke factor. Daartoe werd effectmaat η^2 (proportie verklaarde variantie) geschat. Als ondergrens voor een substantiële bijdrage werd een waarde van 0.14 (Cohen, 1988) gehanteerd.

 DOI: 10.59532/hsnnhe76

Resultaten

In Tabel 2 staat voor elke cel, voor elk van de vier uitkomsten het gemiddelde over de honderd replicaties en in Tabel 3 staat de proportie van de variantie die de factoren wisten te verklaren.

Een toelichting op de vier uitkomsten:

Uitkomst 1. Het percentage vragen waarvoor de p-waarde onder de kritieke waarde viel bleek voornamelijk af te hangen van de verdeling van de moeilijkheid binnen de toets ($\eta^2 = 0.572$), wat uiteraard een triviale uitkomst was. Aangezien de platte itemverdeling ook moeilijke vragen bevatte, werden ze in deze conditie ook gemarkeerd terwijl dat nooit het geval was bij de spitse verdeling. Het gemiddelde percentage vragen dat bij de platte vraagverdeling werd gedetecteerd liep van 8 tot 16 %. De spreiding van de studentvaardigheid had een beduidend kleiner effect ($\eta^2 = 0.175$): bij een homogene groep studenten was het percentage iets hoger dan bij een heterogene groep.

Uitkomst 2. Het percentage vragen met een r_{it} van tussen de 0 en 0.15 werd met name door de spreiding van de vraagmoeilijkheid ($\eta^2 = 0.324$) en de spreiding van de vaardigheid ($\eta^2 = 0.275$) beïnvloed. N speelde een kleinere, maar toch substantiële rol ($\eta^2 = 0.166$) waarbij het percentage afwijkende vragen bij de kleine groepen duidelijk hoger was dan bij de grote groepen. In de condities met zowel een platte verdeling van vraagmoeilijkheid als een homogene verdeling van de vaardigheid, en waar driekeuzevragen werden gebruikt, werd ongeveer één op de vijf vragen gemarkeerd als extreem (dat gold voor beide waarden voor N en beide waarden voor K).

Uitkomst 3. Het percentage vragen met een negatieve r_{it} liet een soortgelijk patroon zien, waarbij het percentage gemarkeerde vragen het meest afhing van N ($\eta^2 = 0.349$), gevolgd door, respectievelijk, de spreiding van de vraagmoeilijkheid ($\eta^2 = 0.301$) en de spreiding van de vaardigheid ($\eta^2 = 0.198$). Daar waar in de condities met kleine groepen, bij een platte verdeling van de vraagmoeilijkheid en een homogene in plaats van een heterogene verdeling van de vaardigheid, percentages vragen tussen de 6 en 10 % werden gemarkeerd, was dit percentage dicht bij 0 in dezelfde condities in het geval van grote groepen.

Uitkomst 4. De waarde van de betrouwbaarheid, zoals geschat met α, was vooral een functie van de verdeling van de vaardigheid ($\eta^2 = 0.484$), waarbij heterogene groepen studenten hogere waarden hadden dan homogene groepen, een bekend fenomeen uit de testleer (bijvoorbeeld, Van Naerssen, 1969, p. 238). Het soort vraag maakte ook iets uit ($\eta^2 = 0.184$), waarbij meerkeuzevragen wat lagere waarden hadden dan open vragen. In zes van de 32 cellen was het gemiddelde van de geschatte α onder de grens van 0.70 (voor een 'middelmatige' toets, zie Tabel 1). De laagste waarde, 0.53, werd gevonden in het geval van 30 driekeuzevragen in toetsen met een platte verdeling van de moeilijkheid en een homogene populatie. De hoeveelheid vragen speelde ook een rol ($\eta^2 = 0.159$), maar dat is een triviale uitkomst bij toetsanalyses.

Tabel 2 Uitkomsten van de simulatiestudie met de *vijf factoren*: aantal studenten (N), soort vragen (Vraagtype), toetslengte (K), spreiding van de vraagmoeilijkheid in de toets (Verdeling β), en spreiding van de vaardigheid van studenten (Populatie: homogeen of heterogeen), en de *vier centrale uitkomsten*: het percentage vragen waarvoor de ondergrenzen werden overschreden ($p < p'$), met een r_{it}-waarde tussen 0 en 0.15 (de ondergrens van het onderscheidingsvermogen), met een negatieve r_{it}-waarde, en de geschatte Cronbach's α van de itemscores

N	**Vraagtype**	K	**Verdeling** β	**Populatie**	$p < p'$	$0 \leq r_{it} < 0.15$	$r_{it} < 0$	α
50	driekeuze	30	plat	heterogeen	12	6	3	0.78
50	driekeuze	30	plat	homogeen	16	18	8	0.53
50	driekeuze	30	spits	heterogeen	0	2	0	0.87
50	driekeuze	30	spits	homogeen	0	12	2	0.66
50	driekeuze	60	plat	heterogeen	11	7	3	0.88
50	driekeuze	60	plat	homogeen	14	22	10	0.67
50	driekeuze	60	spits	heterogeen	0	2	0	0.93
50	driekeuze	60	spits	homogeen	0	15	3	0.81
50	open	30	plat	heterogeen	9	3	2	0.89
50	open	30	plat	homogeen	14	13	6	0.71
50	open	30	spits	heterogeen	0	0	0	0.93
50	open	30	spits	homogeen	0	5	1	0.78
50	open	60	plat	heterogeen	8	3	1	0.95
50	open	60	plat	homogeen	14	14	7	0.83
50	open	60	spits	heterogeen	0	0	0	0.97
50	open	60	spits	homogeen	0	7	1	0.88
200	driekeuze	30	plat	heterogeen	9	2	0	0.79
200	driekeuze	30	plat	homogeen	14	17	2	0.53
200	driekeuze	30	spits	heterogeen	0	0	0	0.87
200	driekeuze	30	spits	homogeen	0	1	0	0.68
200	driekeuze	60	plat	heterogeen	9	2	0	0.88
200	driekeuze	60	plat	homogeen	14	21	2	0.69
200	driekeuze	60	spits	heterogeen	0	0	0	0.93
200	driekeuze	60	spits	homogeen	0	3	0	0.81
200	open	30	plat	heterogeen	9	1	0	0.90
200	open	30	plat	homogeen	14	11	1	0.72
200	open	30	spits	heterogeen	0	0	0	0.93
200	open	30	spits	homogeen	0	0	0	0.79
200	open	60	plat	heterogeen	8	0	0	0.95
200	open	60	plat	homogeen	14	12	1	0.84
200	open	60	spits	heterogeen	0	0	0	0.96
200	open	60	spits	homogeen	0	0	0	0.88

DOI: 10.59532/hsnnhe76

Tabel 3 Geschatte effectgrootes van de *vijf factoren*: aantal studenten (N), soort vragen (Vraagtype), toetslengte (K), spreiding van de vraagmoeilijkheid in de toets (Verdeling β), en spreiding van de vaardigheid van studenten (Populatie: homogeen of heterogeen) op de *vier centrale uitkomsten, de indicatoren*: het percentage vragen waarvoor de ondergrenzen werden overschreden ($p < p'$), met een r_{it}waarde tussen 0 en 0.15 (de ondergrens van het onderscheidingsvermogen), met een negatieve r_{it}-waarde, en de geschatte Cronbach's α van de itemscores

Indicator	N	**Vraagtype**	K	**Verdeling** β	**Populatie**
$p < p'$	0.006	0.012	0.004	0.572	0.175
$0 \leq r_{it} < 0.15$	0.166	0.083	0.001	0.324	0.275
$r_{it} < 0$	0.349	0.028	0.000	0.301	0.198
α	0.001	0.184	0.159	0.094	0.484

Sterker nog, de verdubbeling van vragen (van 30 naar 60) liet een toename in de gemiddelde α's zien die nagenoeg perfect te beschrijven was met de Spearman-Brown regel voor toetsverlenging.

Conclusies voor toetsconstructeurs

In deze studie bestonden de toetsen uit alleen kwalitatief goede items. Deze items werden bij een steekproef met een specifieke omvang, spreiding in vaardigheid etcetera, afgenomen. Door steekproeffluctuaties kan het zijn dat bij een bepaalde afname de vuistregels aangeven dat een bepaald item of de hele toets 'slecht' is, zelfs als dit item of deze toets op populatieniveau goed functioneert. Als dat gebeurt is er sprake van een 'vals alarm'. Hieronder wordt, *in volgorde van verklarende kracht* in de simulatie, de impact van iedere factor besproken, en ook in welke mate toetsconstructeurs er invloed op kunnen uitoefenen.

1. De spreiding van de moeilijkheid van de vragen (als alle vragen ongeveer even moeilijk zijn, dan is er weinig spreiding in moeilijkheid). Toetsen met meer spreiding geven slechtere indicatorwaarden dan toetsen met minder spreiding (zie bijv. Crocker & Algina, 1986, pp. 97–98), al was dit effect in de simulatie duidelijker op vraag- dan op toetsniveau. Toetsconstructeurs in het hoger onderwijs hebben vaak geen controle over de spreiding van vraagmoeilijkheid. Bijvoorbeeld, als de leerdoelen een oplopende moeilijkheidsgraad hebben zoals onthouden, toepassen en evalueren, of als cursusonderdelen op elkaar voortbouwen, is het lastig om vragen van gelijke moeilijkheid te ontwikkelen. Bovendien is eerder betoogd dat

het manipuleren van de vraagmoeilijkheid van criteriumgeoriënteerde toetsen de inhoudsvaliditeit schaadt (Ebel & Frisbie, 1991, p. 223).

2. De spreiding van vaardigheid van studenten. Hier geldt hoe kleiner de vaardigheidsverschillen hoe vaker alarmerende indicatorwaarden optreden. In de simulatie bleek dit uit extremere p-waarden enerzijds, en lagere r_{it}'s en α's anderzijds. Ook deze factor staat niet onder de controle van de toetsconstructeur (en is vaak onbekend), maar kan wel tot opvallende uitkomsten leiden. Bijvoorbeeld, bij opleidingen die een strenger toelatingsbeleid op studentcapaciteiten hanteren, zullen bij toetsafnames de indicatoren vaker alarmerende waarden hebben. Het tegenovergestelde kan ook: indien naast een gebruikelijke groep studenten enkelen de afname onvoorbereid bezoeken (wellicht om te kijken hoe de toets eruitziet), kunnen de indicatoren juist kunstmatig rooskleurig worden (Van den Brink, 1977).
3. Het aantal studenten. Schattingen van statistische indicatoren variëren van steekproef tot steekproef. De spreiding in deze schattingen wordt groter naarmate de groep studenten kleiner is. De simulatie toonde dat bij een kleine N vooral het onderscheidingsvermogen vaak alarmerende waarden vertoont, terwijl dat beduidend minder het geval is voor p-waarden en betrouwbaarheid (zie bijv. ook Crocker & Algina, 1986, pp. 322–325; De Groot & Van Naerssen, 1969, par. 17.6). Het aantal studenten is voor de toetsconstructeur een gegeven, maar het is belangrijk dat men weet dat bij kleine groepen de spreiding van sommige indicatoren zo groot is dat afwijkende waarden vaak kunnen voorkomen.
4. Het soort vraag. In vergelijking met open vragen bieden gesloten vragen de mogelijkheid om te gokken waardoor een student die het antwoord niet weet de vraag soms tóch goed maakt. Dit verhoogt de spreiding van de meetfouten (Crocker & Algina, 1986, H. 6), wat in de simulatie tot uiting kwam in (iets) lagere betrouwbaarheidsschattingen bij driekeuzevragen dan bij open vragen. Het soort vragen dat in een toets wordt gebruikt staat onder de controle van de toetsconstructeur en hangt af van de hoeveelheid constructie- en nakijktijd die beschikbaar is. Gesloten vragen kosten meer tijd om te ontwikkelen, maar zijn sneller nagekeken dan open vragen.
5. Het aantal vragen. Dit is waarschijnlijk de bekendste factor. Bij meer vragen neemt de spreiding in ware scores relatief sterker toe dan de spreiding in meetfouten (zie bijv. Crocker & Algina, 1986, H. 6). Het zal dan ook niet verrassen dat langere toetsen een hogere betrouwbaarheid hadden. Over deze factor heeft de toetsconstructeur vaak geen controle omdat de tijd die beschikbaar is voor het construeren van vragen en het afnemen van de toets doorgaans vaststaan. Wellicht is het wel een eye-opener dat de spreiding in de populatie meer impact kan hebben dan het aantal vragen.

De simulatie toonde aan dat vooral de combinatie van weinig spreiding in vaardigheid, veel spreiding in vraagmoeilijkheid in een kleine groep studenten ten onrechte kunnen

leidden tot markeringen van 'slechte' vragen. Met name in situaties waarin de toetsconstructeur over deze factoren weinig controle heeft, is het aan te bevelen om de vuistregels minder strikt te nemen.

Discussie

In deze bijdrage werd aan het licht gebracht dat gangbare vuistregels voor indicatoren van toetskwaliteit in veel gevallen te strikt zijn. Ze zijn oorspronkelijk vastgelegd voor de selectie van vragen voor een gestandaardiseerd instrument in een normgeoriënteerde toetsituatie waarbij er sprake is van veel spreiding in vaardigheid, weinig spreiding in itemmoeilijkheid, en een afname in een grote groep studenten. De gangbare vuistregels zijn daardoor niet algemeen bruikbaar. Er werden vier vragen geïntroduceerd die toetsconstructeurs kunnen stellen om te bepalen hoeveel gewicht er aan de vuistregels moet worden toegekend. Indien er geen aanwijsbare oorzaak is, er geen sprake is van (puur) normgeoriënteerd toetsen, het geen geijkte toets betreft, er weinig spreiding is in de vaardigheid, of het een kleine groep studenten betreft, kan men uitkomsten die volgens de gangbare vuistregels afwijken met een flinke korrel zout nemen.

Een te rigide verkondiging van de vuistregels en de suggestie dat afwijkende indicatorwaarden vragen om aanpassingen heeft als risico dat toetsconstructeurs '*confirmation bias*' (Nickerson, 1998) ontwikkelen en vaak goede vragen verwijderen waardoor de kwaliteit van toetsen juist wordt aangetast. Er wordt hier dan ook de wens uitgesproken dat toetshandboeken en -software de verwoording minder stellig maken, en dat er ook wordt verwezen naar de alternatieve verklaringen die hiervoor werden beschreven. We willen benadrukken dat we geen tegenstanders zijn van het gebruik van statistische indicatoren. Integendeel, we achten het, net als het opstellen van een toetsmatrijs, een essentiële stap in het toetsproces, maar ze moeten wel op een verstandige wijze worden ingezet.

Daarnaast zouden toetsconsructeurs enorm geholpen kunnen worden indien de verzameling van gangbare indicatoren zou worden uitgebreid en/of anders zou worden ingezet:

Ten eerste, studietoetsen zijn vaak opgebouwd uit vraagclusters die horen bij specifieke onderdelen van de leerstof zoals hoofdstukken uit een studieboek. Hierdoor hangen vragen uit hetzelfde cluster meer met elkaar samen dan met vragen uit andere clusters, en zijn de itemscores dus niet zuiver unidimensioneel. Dit kan tot een onderschatting van de indicatoren leiden. Een oplossing zou zijn om de indicatoren (ook) voor afzonderlijk clusters te berekenen of indicatoren toe te voegen die rekening houden met meerdimensionaliteit.

Ten tweede, zoals de simulatie liet zien, zijn de indicatoren sterk afhankelijk van statistische grootheden als N en spreiding waardoor er in de schattingen veel onzekerheid

zit. Om valide uitspraken over de kwaliteit van toetsen te doen, is het noodzakelijk dat de puntschattingen van de indicatoren worden uitgebreid met betrouwbaarheidsintervallen die deze onzekerheid kwantificeren (zie bijv. Van der Ark, 2025).

Ten derde, de gangbare indicatoren zijn afkomstig uit de klassieke testleer, en daarbij geldt dat ze populatieafhankelijk zijn. Zoals uit de simulatie bleek heeft een lage vaardigheidsspreiding een negatieve impact op de indicatoren, en lijken toetsen vaak kunstmatig slecht. De oplossing is om aanvullend de standaardmeetfout te evalueren. Deze is veel minder gevoelig voor spreiding (Standards, 2014, p. 39), en kan bijvoorbeeld worden gebruikt om rond de cesuurscore te bepalen hoe precies zak/slaagbeslissingen waren.

Een kanttekening aangaande de spreiding van itemmoeilijkheid is hier op zijn plaats. In het verleden is vaak aanbevolen om te zorgen voor gemiddeld moeilijke vragen met onderling weinig verschillen in moeilijkheid. In de huidige simulatiestudie werd dat bevestigd. Echter, dat komt omdat zowel die aanbeveling als de gangbare indicatoren afkomstig zijn uit de klassieke testleer. In de moderne testleer is de meetprecisie niet langer voor een populatie personen gedefinieerd, maar voor de score van een individu waarbij de precisie kan variëren over de schaal van de toetsscore (Van den Brink & Mellenbergh, 1998). Daarom is deze aanbeveling achterhaald. In de moderne aanpak hangt het bijvoorbeeld van het toetsdoel af hoe de vraagmoeilijkheid verdeeld moet zijn: Voor normgeoriënteerd toetsen is veel spreiding in moeilijkheid gewenst terwijl voor criteriumgeoriënteerd toetsen de moeilijkheid in de buurt van de cesuurscore moet liggen (zie bijv. Luecht, 2011).

Dankwoord

De auteurs danken Prof. Dr. Andries van der Ark voor zijn waardevolle commentaar op een eerdere versie van deze bijdrage.

Literatuur

Ans. (2025). *Question insights*. Geraadpleegd op 3 maart 2025 van: https://support.ans.app/hc/en-us/articles/360027234814-Question-insights.

Biggs, J., & Tang, C. (2011). *Teaching for Quality Learning at University: What the Student Does* (4de dr.). Open University Press.

Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2de dr.). Lawrence Erlbaum Associates.

Crocker, L.M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart; Winston.

De Groot, A.D., & Van Naerssen, R.F. (1969). *Studietoetsen: construeren, afnemen, analyseren*. Mouton.

Di Bello, M. (2021). When statistical evidence is not specific enough. *Synthese*, *199*(5), 12251–12269.

Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of Educational Measurement* (5de dr.). Prentice-Hall.

Eggen, T.J.H.M., & Sanders, P.F. (1993). *Psychometrie in de praktijk*. Cito Instituut voor Toetsontwikkeling. https://cito.nl/kennisbank-stichting-cito/psychometrie-in-de-praktijk/

Elffers, L. (2020). Het diploma als mythe. *Didactief*, *50*. 39.

Gigerenzer, G., Todd, P.M., & the ABC Research Group. (2000). *Simple heuristics that make us smart*. Oxford University Press.

Luecht, R.M. (2011). Designing tests for pass–fail decisions using Item Response Theory. In M.S. Downing & T.M. Haladyna (Red.), *Handbook of test development* (pp. 575–596). Lawrence Erlbaum Associates.

Maassen, N.A.M., Otter, D. den, Wools, S., Hemker, B.T., Straetmans, G.J.J.M., & Eggen, T.J.H.M. (2015). Kwaliteit van toetsen binnen handbereik: Reviewstudie van onderzoek en onderzoeksresultaten naar de kwaliteit van toetsen. *Pedagogische studiën*, *92*(6), 280–293.

Martens, R., & Moerkerke, G. (2017). Ontwikkelen van toetsbeleid. In H. Van Berkel, A. Bax, & D. Joosten-ten Brinke (Red.), *Toetsen in het hoger onderwijs* (4de dr., pp. 51–63). Bohn Stafleu Van Loghum.

Molkenboer, H., & Kempers, A. (2017). Oppassen met hergebruik van tentamenvragen. *Examens: Tijdschrift voor de Toetspraktijk*, *1*, 7–10.

Montgomery, D.C. (2019). *Introduction to Statistical Quality Control* (8e dr.). Wiley.

Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Rudolph, M.J., Daugherty, K.K., Ray, M.E., Shuford, V.P., Lebovitz, L., & DiVall, M.V. (2019). Best Practices Related to Examination Item Construction and Post-hoc Review. *American Journal of Pharmaceutical Education*, *83*(7), 7204. https://doi.org/10.5688/ajpe7204

Schaap, S. (2023). *Examencommissies universiteiten zuchten onder stijgend aantal studenten*. Geraadpleegd op 3 juli 2023 van: https://nos.nl/artikel/2481095-examencommissies-universiteiten-zuchten-onder-stijgend-aantal-studenten.

Sigal, M.J., & Chalmers, R.P. (2016). Play It Again: Teaching Statistics With Monte Carlo Simulation. *Journal of Statistics Education*, *24*(3), 136–156. https://doi.org/10.1080/10691898.2016.1246953

Standards. (2014). *Standards for educational and psychological tests* [American Psychological Association and American Educational Research Association and National Council on Measurement in Education]. American Psychological Association.

TestVision. (2024). [Digitaal toetssysteem]. Teelen.

Thompson, N.A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, *69*(5), 778.

Van Belle, G. (2008). *Statistical rules of thumb* (2de dr.). Wiley.

Van Berkel, H. (2011). Meten is weten; vergeet het maar! *Examens: Tijdschrift voor de Toetspraktijk*, *8*, 10–14.

Van Berkel, H. (2017). Toetsen en de rechtsbescherming van examinandi. In H. Van Berkel, A. Bax, & D. Joosten-ten Brinke (Red.), *Toetsen in het hoger onderwijs* (4de dr., pp. 317–322). Bohn Stafleu Van Loghum.

Van Berkel, H., & Bax, A. (2017). Het meten van toetskwaliteit. In H. Van Berkel, A. Bax, & D. Joosten-ten Brinke (Red.), *Toetsen in het hoger onderwijs* (4de dr., pp. 24–36). Bohn Stafleu Van Loghum.

Van Berkel, H., & Bax, A. (2023). Het meten van psychometrische toetskwaliteit. In H. Van Berkel, A. Bax, D. Joosten-ten Brinke, K. Beekman, & T. van Schilt-Mol (Red.), *Toetsen in het hoger onderwijs* (5de dr., pp. 81–94). Bohn Stafleu Van Loghum.

Van den Brink, W.P. (1977). Het verken-effect. *Tijdschrift voor Onderwijsresearch*, *5*(2), 253–261.

Van den Brink, W.P., & Mellenbergh, G.J. (1998). *Testleer en testconstructie*. Boom.

Van der Ark, L.A. (2025). Standard errors for reliability coefficients. *Psychometrika*. https://doi.org/10.1017/psy.2025.10050

Van der Kolk, B. (2021). *De meetmaatschappij*. Atlas Contact.

Van Loosbroek, S. (2019). *Gerecycled tentamen psychologie geldig verklaard* [Mare: Leids Universitair Weekblad]. Geraadpleegd op 3 juli 2025 van: https://www.mareonline.nl/nieuws/gerecycled-tentamen-psychologie-geldig-verklaard.

Van Naerssen, R.F. (1969). De interpretatie van indices. In A.D. De Groot & R.F. Van Naerssen (Red.), *Studietoetsen: construeren, afnemen, analyseren*. Mouton.

Woods, C.M. (2008). Consequences of Ignoring Guessing When Estimating the Latent Density in Item Response Theory. *Applied Psychological Measurement*, *32*(5), 371–384. https://doi.org/10.1177/0146621607307691

The current rules of thumb for test quality indicators are too strict

Abstract In this article, it is claimed that, although the application of indicators of test quality in Higher Education is expedient, the common rules of thumb for evaluating them are often too strict. They were originally specified for the selection of questions for a standardized instrument in a norm-oriented testing situation in which there is a large difference in ability among students, questions that are all of average difficulty, and administrations of large groups of students, making them inapplicable for general use. Four questions were introduced that test constructors can ask to determine how seriously these rules of thumb should be taken: (i) What is the explanation? (ii) What is the test goal? (iii) What is the type of instrument? and (iv) What about variability? Finally, suggestions for adjusting the rules of thumb are presented.

Keywords educational tests, test construction, test quality, reliability