Statistics & Methods

Onderwerp: Itemanalyse,betrouwbaarheid en toetskwaliteit.

INHOUD

Itemanalyse,betrouwbaarheid en toetskwaliteit.

LINK: EXPLANATION ITEM ANALYSIS PARAMETERS: STATSOFT
http://www.statsoft.com/textbook

ITEM ANALYSE

Reliability: Inleiding

De reliability (meestal vertaald als "betrouwbaarheid") van een instrument dat wordt gebruikt om een bepaald fenomeen te meten is van belang omdat dit dient als een indicator over datgene wat men werkelijk meet of wat men wil meten: "kan ik er van op aan". De reliability is in feite een maat voor het vertrouwen dat men in een meetinstrument mag hebben en beschrijft de consistentie waarmee het fenomeen wordt gemeten, bijvoorbeeld met een tentamen of toets met vragen gericht op een bepaald (kennis-)domein. Misschien is het verstandiger om hier niet te spreken van toetsen van een bepaald kennisdomein maar van het toetsen van het "framework of knowledge", gerelateerd aan de totale context van de "body of knowledge". Immers in het beste geval kan men slechts een essentiele fractie (skeleton) van het totale kennisgebied toetsen. Het te testen domein kan op diverse manieren en met diverse toetsvormen worden onderzocht, bijvoorbeeld met een reeks items (vragen) over een bepaald onderwerp. De reliability van deze meting wordt uitgedrukt als de verhouding tussen de variabiliteit van de gemeten scores [1] ten opzichte van de ware scores [2]. De ware score is het gemiddelde en is dus de eindscore die is behaald. De variabiliteit wordt uitgedrukt in de gestandaardiseerde spreiding, zie hierna. Hoe groter deze ratio, des te "betrouwbaarder" is het gemeten resultaat, althans in theorie. De mate waarin binnen het te toetsen domein de toetsvragen correleren met de ware scores is derhalve een maat voor de betrouwbaarheid van die vragen en dus de gehele toets.
Het vaststellen van de betrouwbaarheid kan men toepassen op (korte) open vragen, Ja/Nee vragen en in het bijzonder op Multiple Choice toetsen met gesloten vragen en afleiders (alternatieven), waarbij de benodigde indicatoren voor een toets- en itemanalyse volledig kunnen worden benut.

De itemanalyse geeft antwoord op de volgende vragen:

° Hoe moeilijk was een bepaald item voor een groep respondenten ?

° Hoe aantrekkelijk waren de afleidende keuzes of antwoorden.

° Zijn het vooral personen die verondersteld worden kennis van zaken te hebben die een bepaald item goed beantwoord hebben : -> hoe goed is het onderscheidend vermogen van dat bepaalde item?

° Hoe goed past het item bij de test of toets ?

Bij itemanalyses, bijvoorbeeld uitgevoerd op toets vragen, spelen de betrouwbaarheid van het meetsysteem en de bijbehorende vragen (items) een grote rol. Om hier inzicht in te krijgen is een (computer gestuurde) itemanalyse vereist.

Tijdens een (MC-)toets afname komen als het ware van meerdere kanten een grote hoeveelheid informatie, kennis, en ervaring samen (de leerstof, de items, de afleiders,

de kandidaten). Het is zaak voor de toets creator (docent) om een zodanige toets constructie te ontwerpen dat daarmee met zo min mogelijk ruis en andere storende factoren de kernzaken uit de leerstof gemeten kunnen worden. Slaagt men hierin dan mag men spreken van een betrouwbare en kwalitatief aanvaardbare (MC)toets waarmee het competentie niveau van de kandidaten zo goed mogelijk geschat wordt. Slaagt men hierin niet, dan is de toetsbeoordeling in feite zinloos.

Men verkrijgt ook geen hogere betrouwbaarheid door alleen maar gebruik te maken van bepaalde software (bv. QuestionMark-Perception al of niet gekoppeld aan een Elektronische leeromgeving) waarmee een itembank van vragen aangemaakt en gegenereerd wordt als men niet van te voren ervaring heeft opgedaan met het construeren van valide (toets)vragen en rekening houdt met de verschillende cognitieve niveaus die men kan toetsen.
De itemanalyse is een onderdeel van de **Psychometrie** (kwantitatieve - getalsmatig uitdrukbare - klassieke test-theorie) en deze houdt zich onder meer bezig met het verschaffen van inzicht in de validiteit van de vragen en betrouwbaarheid van het meetsysteem (bv. bij Multiple Choice-items). Het beschrijft de metingen zelf , dus niet de personen, onderwerpen of deelnemers, maar het domein. Hierbij is essentieel het vaststellen van de validiteit en de betrouwbaarheid van vragen. Men tracht onder meer van de items (itemscores) de meetfouten (dwz de varianties) tussen [1] de waargenomen scores en [2] de werkelijke scores zo veel mogelijk te beperken.

De verhouding [1] / [2] kan worden geschat met de **Cronbach´s alpha of de Kuder Richardson coefficient** (zie hierna). Men kan een redelijk goede itemanalyse uitvoeren bij een steekproef vanaf tenminste 30 proefpersonen of kandidaten die aan een totaal van tenminste 20 vragen over eenzelfde onderwerp (kennisdomein) worden onderworpen. Vereist is een groep kandidaten die refereert aan een normaal verdeelde populatie. Dat betekent dat men moet zorgen voor een testgroep van kandidaten die voorwat betreft hun te testen kennis niet eenzijdig asymmetrisch is verdeeld en niet te klein is. Bij een itemanalyse wil men vooral weten of middels een aantal metingen (items, vragen) die allen gebaseerd zijn op dezelfde meetmethode, er sprake van is dat men hetzelfde kenmerk of verschijnsel meet. Hierbij gaat men globaal als volgt te werk:

Men stelt eerst de variantie vast tussen de gemeten itemscores en vervolgens past men een analyse toe om na te gaan in hoeverre de metingen per item onderling van elkaar verschillen. Indien de afwijkingen tussen de metingen gering is spreekt men van een consistentie meetmethode. Hieruit concludeert men dat de meetmethode betrouwbaar is Dit betekent in de praktijk dat men er op mag vertrouwen dat er nauwelijks sprake is van “toevallig goed beantwoorde vragen”. Deze situatie wordt bijvoorbeeld toegepast op een polytome situatie : Multiple Choice vragen met (3 of meer) discrete alternatieven.

TOP

Validiteit:

Een test-meet systeem is valide indien het meetinstrument meet wat het moet meten. Tegelijk wil de geteste kandidaat weten of de meting (de beoordeling) op terechte gronden is gebeurd. Elk meetinstrument is geconstrueerd met een bepaald DOEL. Bv, een weegschaal is bedoeld om (lichaams)gewicht vast te stellen en geen lichaamslengte. Bovendien, men wil niet alleen weten DAT men iets weegt, maar ook HOEVEEL precies. De begrippen validiteit en betrouwbaarheid zijn dus aan elkaar gekoppeld.
Een meetinstrument moet daarom zo nauwkeurig (precies) mogelijk meten. Om te onderzoeken of een toets zowel betrouwbaar als ook valide is heeft men in feite meerdere metingen nodig (herhaling van de zelfde meetsituatie bij dezelfde of een vergelijkbare groep proefpersonen) of men vergelijkt twee verschillende meet-methoden met elkaar (een geijkte en een experimentele methode) in eenzelfde testsituatie of onderzoeksopzet.

In de regel wordt geadviseerd om in een test-meet situatie met toetsvragen tenminste 20 MC vragen aan te bieden bij een (steekproef)populatie van tenminste 30 proefpersonen.

Verzamelingen, Ruis, Meetfouten, Systeemfouten:

De data-matrix van een toets met MC items gebaseerd op K (bv.K=4) alternatieven bestaat ruimtelijk bezien uit de volgende dimensies (domeinen of verzamelingen):

1) Verzameling [N] items [de toetsvragen]
2) Verzameling [N,K] item-alternatieven (inclusief Item-sleutel)
3) Deel verzameling [N] Item-sleutels
4) Verzameling [P] kandidaten
5) Deel verzameling [P,N] correcte responsen [zie hierna, score-matrix].
6) Deel verzameling [P,N,K] alternatieve responsen (keuzes voor een alternatief)

In elk van deze (deel)verzamelingen kunnen onvolkomenheden of fouten optreden. Dit betreft toevallige fouten "random errors" of ruis) maar ook systeemfouten : fouten die te maken hebben met de constructie van de toets als meetinstrument. Daarnaast kunnen er fouten optreden indien men de toets handmatig (tijdrovend,vermoeiend) nakijkt. Dit nog afgezien van de situatie dat het te toetsen onderwerp soms niet altijd volledig wordt "afgedekt" door de aangeboden selectie van de items zelf (niet representatief).

De enige manier om hier achter te komen is een computer gestuurde analyse, die snel verloopt en een betrouwbare "output" van de toets-indicatoren geeft.

De verkregen toets score X (of overall P-waarde) is het gemiddelde van een normale verdeling. Maar elke meting met behulp van een vraag is onderhevig aan meetfouten (ruis, Random errors, Bias). Hierdoor wordt de score-spreiding rondom het score-gemiddelde X sterk beinvloed. Met een MC toets- meetsysteem heeft men te maken met twee soorten spreidingen, uitgedrukt in varianties, die in feite deel uitmaken van dezelfde verzameling:
De **toets variantie** (variantie gericht op de correcte responsen van de P kandidaten) en de **item variantie** (variantie gericht op alle N x P item-responsen). Beide varianties zijn onderheving aan ruis. Bij een MC-toets is de mate van ruis van alle aangeboden vragen en alle kandidaten die de toets maken van invloed op de indicatoren voor de betrouwbaarheid van een toets : de Cronbachs alpha coefficient of de Kuder Richardson coefficient (KR-20). Hoe lager de waarde van deze indicatoren, des te meer ruis (R) en / of systeemfouten (S) er kunnen bestaan. Deze beïnvloeden de standaardmeetfout en de totale toetsbetrouwbaarheid.

°te weinig toetsvragen :kans op toevallig "goed" scoren is hoog (R) °te korte toetstijd (R)
°te grote gokkans bv. bij Ja/Nee vragen, te moeilijke vragen (R en S)
°de mate van heterogeniteit van de samenstelling van de groep (R)
°onjuiste selectie van de groep (R)
°ondeugdelijke, niet eenduidig geformuleerde toetsvragen (S)
°geen goede spreiding tussen moeilijke en makkelijke toets-items (R)
°ongewenste storingen tijdens de toetsafname (R)
°de aanwezigheid van herkansers in een populatie (leereffect) (R)
°slechte afleiders (S)
°vragen die niet relevant zijn voor de toetsstof (S)

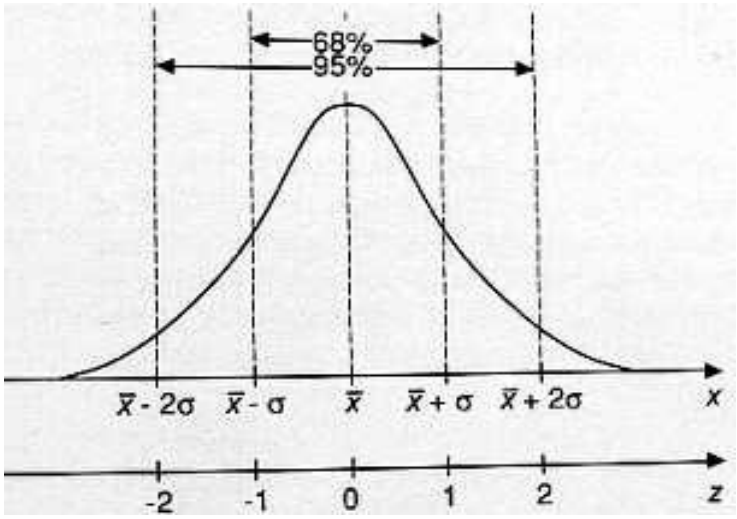
Aan factoren die de ruis (R) beinvloeden is soms wel snel iets te doen, maar systeemfouten maken de toets als meetinstrument onbetrouwbaar en vooral minder valide.

Opmerking:

Hoe heterogener de te testen groep (maar wel binnen de grenzen van een normale verdeling), des te hoger is de toetsvariantie en des te beter kan het meetsysteem schatten tussen verschillen in toetsbeheersing. Maar hoe homogener de items zelf, des te lager is de itemvariantie.

Indien toetsvariantie = itemvariantie dan is de toetsbetrouwbaarheid of KR-20 =0.

De score verzameling:



Bekijkt men de verzameling van alle toetsscores van alle deelnemers dan schommelen deze met een zekere spreiding rondom een bepaald toetsgemiddelde, "de overall P-waarde, of gemiddelde toetsscore X". Een dergelijke situatie is karaktersitiek voor een groep mensen waarvan een meetbaar kenmerk "normaal verdeeld" is. Men kan nu een gebied of interval (zie figuur hierboven) van 1x (68%) of 2x (95%) de standaard-score-meetfout (SE,s) rondom deze overall P-waarde (X) bepalen om de kandidaten alsnog in aanmerking te laten komen voor het slagen van de toets. Kiest men voor een 95 % interval dan betekent dit dat men er maximaal 95 % zeker van kan zijn van alle kandidaten van wie de toetsscore binnen dit interval ligt, hun score representatief is voor de gemiddelde totaaltoetscore X, en in statistisch opzicht onderling niet van elkaar verschillen!. Voor meer informatie over meetfouten KLIK : Meetfouten & Steekproeven

TOP

INDICATOREN:

Om vast te stellen of een toets betrouwbaar, representatief en valide is moet men een groot aantal indicatoren (parameters) bepalen.

De onderstaande indicatoren (item-parameters) worden dikwijls gebruikt om te beoordelen of een bepaald toets-item bruikbaar is. Deze indicatoren zijn (om allerlei redenen) echter NIET stabiel in de tijd. Bij elk nieuw ontwikkeld vraagsysteem of toetssysteem behoren deze coefficienten opnieuw te worden bepaald.

Informatie over de kenmerken,indicatoren en operationaliseren van een toets, zie link 1 :Toetsmatrijs, Kennisnet link 2 :Toetswijzer **https://www.toetswijzer.nl/**

P-waarde en Gokkans

De p-waarde van een item is de proportie van het aantal personen dat op een item goed heeft gescoord (S) en de totale populatie deelnemers (P), en geeft de moeilijkheidsgraad van een item aan.

In formule: p-waarde = S / P

Ook heeft natuurlijk de toets zelf een P-waarde. Dit noemt men ookwel de Overall P-waarde (het toetsgemiddelde X). Deze Overall P-waarde is uit twee domeinen te bepalen: ofwel middels alle items zoals zij gescoord zijn door de kandidaten, ofwel via alle kandidaten zoals deze scoren op de items. In beide gevallen verkrijgt men uiteraard hetzelfde getal, het Overall gemiddelde. Maar de spreiding rondom het gemiddelde van deze domeinen is niet hetzelfde.

De betekenis van de p-waarde (van een item maar ook van de toets zelf) wordt verhoogd indien de groep kandidaten groter is. Bij minder dan 20 kandidaten is de p-waarde nog geen betrouwbare norm.

De p-waarde van een item varieert van 0 tot 1. Indien deze p-waarde onder het theoretische kansniveau ligt (bij 4-keuzevragen =0,25, 3-keuze vragen = 0,33 en 2-keuzevragen = 0,50) dan is de vraag op zijn minst "verdacht".
Er wordt in het algemeen gesteld dat een item met een p-waarde = 0.5 (50 % van de kandidaten scoort dit item goed) een maximale bijdrage levert aan de summatieve functie van de toets. Men moet er naar streven om items te construeren waarvan de p-waarde ligt tussen 0.25 en 0.85. Een item met een p-waarde =1 moet men verwijderen. Een gokkans op succes van 25 % bij een enkele MC-vraag met 4 alternatieven betekent dat men er rekening mee houdt dat een kandidaat 1 op de 4 vragen gokt (IK WEET HET NIET, DUS "dan maar ofwel fout ofwel goed gegokt"). De "norm" (of wenselijke norm) voor een gemiddelde P-waarde van de totale MC toets (bij een summatieve toets) met 4-MC-items = 0.625 (dit is de optimale p-waarde van een item met 4 alternatieven). Deze ligt dus hoger dan de cesuur "voldoende" = 0.6 die soms wordt gehanteerd en dat komt omdat men hier de gokkans mee rekent.

De formule om te p-waarde van een item te corrigeren voor de gokkans luidt:
[Pc = Po - (1 - Po)/(k-1)] of anders geschreven [Pc = Po + (P+1)/(k-1)]
Waarin Pc = gecorrigeerde p-waarde, Po = oorsponkelijk gemeten p-waarde, k = aantal alternatieven per MC-item. Een item gecorrigeerd voor de gokkans op dat item valt dus altijd lager uit. Bijvoorbeeld:

Stel een item met 4 alternatieven heeft een p-waarde Po=0.7 en gecorrigeerd voor de gokkans wordt de gecorrigeerde p-waarde Pc dan:
Pc = 0.7 + (0.7-1)/(4-1) = 0.7 - 0.1 = 0.6

Verder kan men met behulp van de zg. **binomiaal-verdeling** berekenen, dat de kans op succes om bijvoorbeeld uit een mc-toetsaanbod van 40 vragen, elk met 4 alternatieven, 22 vragen willekeurig GOED/CORRECT te gokken (gemiddelde "overall" P-waarde = 55%) kleiner is dan 0,01 %. De formule voor deze berekening luidt :

$$P(g) = \frac{N!}{(N-K)!\,K!}\left[\frac{1}{A}\right]^N$$

P(g) = kans op goed "gokken"
N = totaal aantal items
K = Aantal correct
A = Aantal alternatieven per item

Een complicatie is dat kandidaten die de toetsstof redelijk (tot goed) beheersen en een gedeelte van de vragen bewust goed beantwoorden desondanks bij vragen waarop men het antwoord NIET bewust weet, gaan gokken en met wat geluk toch hierbij goed scoren. Bij een persoon die bijvoorbeeld bij een toets met 40 MC vragen met 4 alternatieven het antwoord op 20 vragen bewust weet blijven er 20 vragen over om te gokken.Theoretisch is hier de raadscore op een goed antwoord dus 25 % van 20 vragen = 5 vragen en dat is 12.5 % van het totale vragen aanbod. Dit levert derhalve een extra "bonus" op bovenop het aantal terecht goede scores. Hiermee zouden deze personen ten onrechte voldoende behalen indien de cesuur wordt vastgesteld op 55 % (=22 vragen van de 40). Reden te meer om de raadkans te betrekken bij de cesuur en de beoordeling "Voldoende-
Onvoldoende", waarin men uiteraard het aantal aangeboden toetsvragen moet verdisconteren. Hoe hoger het aanbod van vragen, des te geringer zal deze gok(raad) kans van invloed zijn op de beoordeling Voldoende - Onvoldoende. Bij Ja/Nee vragen is deze kans (zie hierboven) natuurlijk groter dan bij MC vragen met 4 alternatieven. Bij het theoretisch examen voor het rijvaardigheidsbewijs is de norm voor 70 vragen: minimaal 60 goed.

A-waarde =

Proportie van het aantal studenten dat een alternatief (of afleider) heeft gekozen. Hoe hoger de Awaarde uitvalt, des te "aantrekkelijker" is deze kennelijk geweest voor de kandidaat.
Indien de A-waarde < 2 % dan is de vraag zeer waarschijnlijk niet correct geformuleerd of de vraag heeft geen betrekking op het te toetsen domein. Dit is zeker het geval indien tevens de Rit-waarde of DI-waarde =< 0 (zie hierna). De vraag discrimineert dan in het geheel niet en behoort in dezelfde vorm geen tweede maal te worden aangeboden.

Rit =

De Rit of item-totaal correlatie (of item-test correlatie, symbool R): dwz de samenhang/correlatie tussen de itemscore (de respons op de vraag) en de hele toets eindscore (zie ook hierna) . Soms wordt hierbij ook nog een correctie toegepast (de rest-waarde of Rir-waarde) omdat het beschouwde item ook deel uitmaakt van de totale verzameling items van de toets zelf. Het is een maat voor het onderscheidingsvermogen van een bepaald item.
Deze varieert van -1 tot +1. Een Rit-waarde >= 0.45 is uitstekend. Een Rit tussen 0.45 - 0.35 is goed. Een Rit-waarde =< 0.15 is matig tot slecht. Het is een totaal discriminatie-index en geeft aan in hoeverre een item differentieert tussen 'goede' en 'slechte' studenten. 'Goede' studenten zijn studenten met een hoge toetsscore, 'slechte' studenten zijn studenten met een lage toetsscore. Men moet wel van te voren vastleggen wat men in de studenten populatie onder goede en slechte studenten verstaat : bv. 25 % van alle studenten die de hoogste scores bezitten en 25 % van de studenten die de laagste score bezitten.
Ook voor de alternatieven is een item-correlatie te bepalen (de Rar-waarde).

KR20 = Kuder Richardson parameter (zie ook hierna)

Dit is een schatting van de "betrouwbaarheid" van een toets met "gedwongen raden" vragen of gesloten vragen, en varieert van 0 tot 1. Een lage KR20 (dwz =< 65 %) betekent dat

a) men niet goed weet wat de toets precies meet
b) men op een kwalitatief even goede / slechte parallelle toets tot andere beslissingen zou zijn gekomen (men meet infeite niet consistent)

Een manier om iets meer te weten te komen over de betrouwbaarheid van een toets is om de zg. Splithalf-waarde te berekenen. Dit doet men door de betrouwbaarheid van de even- versus de oneven-items met elkaar te correleren. Met deze methode kan ook geschat worden hoeveel MC-items er vereist zouden zijn om tot een bepaalde gewenste KR-20 (bv 80 %) uit te komen.

DI = (D-waarde) Algemene discriminerende waarde voor een bepaald Item. Een oudere benaming voor de Rit waarde waarbij meestal 25 % van de hoogstscorende kandidaten als uitgangspunt wordt gebruikt. De DI waarde kan eveneens varieren van -1 tot +1. Indien de DI waarde < 0, dan meet men een correcte score bij iemand die geen verstand heeft van het te testen domein ! (Men test in dat geval "the man in the street"). Zie hierna.

TOP

Cronbach´s Alpha:

Een maat voor de reliability voor een test-meet-systeem met open vragen kan worden uitgedrukt met de Cronbach alpha coefficient. Hiervoor is het vereist dat men vragen met een unidimensionale of multidimensionale structuur beschouwt, zoals vragen met een Likertschaal waarbij sprake is van een rating-beoordeling (1,2,3,4,5). Cronbachs alpha wordt berekend uit de gemiddelde covariantie tussen de scores. Dit is een maat voor de variantie tussen de itemscores binnen het testsysteem.Een waarde van Cronbach´s alpha 0.8 of hoger is vereist voor een goed test-meetsysteem. Cronbach's alpha is geen statistische test - het is de coefficient van consistentie (=reliability). Cronbach's alpha kan worden geschreven als een functie van het aantal test items EN de gemiddelde inter-correlatie tussen de tems. N = aantal items, r = interne correlatie tussen de items. Zie onderstaande formule

$$\alpha = \frac{N \cdot \bar{r}}{1 + (N-1) \cdot \bar{r}}$$

Ookwel geschreven als:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\Sigma S^2 i}{S^2 x}\right)$$

Waarin:

$$\frac{\Sigma S^2 i}{S^2 x}$$

[De som van alle itemvarianties / de toetsvariantie]

. ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ . ▪ .

TOP

Kuder Richardson:

De Kuder Richardson formule en de Cronbach´s alpha formule meten hetzelfde.

Gebruikt men een gedwongen raden systeem of discrete dichotome MC vragen (point biserial) met p=1=Ja ,correct en q = 1 - p = Nee, niet correct, voor de beoordeling dan is de Cronbach´s alpha berekening identiek aan de Kuder Richardson coefficient (KR20) berekening. Zie onderstaande formule:

Kuder-Richardson 20 (KR20)

- formula $\rho_{KR20} = \left[\frac{k}{k-1}\right]\left[1 - \frac{\sum pq}{\sigma_X^2}\right]$

  - where k = number of items
    - p = the proportion of correct responses
    - q = 1 - p

Beide formules zijn afgeleid van de zg. Spearman-Brown formule . Voor de KR20 formule geldt: k = N = aantal items. De KR20 berekent de reliability coefficient van een set items bij een gedwongen raden systeem (waarbij de toetsvariantie = S^2X en de itemvariantie S^2i = Spq) terwijl Cronbach's alpha wordt gebruikt voor multipoint schalen (Likertschaal) en open vragen.

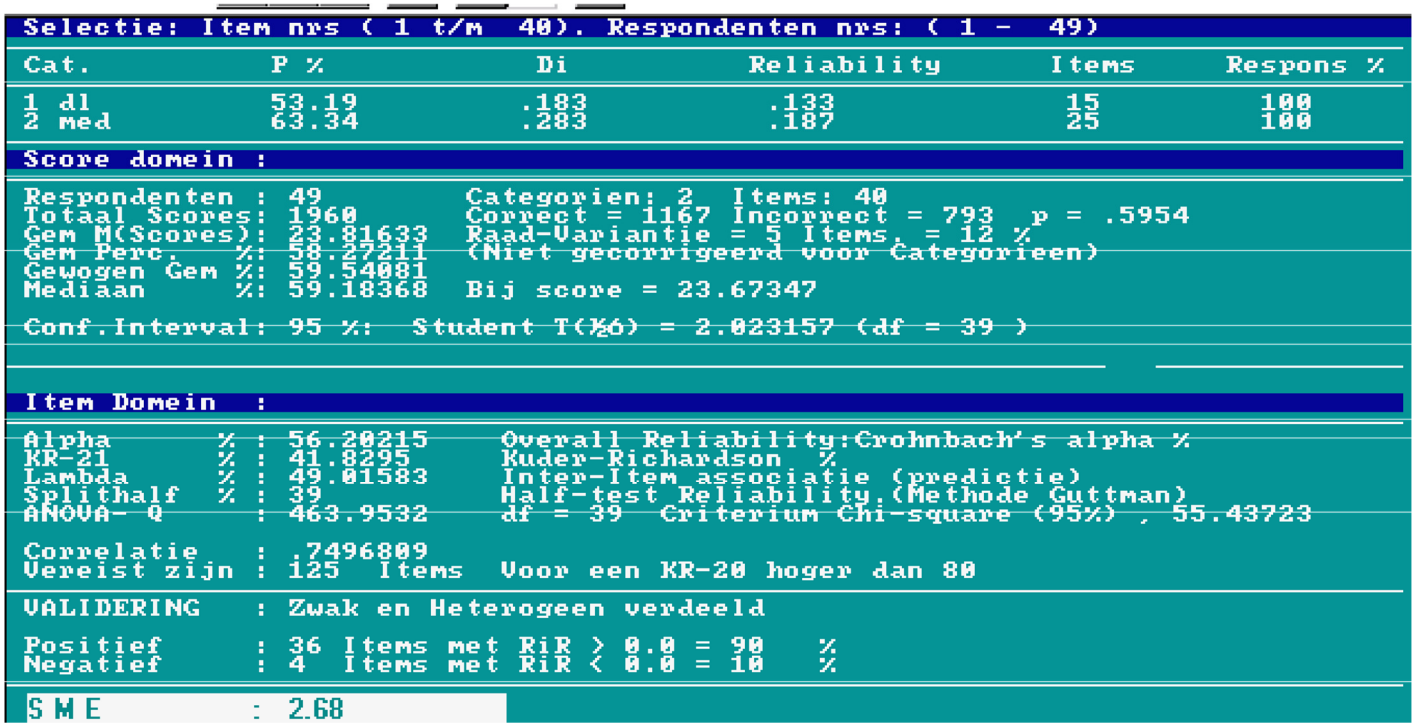
De Cronbach´s alfa en de KR-20 zijn indicatoren voor de kwaliteit van de toets als geheel. Deze indicatoren zijn met een computerprogramma makkelijk te bepalen (Zie Algoritme)

De genoemde maten voor de betrouwbaarheidsschatting zijn echter NIET stabiel in de tijd, en dat betekent dat een eenmaal vastgestelde betrouwbaarheid van een toets niet als een vast gegeven mag worden beschouwd.

TOP

BEOORDELING van de KWALITEIT

Het belangrijkste kriterium voor een goede kwaliteit van een toets behoort te zijn : de toets selecteert de kandidaten op hun competentie, vaardigheid of kennis. Dit ongeacht of de kandidaten een toets nu makkelijk of prettig vinden of niet en ongeacht of de toets een hoge slaagkans heeft of niet. Uitsluitend zij die de stof beheersen behoren te slagen voor een toets.
Hoewel een toets meestal nooit een representant is van het totale kennisdomein, betekent dit wel dat de toets als meetinstrument betrouwbaar en valide moet zijn. Maar, bij een MC-toets is in ieder geval de objectiviteit van de scoring een garantie.
Om de kwaliteit te beoordelen moeten indicatoren worden vastgesteld (bv.gebaseerd op MC-items) en hierover kan uitsluitend iets zinnigs worden gezegd met computer-ondersteunde analyse. Wat men met een dergelijke analyse kan bereiken is dus een betrouwbare indicatie, het behoort geen doorslaggevende beslissing te zijn. Men kan aantonen dat er altijd, zelfs bij een toets met hoge betrouwbaarheid, kandidaten ten onrechte zullen zakken en tenonrechte zullen slagen.
Een toets men een "iets meer dan gemiddelde score",(bv met een overall P-waarde tussen 55% - 65 %) zal echter het beste differentieren tussen kandidaten met een hoge eindscore en kandidaten met een lage score.
Hieronder enige weergaven van indicatoren (parameters) met betrekking tot kwaliteitsbeoordeling van een MC toets met bepaalde vragen uit het domein "Medisch,MED" en "Dieetleer,DL" behorende tot eenzelfde toetsaanbod.
[De berekening van deze parameters zijn uitgevoerd met een automatisch item-toets-analyse software-pakket dat reeds operationeel was in 1980].



```
Selectie: Item nrs ( 1 t/m  40). Respondenten nrs: ( 1 -  49)

Cat.          P %            Di          Reliability      Items      Respons %
1 dl         53.19          .183          .133             15          100
2 med        63.34          .283          .187             25          100

Score domein :

Respondenten : 49          Categorien: 2  Items: 40
Totaal Scores: 1960        Correct = 1167 Incorrect = 793   p = .5954
Gem M(Scores): 23.81633    Raad-Variantie = 5 Items, = 12 %
Gem Perc.   %: 58.27211    (Niet gecorrigeerd voor Categorieen)
Gewogen Gem %: 59.54081
Mediaan     %: 59.18368    Bij score = 23.67347

Conf.Interval: 95 %:  Student T(95) = 2.023157 (df = 39 )

Item Domein  :

Alpha       % : 56.20215    Overall Reliability:Crohnbach's alpha %
KR-21       % : 41.8295     Kuder-Richardson  %
Lambda      % : 49.01583    Inter-Item associatie (predictie)
Splithalf   % : 39           Half-test Reliability.(Methode Guttman)
ANOVA- Q      : 463.9532    df = 39  Criterium Chi-square (95%) , 55.43723

Correlatie    : .7496809
Vereist zijn  : 125  Items  Voor een KR-20 hoger dan 80

VALIDERING    : Zwak en Heterogeen verdeeld

Positief      : 36 Items met RiR > 0.0 = 90   %
Negatief      : 4  Items met RiR < 0.0 = 10   %

SME           : 2.68
```

Uit de hier weergegeven parameters blijkt dat de betrouwbaarheid van deze toets slecht is : De KR20 (hier aangeduid als alpha) is 56.2% en dus lager dan de minimaal toegestane betrouwbaarheid van 65 %. Uit de SPLIT-HALF test blijkt dat er tenminste 125 items aangeboden hadden moeten worden voor een goede betrouwbaarheid ! In bovenstaande situatie is bij 40 MC-items het overall toetsgemiddelde X = 23.81 (ruim 58 %) en blijkt de score-standaard meetfout (SME) = 2.68. Een 95% interval van 2x de SME is dan 5.36. Dit betekent dat kandidaten met een score tussen [18.45 en 29.17] statistisch bezien voor maximaal 95 % zekerheid behoren tot dit toetsgemiddelde X. De ANOVA Q test (Cochran Q, zie hierna) geeft hier aan een kritische waarde van 463.95 en deze is significant hoger dan het kriterium (55.437, bijbehorende p-value < 0.005) onder de geteste omstandigheden, zodat men er van uit mag gaan dat de data-matrix van de beantwoordingen niet op grond van toeval (willekeurig]) tot stand gekomen is.
Voor kandidaten met een score in dit gebied lager dan het gemiddelde betekent dit nogal wat, immers mag men deze dan nog wel laten slagen ?

Parameters met betrekking tot kwaliteitsbeoordeling van MC-items (4 alternatieven):

ITEM Domain 49 Cases 40 Items 2 Categories

Item	P(Mean)	Di	Rir	[KEY]	Item Validity
med					
16	.6326531	.3333	.2997	[4 d]	
17	.4489796	.5	.2744	[4 d]	
18	.4081633	.25	.0537	[1 a]	
19	.8367347	.3333	.4731	[3 c]	
20	.877551	.4166	.3142	[2 b]	
21	.6530612	.25	.0957	[3 c]	
22	.7755102	0	.0416	[2 b]	
23	.7346939	.1666	.0211	[1 a]	
24	.9183673	.25	.2784	[3 c]	
25	.7346939	.4166	.2022	[2 b]	
26	.1020408	-.0834	-.194	[4 d]	Bad
27	.755102	.4166	.3289	[3 c]	
28	.755102	.5833	.4446	[1 a]	
29	.5714286	.1666	.0745	[4 d]	
30	.5510204	0	-.1126	[3 c]	Bad
31	.2857143	.0833	.062	[2 b]	
32	.3673469	.1666	.0302	[3 c]	
33	.8163266	.5833	.4684	[3 c]	
34	.5918368	.5833	.3295	[2 b]	
35	.8979592	.1666	.3218	[1 a]	
36	.7346939	.4166	.1424	[1 a]	
37	.244898	.1666	.0597	[4 d]	
38	.4693878	.3333	.0514	[3 c]	
39	.877551	.3333	.3142	[2 b]	
40	.7959183	.25	.3033	[2 b]	

Score max= .9591837 Score min= .1020408
Rir max= .5476067 Rir min=-.1940444

Parameters met betrekking tot de scoring van 25 MC-items, [MED = "medische vragen", e = 0 = niet ingevuld door de kandidaat].
Key = P-waarde. Alternatieven [a b c d]weergegeven met hun frequentie.Aantal domein-items = 25. Aantal kandidaten N = 49.

Keuze Item	[1] a fr	[2] b fr	[3] c fr	[4] d fr	[5] e fr	GEM.waardering (a/e)	[KEY]
16	11	0	7	31	0	3.18	[d]
17	3	15	9	22	0	3.02	[d]
18	20	1	27	1	0	2.18	[a]
19	2	2	41	4	0	2.95	[c]
20	3	43	0	3	0	2.06	[b]
21	9	7	32	1	0	2.51	[c]
22	4	38	6	1	0	2.08	[b]
23	36	1	10	2	0	1.55	[a]
24	3	1	45	0	0	2.85	[c]
25	12	36	0	1	0	1.79	[b]
26	0	5	39	5	0	3	[d]
27	0	3	37	9	0	3.12	[c]
28	37	2	7	3	0	1.51	[a]
29	3	12	6	28	0	3.20	[d]
30	16	2	27	4	0	2.38	[c]
31	19	14	9	7	0	2.08	[b]
32	16	8	18	7	0	2.32	[c]
33	2	6	40	1	0	2.81	[c]
34	7	29	9	4	0	2.20	[b]
35	44	4	1	0	0	1.12	[a]
36	36	1	8	4	0	1.59	[a]
37	3	32	2	12	0	2.46	[d]
38	9	6	23	11	0	2.73	[c]
39	2	43	2	2	0	2.08	[b]
40	9	39	1	0	0	1.83	[b]

Cat.med :

De Item-respons (zie ook hierna):

Zo valt uit de gegevens in de bovenstaande tabellen met indicatoren het een en ander te zeggen over de kwaliteit van de vragen naar aanleiding van de respons van de 49 kandidaten op de items:

Bijvoorbeeld: Vraag 24, KEY = c, P-waarde = 91.8 [%]. Deze vraag discrimineert goed [Rir=0.27] maar toch is dit een makkelijke vraag en de alternatieven worden nauwelijks gekozen, want

de [a-waarde %] voor alternatief a =
6.1 de [a-waarde %] voor alternatief
b = 2.9
de [a-waarde %] voor alternatief d = 0

Bijvoorbeeld: Vraag 31, KEY = b, P-waarde = 28.6 [%]. Deze vraag discrimineert NIET [Rir=0.062]. Het is een redelijk moeilijke vraag maar de keuzes voor de alternatieven zijn goed gespreid, want de [a-waarde %] voor alternatief a = 38.8 de [a-waarde %] voor alternatief c = 18.4 de [a-waarde %] voor alternatief d = 14.3

TOP

Vergelijking tussen Betrouwbaarheid,Validiteit en Objectiviteit.

Het is een misverstand om te denken dat met MC vragen uitsluitend feiten(kennis) kan worden getoetst. Uit vergelijkend onderzoek en opgedane ervaringen (bv volgens de CITO, het Theoretisch rijexamen, vergelijkende academische studies, marktanalyses etc.), blijkt dat het construeren van diverse (alternatieve) vraagformuleringen om een MC vraag te maken zeer veelzijdig is, als men maar creatief genoeg is.

De denkfout is dat men soms meent dat het construeren van MC vragen gemakkelijk zou zijn en meestal heeft men ervaring opgedaan met slechte voorbeelden in het verleden. Dat wil natuurlijk niet zeggen dat uitsluitend MC toetsing zaligmakend is. Echter, met een MC toets kan men alle graderingen van cognitieve niveaus onderzoeken, zie tabel A hieronder, maar dat is voor de docent een arbeidsintensieve klus. En dit nog afgezien van de vraag of het wenselijk is om zich uitsluitend te verlaten op gebruik van MC toetsen. Recente toepassingen in diverse toets-software applicaties en elektronische leeromgevingen (onder meer met QuestionMark/Perception, WebCT, Blackboard) laten duidelijk zien dat met diverse vraagtypen veel aantrekkelijke toetsen zijn samen te stellen, hoewel dit meestal wordt toegepast voor diagnostische- of zelftoetsing.

Met betrekking tot bovenstaande genoemde indicatoren kan men een vergelijking maken tussen de betrouwbaarheid, de validiteit en de objectiviteit gerelateerd aan een aantal toetsvormen.

In het algemeen (onafhankelijk van het type meting of het type toetsvorm) heeft de betrouwbaarheid van een beoordeling te maken met de precisie van de meting.
Hoe preciezer, des te minder ruis en spreiding tussen de metingen bestaat.
Hoe betrouwbaarder, des te meer zekerheid er bestaat over uitspraken op grond van de toets-uitslag. Met een reeks betrouwbare metingen "treft" men steeds hetzelfde deel van een te toetsen kennisgebied.

De betrouwbaarheid van een toets is ideaal indien de Cronbach´s alpha of KR-20 >= 80 %.

De validiteit van een beoordeling heeft te maken met het doel van de meting. Validiteit is echter sterk afhankelijk van de mate waarin een toets aansluit op het te toetsen kennisdomein, de doelen etc.
Een hoge validiteit garandeert op zich niet dat er ook geen "ruis" in de metingen voorkomt. Daarom moeten betrouwbaarheid en validiteit tesamen worden beschouwd.

° Een niet betrouwbare toets heeft een grote spreiding tussen de metingen.
° Een betrouwbare toets "treft" steeds een zelfde gedeelte van het te toetsen domein.
° Een niet valide toets meet niet wat de bedoeling is.
° Een valide toets meet wat (gemiddeld) de bedoeling is maar tussen de metingen kan veel ruis optreden. ° Een betrouwbare EN valide toets heeft weinig ruis, meet wat het moet meten.

De objectiviteit heeft te maken met de kans op de zuiverheid (het waarheidsgehalte) van de meting. Een groot voordeel van een MC toets is de hoge mate van objectiviteit.

In onderstaande tabel A is een vergelijking gemaakt tussen betrouwbaarheid, validiteit en objectiviteit bij een aantal toetsvormen afgemeten naar de verschillende (cognitieve) niveaus bij toetsing (++ zeer hoog, + hoog, ± redelijk, - laag) al of niet gebruikmakend van extra multimediale middelen zoals grafische afbeeldingen, geluid etc. Men kan uit deze tabel concluderen dat, mits voldaan wordt aan de eisen van Validiteit en Betrouwbaarheid MC vragen een goed middel zijn om diverse aspecten van cognitieve niveaus te toetsen.

TABEL A : Betrouwbaarheid,Validiteit en Objectiviteit bij diverse toetsvormen.

	Feiten kennis	Begrips vragen	Toepassings vragen	Probleemoplossend analyserend	Redenerend Vergelijkend	Belasting voor de student	Betrouwbaar heid	Validiteit	Objectiviteit
MV vragen (met meer dan 2 alternatieven)	+	+	+	+	+	-	+	+	++
Casus met korte open vragen	+	+	+	+	+	±	±	±	±
Essays	±	+	+	+	+	+	±	-	-
Mondelingen	±	+	+	+	+	++	±	±	±

In tabel B wordt een verband getoond tussen de betrouwbaarheid (Cronbach´s alpha of KR-20) in relatie tot de afnametijd van een bepaalde toetsvorm (gerelateerd aan eenzelfde te toetsen kennisdomein).

TABEL B: Betrouwbaarheid in relatie tot het soort toets en de duur van de toets.

Afnametijd (uren)	MC	Casus met korte open vragen	Mondeling
1	0.65	0.45 - 0.56	0.45
2	0.76	0.57 - 0.73	0.45 - 0.50
4	0.88	0.74- 0.84	0.50 - 0.55
8	0.95	0.85 - 0.92	0.50 - 0.55

Uitsluitend bij gebruikmaking van MC vragen kan de betrouwbaarheid exact worden vastgesteld en deze blijkt aanzienlijk hoger in vergelijking tot het aanbieden van andere toetsvormen zoals Open vragen of een mondelinge toetsafname. De tijdsduur is wel van invloed bij MC vragen maar nagenoeg niet bij een mondelinge afname. Dat laatste kan een voordeel zijn omdat men sneller (of directer) dan bij een MC toets kan beoordelen of een kandidaat in staat is geweest zich bepaalde kennis over het te toetsen onderwerp eigen te maken.
Met nadruk wordt er op gewezen dat de betrouwbaarheids coefficient (Cronbach´s alpha, KR-20) geen stabiele maat is in de tijd. Dit tengevolge van mogelijke systematische veranderingen in de te testen doelgroep.

TOP

De score-matrix en de Cochtran Q test

De Cochran Q test (of ANOVA Q) is een nonparametrische statistische analyse die dient om na te gaan in hoeverre de responsen in een test-meet systeem (on-)afhankelijk tot stand is gekomen : [Nul-hypothese: de responsen zijn homogeen verdeeld: dwz verdeeld op grond van het toeval]. Dit is uit te beelden in een data-matrix, score-matrix of respons-matrix, zie hieronder.

N items (j=1...N) en hun -scores ---- Þ

Item

Nr[1,2,3,4,5,6,7,8,9,..,..,..,J.,...,...,N]|

1)[1,0,1,1,0,0,1,1,1,1,0,0,0,0,1,1,...N]|
2)[0,1,1,1,1,0,1,0,0,1,1,1,1,0,1,...N]|
3)[0,1,1,0,1,1,0,0,1,1,0,1,0,0,1,...N]| ß P kandidaten (i=1...P)
4)[1,1,0,0,0,0,1,1,1,0,1,1,1,1,1,...N]|
5)[0,1,0,0,0,1,1,1,1,0,0,0,0,0,0,...N]|
i) [.....N]|

P)[.....N]|

Gebruikt men bij een MC-toets met (2,3 of meer) alternatieven, een "gedwongen raden" of dichotome items (1 = goed, 0 = fout) dan ontstaat een 2-dimensionale score-matrix [P,N] met een bepaald patroon van 1,1,0,1 etc **gebaseerd op een binomiale verdeling. Uit deze data-matrix of score-matrix is veel vast te stellen**.

Indien op de score-matrix een zogenoemde Cochran Q analyse (ANOVA variant) wordt uitgevoerd en de p-waarde van deze analyse hierbij lager is dan p =< 0.01 dan betekent dit dat de responsen niet volgens het toeval "goed" zijn gescoord en derhalve de score-matrix heterogeen is samengesteld. De beantwoording van de personen in de steekproef is statistisch bezien niet willekeurig en men mag er dan van uit gaan dat "kennis" over het te testen onderwerp aanwezig is bij de proefpersonen (het patroon van de score-matrix is dan afhankelijk van de mate van kennis).

Rit en Rir : Item Correlatie

Bovenstaande score-matrix wordt ook gebruikt om te bepalen hoe een item (met de responsen 1,0,0,1,0.. etc) correleert met de responsen van overige items van de totale matrix. Dat wordt uitgedrukt met de Rir of Rit waarde (R = correlatie symbool). De Rit -waarde berekent de (lineaire) correlatie binnen de score-matrix inclusief het beschouwde item zelf. De Rir-waarde fixeert het betreffende item en berekent de correlatie op de rest van de score-matrix zonder dat item.

De Rir is een relatieve correlatie-maat tussen een bepaald item en alle overige items in het testmeetsysteem. De Rir waarde wordt vooral berekend voor een nog niet goed gevalideerd testmeetsysteem. Hoe hoger de RIR waarde, des te beter "past" dit item in het totale test-meetsysteem. Een Rir waarde >= 0.45 wordt aanbevolen voor een goed item. In de meeste testsituaties waarbij men zeker weet dat sprake is van een

valide test-meetsysteem wordt volstaan met de berekening van de zogenaamde Discriminatie waarde index, DI in plaats van de Rit- of Rir-waarde.

Dat wil dus zeggen, een item met een Rit waarde >= 0.45 discrimineert uitstekend tussen een goede score-uitslag (indien kennis aanwezig is) en een slechte score-uitslag indien kennis over het te testen onderwerp NIET aanwezig is bij de kandidaat. Een Rir =< 0 betekent dat men van dit item zeker NIET op aan kan: Willekeurige personen ZONDER kennis van zaken scoren dit item juist goed terwijl geoefende personen MET kennis van zaken dit item juist niet goed scoren. Een dergelijke toestand dient te worden vermeden ! Items met een zeer lage of negatieve Rir waarde dienen uit het test-meetsysteem te worden verwijderd. Ofwel de vraagformulering was niet goed geweest, ofwel de vraag heeft weinig of niets te maken met het te testen onderwerp.

Om een Rir waarde betrouwbaar vast te stellen moet men tenminste 20 items behorende tot hetzelfde kennisdomein onderzoeken, bij tenminste 30 personen.

. . . . . .

TIPS voor een bruikbaar MC toetsontwerp (de constructie).

Een bruikbaar MC toetsontwerp voldoet onder meer aan:

1. Diversiteit (het vragenaanbod meet meerdere cognitieve niveaus).
2. Discriminerend (de vragen kunnen onderscheiden in "goede" en "slechte" kandidaten) => Rit waarden.
3. Homogeniteit (alle vragen hebben te maken met de kern van de leerstof).
4. Betrouwbaarheid (de toets als geheel bestrijkt de leerstof zo precies mogelijk, alle items correleren met de totale toets, dus consistent)=> KR-20 waarde. 5. Representativiteit (de toets is een afspiegeling van de leerstof).
6. Validiteit (de toetsconstructie meet wat de bedoeling is : niets meer en niets minder) = interne validiteit.
7. Specificiteit (de vraag inhoud is gericht op de leerdoelen en "dekt de lading" , de belangrijkste kenmerken of attributen worden bestreken).
8. Sensitiviteit (gericht op de juiste doelgroep en die kennis van zaken heeft).
9. Eenduidigheid (niets is voor meerdere uitleg vatbaar, in dezelfde situatie is de vraagformulering voor de alternatieven identiek, dus consequent).
10. Persuativiteit (overtuigendheid, hoog afleidend vermogen van de alternatieven) => De awaarden van een item zijn >= 0.2
11. Gebalanceerdheid (er is een evenwicht tussen makkelijke en moeilijke vragen. Plaats niet alle moeilijke vragen aan het einde van de toets) => Pi waarden.

12. Proportionaliteit (Realistisch evenwicht tussen toetslengte en toetsafname tijd. De gemiddele tijd om een MC toetsvraag te beantwoorden wordt geschat op 2 á 3 minuten).

° Het is meestal niet vanzelfsprekend dat een bepaalde toets ook in een volgende situatie zonder wijzigingen gebruikt kan worden (niet generaliseerbaar). Dit hangt af van de betrouwbaarheid, de itemcorrelaties, de te meten leerstof (doelen) en de samenstelling van de doelgroep (kandidaten). Bovendien, de toets als geheel zou eigenlijk "geijkt" moeten worden met een "gouden standaard" en deze is meestal niet voorhanden. De externe validiteit van een toets is moeilijk vast te stellen.

° Voer een itemanalyse uit. Verwijder zeer slecht correlerende items en doe de itemanalyse opnieuw.

° Corrigeer alleen dan voor de gisfactor (gokkans) als de toets als geheel een redelijk hoge KR-20 heeft dwz tenminste 70 %, anders niet.

° Geef feedback: niet alleen ten aanzien van de vraaginhoud, maar ook ten aanzien van de validiteit van een toets en de beslissingen die hieruit voortvloeien voor de beoordeling.

. . . . . .

De cesuur:

De toetscesuur (bv iedereen met een toetsscore gelijk aan of hoger dan 50 % of 60 % is geslaagd) kan een richtlijn zijn om vast te leggen wie onvoldoende of voldoende krijgt toebedeeld voor een toets. Soms geeft men als cesuur (norm) voor slagen de waarde P=0.55 (55%) voor de gehele toets. En wat men daarmee aangeeft is dat men in dat geval 55 % van de aangeboden kennis beheerst ! Dit is dus niet identiek met de situatie dat iemand 55 % van de vragen goed heeft gescoord.Hieronder een methode om de cesuur te berekenen.

Vaststellen van de cesuur voor MC vragen met 4 alternatieven. Raadkans = 25 % (¼ of 0.25)

Stel : 50 % van de kandidaten moet voldoende hebben en 50 % onvoldoende bij een maximale score van 100 %.
Cesuur formule voor vaststellen van de ondergrens "Voldoende":
50 % x (Maximale score – raadkans)
100% – [0,5 x (100 – 25)] = 62.5 % als score ondergrens voor voldoende.

Men kan de bovenstaande formule herleiden tot: [maximale score + raadkans] / 2.
Bij een toets met MC-items met 4 alternatieven wordt dit :[1 + 0,25] / 2= 0.625 = 62,5 %
Bij een toets met MC-items met 5 alternatieven wordt dit :[1 + 0,20] / 2= 0.60 = 60 % Bij een toets met MC-items met 2 alternatieven wordt dit :[1 + 0,50] / 2= 0.75 = 75 %

De cesuur kan worden verlaagd bij zeer veel onvoldoendes (bijvoorbeeld bij 40% of meer) totdat 60 % van de kandidaten voldoende heeft.Bij een toets met MC-items met 4 alternatieven wordt dit:
100% – [0,6 x (100 – 25)] = 55 % als score ondergrens voor voldoende.

Berekening bij uitval van items:

Bij de eindbeoordeling moet men soms ook rekening houden met slechte items die worden geelimineerd naar aanleiding van een eerste analyse.
Stel: Geëlimineerd moeten worden 3 slechte items uit een totaal van 40 items en men berekent het cijfer op een schaal van 0 - 10:

Resteren 40-3 = 37 totaal geldige items.
Gecorrigeerde score voor de raadkans = ¼ x 37 = 9,25 (= raadkansscore)

$$\text{Cijfer} = \frac{[\text{gemeten score} - \text{raadkansscore}] \times 10}{(\text{totaal} - \text{raadkans})}$$

$$\text{Cijfer} = \frac{[\text{gemeten score} - 9{,}25] \times 10}{(37 - 9{,}25)} = 27{,}5$$

Hieronder een tabel waarmee men rekening kan houden om voor de raadkans te corrigeren:

N = totaal aantal items
G = Totaal GOED (CORRECT) beantwoord (gemeten)
W = aantal items correct gescoord EN geweten, gecorrigeerd (dwz. verminderd met de kans op gokken).
1/k (N) = aantal dat men uit N items goed gokt = raadkansscore

Voor 40 items (MC-4 alternatieven) geldt voor de raadkans: Het theoretisch aantal items op goed gokken = 10

De correctie formule die rekening houdt met deze raadkans is: W = (G – 10) ¾

G	W	Gegokt			
40	40	0			
38	37	1	36	35	1
34	32	2	32	29	3 30
27	3				
28	24	4			
26	21	5	24	19	5
22	16	6			
20	13	7	18	11	7
16	8	8			
14	5	9			
12	3	9			
10	0	10			

Rekeninghoudend met een cesuur van 62,5 % (de ondergrens voor tenminste goed) zou een voldoende gegeven kunnen worden indien een kandidaat een gemeten eindscore G heeft van 29 of meer items correct.
(W = 25), vooropgesteld dat geen enkel slecht item geëlimineerd behoeft te worden !

. . . . . .

:

Testen met SPSS

Met SPSS kan men makkelijk een item analyse uitvoeren indien men gebruikt maakt van een discrete binomiale situatie: Elk item moet als een aparte variabele (kolom) zijn ingevuld in het SPSS spreadsheet en men onderscheidt per CASE (proefpersoon,kandidaat) de volgende (numerieke) notatie:

1 = correct beantwoord

0 = niet correct beantwoord.

Voorbeeld score-matrix:

	Item1	Item2	Item3
Case 1	1	0	1
Case 2	0	0	1
Case 3	1	1	0

SPSS Procedure Item analyse:

Analysis (of Statistics) -> Scale -> Reliability Analysis.

Breng vervolgens alle variabelen (Item1,Item2,Item3..) over in het dialoogvenster onder Items. Vervolg daarna de SPSS de procedure met te klikken op Model.

LINK http://www.statsoftinc.com/textbook/streliab.html

TOP

GUTTMANschaal

Een andere manier om de reliability van items na te gaan wordt soms uitgevoerd door gebruik te maken van de zg. Guttman schaal.

Zie voor details over de Guttmanschaal (en andere schaaltypen) de volgende LINK: Knowledge Base Bill Trochim. Scaling Social Research Methods **http://socialresearchmethods.net/kb/**

Bij de Guttmanschaal wordt gebruik gemaakt van een deterministisch schaalmodel voor een dichotome items : Ja-Nee vragen.

Het idee is dat de responsen op de items een bepaald triangulair patroon kunnen vormen waarbij het mogelijk is dat indien een bepaald persoon op een eerste vraag JA antwoord, dat deze op de volgvragen eveneens met JA zal antwoorden.

Indien een andere persoon de eerste vraag NIET met JA beantwoord maar alle volg vragen wel, bestaat er toch kans dat deze volgvragen eveneens met JA zullen worden beantwoord, etc. Dit kan visueel worden bereikt door de positieve responsen cumulatief in een tabel uit te zetten.(zie de tabel hierna).
Onderstaande tabel C geeft hiervan een beeld (trapvorm):

TABEL C GUTTMAN triangulair patroon bij JA / NEE vragen:1 = JA respons.

ITEM	PERSOON 1	PERSOON 2	PERSOON 3	PERSOON 4	PERSOON 5	PERSOON 6
a	1					
b	1	1				
c	1	1	1			
d	1	1	1	1		
e	1	1	1	1	1	
F	1	1	1	1	1	1

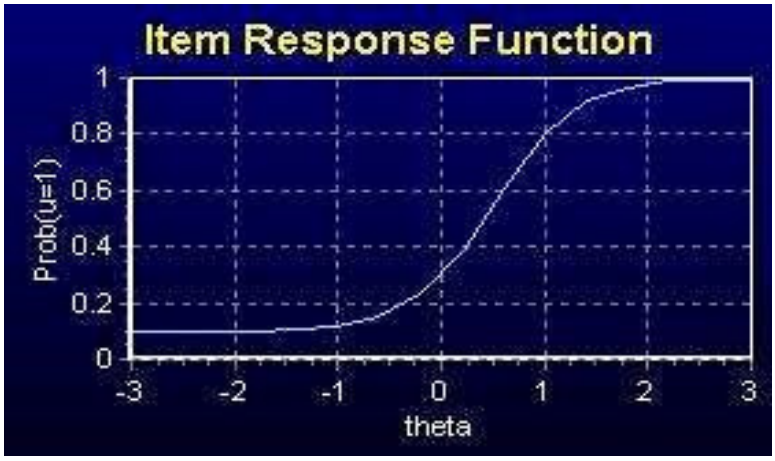
Wat men hierbij moet doen is de items voorleggen aan een groep personen en vervolgens de verkregen responsen cumulatief rangschikken volgens het patroon in de bovenstaande tabel. Het is echter zeer onwaarschijnlijk dat men in werkelijkheid een dergelijke fraaie trapvorm kan verkrijgen. Algemeen wordt echter aangenomen dat een dergelijke Guttman verdeling niet erg realistisch is. Het blijkt erg moeilijk een dergelijk patroon te kunnen reproduceren bij herhaling.

. . . . . .

TOP ITEM RESPONS

De Item Respons Theorie (IRT) legt de focus niet op de validiteit van de vraag zoals bij de klassieke item-analyse (zie hierboven) maar houdt zich vooral bezig met vragen als
(1) "Hoe waarschijnlijk is het dat de itemrespons een betrouwbare weergave is van de competentie van de tester (kandidaat)"?
(2) "Hoe kan de competentie zo nauwkeurig mogelijk worden geschat ?"

Het model houdt rekening met de (theoretische) gokkans, de discrimineerbaarheid en de moeilijkheidsgraad van een item en legt hiermee een mathematisch verband tussen de werkelijke score en de geobserveerde score van een item of een set items. Het doel is om te bepalen op welke punt van de competentie-schaal de kandidaat zich bevindt. Hiermee kan een kandidaat een zekere "ability score" verkrijgen die afleesbaar is op een zogenoemde Item-karakteristiek-curve. De IRT beschrijft dit met mathematische- / logistische modellen.



Curve:Item Respons Karakteristiek

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta - b_j)}}$$

Kansfunctie:Item Respons Logistiek Model

P(θ)= de kans (waarschijnlijkheid) dat een kandidaat met een bepaalde competentie "Theta" of ookwel de proportie kandidaten die competent zijn om het item correct te beantwoorden (θ) het item j correct beantwoordt

b_j =de moeilijkheidsgraad van het item (P-waarde) a_j =de discriminatie-index van het item (DI waarde,Rir- of Rit-waarde) c_j =de gokkans (bij een 4-MC item is deze 0.25) e = 2,718281 θ= "competentie"- (ability)-factor, volgens de standaard-normaalverdeling varierend van -3 .. +3

In de praktijk kan de IRT zeer goed worden toegepast bij computer adaptieve testen (CAT) bv. voor onderwijskundige doeleinden.

[Link 1: ITEM RESPONS THEORIE en CAT]
[Link 2: Item Response Theory (*pdf documents)]

[Link 3: IRT Overzicht]
[Link 4: CITO: Computer gestuurde en Adaptieve toetsing]

TOP

MOKKEN schaal

Robert Mokken (UvA) heeft een Nonparametrisch schalingsmodel ontwikkeld voor analyse van (dichotome) responsen.Dit model wordt (onder meer) toegepast bij vraagconstructies ten behoeve van de sociale en gedragswetenschappen.

MSP, Mokken Scaling for Polytomous items, offers scaling facilities for the cumulative nonparametric item response theory developed since 1971 by Mokken and other Dutch researchers. Measurement of latent attributes occurs in all social and behavioral sciences. Such scales are often used to measure properties like academic achievement, personality traits like extraversion, or personal attitudes about political, moral or social issues, or a patient's views on quality of life. From the responses of persons to a number of items or questions indicative of the latent trait, scale scores for the persons are obtained and the quality of the measurement instrument is evaluated. MSP is a simple and flexible tool for this task, widely applicable because it is based on mild statistical assumptions.

Het software pakket MSP5 voor windows is te downloaden via de volgende weblink Item Response Theory Analysis, Rijks Universiteit Groningen

TOP

Algoritme ten behoeve van de berekening van de KR-20 voor een willekeurig toetsdomein, inclusief een Studenten-Cohort, toets-item-sleutel en 4 MC-alternatieven, met een eenvoudig computerprogramma:
Let op het aantal verschillende verzamelingen (dimensions) dat nodig is.

```
//: Algorithm RELIABILITY      Warnar Moll 1995
Item,Category,Key,ID,Answer: integer;
Cohort,Participant,Category: string;
vn      = total items
k       = total keys
p       = total participants
vk      = total within domains [ categories ]
som     = total (SCORE SUM)
//: ARRAYS
REM:N$(i)       = dimension of COHORT
REM:L$(i)       = dimension of PARTICIPANTS
REM:V$(k)       = dimension of DOMAINS
REM:ID(i)       = dimension of ID numbers (account)
REM:m(i,4)      = dimensions of ANSWERS (MC-4)
REM:k%(j)       = dimension of KEYS (CORRECT)
REM:a%(i,j)     = dimensions of answers of all participants
REM:p(j)        = dimension of correct answers,  item-p-values
REM:S%(i)       = dimension of partial item-score-sum correct answers
REM:va(k)       = dimension of categories
REM:v%(i,k)     = dimensions of correct item-scores of –within- categories
FOR j = 1 TO vn
        pw = 0
                FOR i = 1 TO p
                        IF k%(j) = a%(i, j) THEN s%(i) = s%(i) + 1: p(j) = p(j) + 1
                                FOR k = 1 TO vk
                                        FOR j = 1 + va(k - 1) TO va(k - 1) + a(k)
                                                IF k%(j) = a%(i, j) THEN z = z + 1
                                        NEXT
                                        v%(i, k) = z: z = 0
                                NEXT
                NEXT
        pw = p(j) / p
NEXT
REM:kr20        = KR-20 reliability
REM:kr21        = KR-21,reliability (under-estimation of KR-20)
REM:smf         = 1X standard error of measurement (68 % of the total score distribution)
REM:som         = sum of all correct scores : participants
REM:gem         = mean testscore X :  Overall P-value
REM:sp          = sum of the product : correct scores  X incorrect scores: itemvariance
REM:var         = testscore variance

FOR i = 1 TO p
        som = som + s%(i)
NEXT

FOR j = 1 TO vn
        sp = sp + p(j) * (1 - p(j))
NEXT

gem     = som / p
vr      = square sum

FOR i = 1 TO p
        vr = vr + (s%(i) - gem) * (s%(i) - gem)
NEXT

var     = (1 / p) * vr
kr20    = ((vn / (vn - 1)) * (1 - sp / var))
kr21    = (vn / (vn - 1)) * (1 - ((gem - gem * gem / vn) / var))
smf     = SQR(var) * SQR(1 - kr)
```

Correlatie tussen de toetsbetrouwbaarheid en het aantal toetsvragen (items) dat per kennisdomein werd aangeboden.
Vraagtype: MC met 4 alternatieven. Slechts 1 keuze is correct.

De Kuder-Richardson Formule 20 (KR20) wordt onder meer gebruikt voor MCtoetsen als maat voor de correlatie tussen de gemiddelde item-score en (toets)meetbetrouwbaarheid (=Reliability).Men bepaalt hiermee de invloed van de meetfouten en derhalve de betrouwbaarheid en validiteit van de toets. Een MC-toets met een Kuder-Richardson waarde (KR20) < 0.65 is niet acceptabel. Dit punt (zie figuur) is in de regel gelokaliseerd bij een aanbod van maximaal 20 toetsitems per toets.

Kennisgebied: Medische Microbiologie/ Toxicologie / Biochemie / Statistiek & Methodologie.
(De Rir waarden per item varieerden van -0.14 tot + 0.58)

De gegevens zijn gebaseerd op in totaal 65 toetsperioden gedurende 15 jaar digitale toetsmeting en itemanalyse.
