Basiese Psigometrika

IPEN – opleidingsreeks nr. 1

Basiese Psigometrika

Redakteurs:

A.R. van den Berg
J.F. Vorster

Pretoria
Raad vir Geesteswetenskaplike Navorsing
1982

Die volgende skrywers het meegehelp om hierdie handboek tot stand te bring:

F.W.O. Heinichen, B.A., D.Ed., voormalige assistent-direkteur, IPEN
O.V. Kilpert, B.Sc., B.Ed., voormalige senior navorsingsbeampte, IPEN
Elizabeth M. Madge, M.A., senior hoofnavorsingsbeampte, IPEN
W.B.J. Prinsloo, B.A.(Hons.), assistent-direkteur, IPEN
A. Schoeman, B.Sc., M.A., D.Litt. et Phil., senior hoofnavorsingsbeampte, IKOMM
A.R. van den Berg, B.Sc., assistent-direkteur, IPEN
H.S. van der Walt, M.A., D.Phil., senior hoofnavorsingsbeampte, IPEN
J.F. Vorster, B.Sc., M.Ed., M.A., assistent-direkteur, IPEN

Instituut vir Psigologiese en Edumetriese Navorsing

Direkteur: Dr. J.D. van Staden

ISBN 0 86965 869 7

Prys: R9,85
(AVB ingesluit)

Gedruk deur V&R Drukkery, Pretoria

VOORWOORD

Die afgelope tiental jare is interne lesingreekse deur ervare navorsers oor aspekte van toetskonstruksie en toetsteorie, dit wil sê aspekte van die psigometrika, in die Instituut vir Psigologiese en Edumetriese Navorsing aangebied. Sommige van dié lesings word nou in verwerkte vorm saamgevat in hierdie handboek. Die handboek is dus primêr bedoel vir navorsers wat sielkundige of opvoedkundige toetse moet opstel, standaardiseer en valideer.

Na die inleidende hoofstuk volg die onderwerpe van klassifikasie, funksies en ontwerp van toetse in hoofstukke 2 en 3. Hoofstuk 4 handel oor die skepping en skryf van items. In hoofstuk 5 word riglyne en wenke gegee met betrekking tot die eksperimentele toepassing van items in grootskaalse toetsprogramme tydens standaardisering. In hoofstuk 6 volg metodes en tegnieke van itemontleding en -seleksie terwyl hoofstuk 7 handel oor normskale en normbepaling. In hoofstuk 8 word die klassieke toetsteoriemodel bespreek terwyl hoofstukke 9 en 10 handel oor betroubaarheid en veralgemeenbaarheid. Geldigheid van toetse word in hoofstuk 11 bespreek en itemresponsteorie in hoofstuk 12.

Daar is gepoog om die leser nie te oorlaai met lang wiskundige of statistiese afleidings nie. In hoofstukke 6 tot 12 word egter aanvaar dat die leser 'n kennis van hoërskoolwiskunde en eenvoudige statistiese begrippe sal hê. Daar bestaan talle goeie inleidende Statistiek handboeke wat geraadpleeg kan word in dié verband. Sommige daarvan is opgeneem in die literatuurverwysings aan die einde van hoofstukke 6 tot 12 (hoofstuk 6, nommer 8; hoofstuk 8, nommer 7; hoofstuk 10, nommer 2). Diegene wat dalk sou belangstel om meer gevorderde kennis oor 'n onderwerp te bekom, sal waarskynlik ook die literatuurverwysings nuttig vind.

Konstruktiewe kommentaar op die inhoud en formaat van die handboek sal deur die redakteurs en skrywers van die verskillende hoofstukke waardeer word.

J. D. van Staden

DIREKTEUR: IPEN

INHOUDSOPGAWE

HOOFSTUK 1
INLEIDING

1.1 AGTERGROND

Sedert die verskyning van die eerste volwaardige intelligensieskaal in 1904, naamlik dié van Alfred Binet, het die tegnologie van die meting van menslike vermoëns in die kort tydsbestek van 75 jaar met sulke rasse skredes vooruitgegaan dat daar nie nog 'n vertakking van die eksperimentele sielkunde is wat so 'n uitgebreide literatuur opgebou, so 'n groot weerklank en wye toepassing op feitlik alle terreine van die samelewing gevind het en so die belangstelling van sowel die leek as die ingeligte gaande gemaak het nie. Nie alleen het sielkundige meting 'n wye toepassing op die terrein van die onderwys en opvoeding gevind nie, maar is daar ook prakties gebruik gemaak van toetsing in die leër, in die nywerhede, ja in feitlik elke vertakking van die samelewing waar daar op een of ander wyse 'n beslissing in verband met die plasing, bevordering, keuring of sifting van persone gemaak moes word. Duisende toetse het reeds die lig gesien - sowel individuele as groeptoetse. Hieronder val toetse vir die meting van intelligensie, aanleg en bekwaamheid, toetse vir die bepaling van persoonlikheidseienskappe soos angs, aggressie, ensovoorts, prestasietoetse, houdingskale, vraelyste in verband met belangstelling en dergelike meer. Statistiese metodes is in so 'n mate in die tegnologie van sielkundige meting betrek dat daar met die verloop van jare 'n eie wetenskap naamlik die Psigometrika, ontwikkel het.

1.2 METING IN DIE FISIESE EN IN DIE GEESTESWETENSKAPPE

Meting bestaan uit reëls waarvolgens bepaalde kenmerke of eienskappe op so 'n wyse gekwantifiseer word dat dit as 'n hoeveelheid in die vorm van 'n getal uitgedruk kan word. In sommige gevalle is hierdie reëls voor die hand liggend soos die by bepaling van die afstand tussen twee dorpe. Waar dit egter om die meting van sielkundige eienskappe gaan soos intelligensie, belangstellings, houdings en so meer, is die reëls nie altyd so eenvoudig nie en is dit belangrik dat die "reëls" of die prosedures waarvolgens die kwantifisering plaasvind, eksplisiet geformuleer moet word. Meting in die fisiese wetenskappe verskil van meting in die geesteswetenskappe ten opsigte van onder andere die volgende:

1.2.1 In die fisiese wetenskappe kan 'n meting gewoonlik verskeie kere herhaal word sonder dat die resultaat enigsins deur die herhaling beïnvloed word. As die breedte van 'n tafel keer op keer bepaal word, sal slegs geringe variasies in die resultaat waargeneem word. Wanneer 'n siel= kundige meetinstrument egter verskeie kere op dieselfde leerling toe= gepas word, kan faktore soos vermoeidheid en die effek van oefening die resultate aansienlik beïnvloed.

1.2.2 Die tydsverloop tussen herhaalde toepassings speel in die fisiese wetenskappe dikwels geen rol nie. As 'n toets egter na 'n kort tussen= pose weer op 'n leerling toegepas word, kan faktore soos geheue en oefe= ning 'n groot rol speel. As die tydsverloop weer lank is, kan die ont= wikkeling, ryping of ervaring van die individu weer 'n groot rol speel.

1.2.3 In die fisiese wetenskappe word gewoonlik een enkele eienskap op 'n keer gemeet en bestudeer. As die afstand tussen twee plekke bepaal word dan is enkel en alleen *afstand* ter sake en speel faktore soos die breedte van die pad of die materiaal waarmee dit gebou is geen rol nie. Meting in die geesteswetenskappe is egter ingewikkelder. Wanneer 'n intelligen= sietoets op 'n leerling toegepas word, ontstaan die vraag ook of ons seker is dat dit wel intelligensie is wat gemeet word. Verdere vrae soos die volgende duik ook op:

Watter rol speel ervaring en watter oorerwing?
In watter mate is die resultaat beïnvloed deur angs, fisiese gesteldheid, geestelike versteurings, huislike omstandighede van die leerling, en so meer?

1.3 VLAK VAN METING

1.3.1 Die gebruik van getalle

Om kenmerke of trekke wat bestudeer moet word, te kan kwantifiseer, is dit 'n vereiste dat daar 'n homologie gevind moet word tussen sekere eien= skappe daarvan en dié van die getallesisteem wat gebruik word. Drie eienskappe van getalle wat die vlak van meting bepaal, is dié van iden= titeit, rangorde en optelling.

1.3.2 Indeling van meting in vier metingsvlakke

a. Nominale meting

Nominale meting gee die laagste vlak van meting en daar kan twee tipes onderskei word:

(i) 'n Blote nommer of etiket

Die meting bestaan bloot uit die toeken van 'n nommer aan 'n element of voorwerp vir identifikasiedoeleindes. Byvoorbeeld, die spelers in 'n voetbalspan word van nommers voorsien. Dit kan wees van 1 tot 15 of enige arbitrêre reeks nommers, sê 21 tot 35. Sinvolle berekeninge kan gewoonlik nie met hierdie nommers gemaak word nie. Dit is byvoorbeeld onsinnig om te beweer dat die 2de speler plus die 3de speler gelyk sal wees aan die 5de. Die doel is ook nie om berekeninge te kan maak nie. Die nommers word slegs gebruik om voorwerpe met bepaalde kenmerke te etiketteer.

(ii) Kategorieë

Hier behels meting die indeling van voorwerpe in groepe. Leerlinge kan byvoorbeeld in verskillende groepe of versamelings ingedeel word:

Groep 1: Al die leerlinge in Transvaal
Groep 2: Al die leerlinge in Natal
Groep 3: Al die leerlinge in die Kaapprovinsie

In die samelewing is daar eindelose voorbeelde van sulke versamelings, byvoorbeeld persone in verskillende professies, die twee geslagte, ensovoorts om maar twee te noem.

Albei hierdie metodes van meting is baie eenders. Die verskil is dat in laasgenoemde metode meer as een element in 'n versameling tuisgebring kan word en dat alle elemente met dieselfde nommer minstens een gemeenskaplike eienskap besit.

b. Ordinale meting

Ordinale meting bestaan daarin dat

(i) 'n versameling elemente gerangskik word van "meeste" tot "minste" met betrekking tot 'n bepaalde kenmerk;

(ii) daar geen aanduiding gegee word van "hoeveel" (in 'n absolute sin) van die bepaalde kenmerk by enigeen van die elemente in die versamelings aanwesig is nie;

(iii) geen aanduiding gegee word van hoeveel die elemente in die versameling van mekaar ten opsigte van die bepaalde kenmerk verskil nie. Wanneer dit slegs bekend is dat in 'n versameling sê $S_1 > S_2 > S_3 \ldots > S_n$ is ten opsigte van 'n bepaalde kenmerk, dan het ons met ordinale meting te doen. Wanneer 'n versameling seuns van die langste tot die kortste gerangskik word, word van ordinale meting gebruik gemaak. Geen informasie word met betrekking tot die gemiddelde hoogte verkry nie en ook nie van hoeveel elke seun van die ander in lengte verskil nie.

c. Intervalmeting

By intervalmeting is

(i) die elemente in 'n versameling volgens rangorde met betrekking tot 'n bepaalde kenmerk gerangskik;

(ii) die "hoeveelheid" bekend wat een element van 'n ander in dieselfde versameling ten opsigte van 'n bepaalde kenmerk verskil, en

(iii) geen informasie in verband met die absolute waarde van die bepaalde kenmerk by elke element bekend nie, dit wil sê, 'n absolute nulpunt bestaan nie.

Numeries gelyke verskille by intervalmeting verteenwoordig empiries gelyke verskille ten opsigte van die kenmerk wat by die elemente in die groep gemeet word. Twee elemente waaraan die syfers 5 en 10 toegeken is, is op die skaal net so ver uitmekaar as twee elemente waaraan die

syfers 15 en 20 toegeken is. Die "afstand" van A na B plus die afstand B na C sal gelyk wees aan die "afstand" van A na C.

d. Verhoudings- of ratiometing

By ratiometing is

(i) die rangorde van die elemente in 'n versameling met betrekking tot 'n bepaalde kenmerk bekend,

(ii) die intervalle tussen die elemente bekend, en

(iii) die afstand vanaf 'n vaste nulpunt vir ten minste een van die elemente bekend.

Hierdie nulpunt is 'n absolute nulpunt, naamlik die onderste grens waar die eienskap wat gemeet word, heeltemal verdwyn, met ander woorde nie meer of nie minder as absoluut niks is nie.

1.3.3 Vlak van meting in die sielkunde

Die bogemelde vlakke van meting word almal op een of ander wyse in sielkundige ondersoeke gebruik. Daar is ondersoeke waar slegs 'n syfer nodig is vir meting van 'n eienskap soos in die geval van pasiënte wat in sielkundige klinieke vir een of ander geestesversteuring geklassifiseer word. Persone moet dikwels op een of ander eienskap in rangorde geplaas word, soos wanneer proefpersone 'n aantal beroepe in volgorde van gewildheid moet plaas. Verhoudingmeting word veral gebruik in eksperimentele werk soos met ondersoeke na reaksietye op bepaalde stimuli of na die tyd wat dit 'n persoon neem om woordpare te memoriseer. Hier is dit sinvol om te praat van "niks" geleer nie, of een proefpersoon het twee keer soveel woordpare as 'n ander een gememoriseer.

'n Wanopvatting wat soms bestaan, is dat die gebruik van sekere statistiese prosedures sekere skaaleienskappe van die betrokke veranderlikes vereis. Gaito (1980) wys dat hierdie opvatting gebaseer is op die verwarring wat by sommige skrywers bestaan tussen metings- en statistiese teorie. Gaito (1980) stel dit soos volg: *For statistical tests of null hypothesis, as Lord stated, "the numbers do not know where they came from".*

1.3.4 Wiskundige transformasies

Dit is soms nodig dat metings op een of ander skaal en op 'n sekere vlak van meting om bepaalde redes wiskundig na 'n ander skaal getransformeer moet word. Onderhewig aan sekere voorwaardes is dit moontlik om vir al die vlakke van meting hierbo genoem transformasies van skaal deur te voer en nog dieselfde vlak van meting te behou. Sodanige transformasies gaan dikwels om 'n wysiging in die eenhede, die posisie van die nulpunte, die gemiddelde of die standaardafwyking. In die geval van nominale me= ting is daar oneindig baie moontlikhede vir transformasie aangesien enige syfer gebruik kan word om 'n element in 'n versameling of 'n ver= sameling self as geheel aan te dui. In die geval van ordinale meting kan enige transformasie wat nie die rangorde van elemente versteur nie, toegepas word. Elke telling kan byvoorbeeld met 'n konstante vermenigvul= dig word of 'n konstante kan bygetel word. Selfs vierkante, vierkants= wortels of logaritmes kan bepaal word en die resultate sal nog in die= selfde rangorde wees. Laasgenoemde bewerkings sal gewoonlik in die geval van intervalmeting nie wenslik wees nie, aangesien slegs lineêre transformasies die eienskappe van intervalmeting behou. Lineêre trans= formasies, wat voorgestel kan word deur die formule $y = mx + c$ het 'n verandering in die eenheid en die nulpunt tot gevolg. Die getransfor= meerde tellings sal ook op die vlak van intervalmeting wees. In die geval van metings op die vlak van verhoudingsmeting sal die getrans= formeerde metings weer op die vlak van verhoudingsmeting wees, slegs as al die metings met 'n konstante vermenigvuldig word.

.4 SKALINGSMETODES

Meting in die Sielkunde het te doen met die toekenning van getalle aan psigologiese eienskappe van persone. In paragraaf 1.3 is vier vlakke van meting gedefinieer en aangetoon aan watter vereistes die getalle vir die vier vlakke van meting moet voldoen. 'n Belangrike probleem is egter hoe te werk gegaan kan word om te verseker dat die getalle wat toegeken word aan die vereistes van die hoogs moontlike vlak van meting sal voldoen.

In die algemeen het die psigometriese metingsprobleem te doen met per= sone, stimuli en response. Daar word aanvaar dat 'n aantal psigologiese eienskappe in variërende hoeveelhede by elke persoon teenwoordig is.

Die probleem van skaling is dan: Watter prosedures kan gebruik word om te bepaal

(a) wat die minimum aantal psigologiese eienskappe van persone is wat 'n rol speel by hulle response op die stimuli,

(b) hoe 'n waarde vir elk van die psigologiese eienskappe aan elke per= soon toegeken kan word sodat die skaaleienskappe aan die hoogste moontlike vlak van meting sal voldoen, en

(c) hoe die inligting in die response maksimaal benut kan word?

Ongelukkig is daar nie klinkklare antwoorde op die vorige vrae nie. Daar bestaan verskeie skalingsprosedures en enkeles word hieronder ge= noem. Uiteindelik is die enigste werklike toets vir 'n geskikte ska= lingsmetode om te bepaal hoe die verkreë skaal verband hou met ander veranderlikes en of die skaal inpas in 'n teoretiese raamwerk van ver= wantskappe met ander veranderlikes. Die verkreë skaal moet, met ander woorde, kan help om natuurlike verskynsels te verklaar.

Vir die doel van hierdie handboek is dit slegs nodig om aandag te gee aan stimuli wat in die vorm van toetsitems aangebied word. Die skalings= probleem word dus: Hoe moet itemtellings toegeken word en watter funk= sie(s) van die itemtellings gee die beste meting van die onderliggende dimensie(s).

Die eerste stap in die skalingsproses is gewoonlik om te bepaal wat die minimum aantal psigologiese eienskappe is wat die response van persone op 'n gegewe stel items bepaal. Verskeie wiskundige hulpmiddels soos multidimensionele skaling en faktorontleding kan hiervoor gebruik word. In gewone toetskonstruksie word egter dikwels eenvoudig aanvaar dat 'n enkele psigologiese eienskap hoofsaaklik die response van alle persone op die items van 'n toets bepaal. Ander faktore wat 'n invloed op die response van persone op 'n stel items kan hê, word gesien as kans- en/of steuringsfaktore.

In die skalingsproses word gewoonlik as tweede stap (soms gelyktydig met die eerste stap) gepoog om itemtellings vir elke item op 'n skaal te kry wat lineêre transformasies sal wees van die skaal wat uitein=

delik verkry word uit die informasie wat in al die items van die toets vervat is. Soveel moontlik van die informasie in die itemresponse moet benut word. 'n Metode om dit te bewerkstellig, is om te sorg dat die itemtellings van 'n item tussen soveel persone moontlik in die groep vir wie die toets bedoel is, sal onderskei. 'n Voorbeeld is die volgende: Gestel 'n item in 'n intelligensietoets behels dat die toetsling 'n legkaart binne 100 sekondes moet voltooi. 'n Puntetoekenning van 1 (suksesvol) of 0 (onsuksesvol) sal tussen minder persone onderskei as 'n puntetoekenning van 1 (suksesvol binne 50 sekondes), 0,5 (suksesvol binne 75 sekondes), 0,25 (suksesvol binne 100 sekondes) en 0 (onsuksesvol binne 100 sekondes). Die beste metode van puntetoekenning sal wees om tydsintervalle te kies sodat naastenby ewe veel toetslinge in elke tydsinterval die item sal voltooi. Hoe meer intervalle, hoe meer kan tussen persone onderskei word, maar praktiese oorwegings beperk natuurlik die aantal tydsintervalle.

Die volgende stap is om 'n saamgestelde meting uit die itemtellings te verkry. 'n Probleem wat dan na vore kom is: Hoe groot moet die relatiewe bydrae van elke item wees? Anders gestel: Hoe word 'n meting van 'n vermoë ten beste uit die itemtellings verkry?

Die skalingsmodel wat vervolgens bespreek gaan word, bied 'n oplossing vir bogenoemde probleem en staan bekend as die sommodel. In die praktyk is gevind dat die sommodel bevredigende resultate vir byna alle soorte toetse lewer en gevolglik is dit die model wat die meeste gebruik word. Die eienskappe van die sommodel is soos volg:

(a) Itemtellings word so toegeken dat die potensiële maksimum- en minimumtellings vir al die items naastenby dieselfde is, en

(b) 'n lineêre kombinasie word gemaak van itemtellings (gewoonlik word hulle slegs bymekaar getel) waar elke itemtelling 'n sekere gewig het. Hierdie som van itemtellings dien dan as die meting van die onderliggende konstruk. Itemtellings moet so geskaal wees dat al die items 'n positiewe korrelasie met die onderliggende konstruk sal hê voordat die itemtellings gesommeer word.

'n Verduideliking van die reël van gelyke maksimum itemtellings is miskien nodig. Indien items variërende maksimumtellings het, sal 'n

item se bydrae tot die totaaltelling van items naastenby eweredig aan die maksimumtelling van die item wees. Items met hoë maksimumtellings sal dus 'n hoër bydrae in die totaaltelling hê. Indien die toetsopsteller rede het om sulke implisiete differensiële gewigte van items te aanvaar, is dit sy goeie reg. In die algemeen word egter goeie resultate verkry wanneer itemtellings met naastenby gelyke gewigte gesommeer word.

Die sommodel geld vir items wat dichotomies geskaal is, dit wil sê items waarin slegs tellings van 0 of 1 behaal kan word, en vir items wat 'n wyer strek van itemtellings, byvoorbeeld van -6 tot +6, kan hê. Likert (1932) het die voordele van die sommodel vir houdingskale aangetoon. Die sommodel word dan ook soms 'n Likertskaal genoem wanneer houdings gemeet moet word. Die welbekende semantiese differensiaal skalingsmetode (kyk paragraaf 4.2.4) is ook 'n voorbeeld van die sommodel. Die itemontledingstegnieke wat in hoofstuk 6 bespreek word, berus op die sommodel van skaling.

In hoofstuk 12 word 'n elementêre uiteensetting van itemresponsteorie (latente trekteorie) gegee waarin 'n teoreties bevredigender oplossing van die skalingsprobleem as die sommodel vir dichotomiese items gegee word. Die Guttman- en Thurstoneskalingsmetode is twee ander skalingsmetodes wat soms in die literatuur genoem word. Heelwat besware kan egter teen hierdie twee metodes geopper word.

LITERATUURLYS VIR HOOFSTUK 1

1 COOMBS, C.H. A Theory of Data. *Psychological review* 67, 1960: 143 - 159.

2 GAITO, J. Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin* 87, 1980: 564-567.

3 GLASS, G.V. & STANLEY, J.C. *Statistical methods in education and psychology.* New Jersey, Prentice Hall, 1970.

4 GUILFORD, J.P. *Psychometric methods.* New York, McGraw-Hill Book Co., 1954.

5 LIKERT, R. A technique for the measurement of attitudes. *Arch. Psychol.*, 1932, No. 140.

6 LORD & NOVICK. *Statistical theories of mental test scores.* Massachusetts, London, Addison-Wesley Publishing Co., 2nd Printing 1974.

7 NUNNALLY, Jum C. *Introduction to psychological measurement.* New York, McGraw-Hill, ©1967.

8 NUNNALLY, Jum C. *Psychometric methods.* New York, McGraw-Hill, ©1970.

9 TORGERSON, W.S. *Theory and methods of scaling.* New York, John Wiley, 1958.

HOOFSTUK 2
KLASSIFIKASIE EN FUNKSIES VAN TOETSE

2.1 WAT IS 'N TOETS?

In beginsel is alle verskille wat daar tussen twee persone bestaan, meetbaar. Sodra 'n verskil ten opsigte van 'n bepaalde kenmerk merkbaar is, en een persoon "meer as" of "minder as" 'n ander een van hierdie kenmerk kan besit, dan kan daar op grond van hierdie merkbare verskil "gemeet" word. As daar aan die gewig, ouderdom, lengte, vriendelikheid, hulpvaardigheid, eerlikheid, ensovoorts van 'n individu gedink word as bepaalde kwaliteite wat hy besit en hierdie kwaliteite hom in enige opsig van 'n ander individu onderskei, kan dit as kwantifiseerbare kwaliteite beskou word. E.L. Thorndike soos aangehaal deur Downie (1959), se stelling is dat enigiets wat bestaan, bestaan in 'n hoeveelheid en enigiets wat in 'n hoeveelheid bestaan, kan gemeet word. Primêr is 'n toets dus 'n instrument wat in staat is om daardie "meer as" of "minder as" ten opsigte van 'n bepaalde eienskap by 'n individu te bepaal sodat proefpersone ten opsigte van hierdie eienskap in rangorde geplaas kan word.

Volgens Cronbach (1970) is daar nie eintlik 'n bevredigende definisie vir die woord "toets" nie en is die algemene opvatting dat dit 'n prosedure vir die stel van 'n reeks standaardvrae is. Hy beskryf dit egter as 'n sistematiese prosedure om die gedrag van 'n persoon te bestudeer en dit met behulp van 'n numerieke skaal of 'n sisteem van kategorieë te beskryf. Vir Anastasi (1961) is 'n sielkundige toets hoofsaaklik 'n objektiewe en gestandaardiseerde meetmiddel met behulp waarvan 'n monster van menslike gedrag gemeet kan word. 'n Aanvaarbare definisie vir die doel van hierdie handboek, is dié van Freeman (1962) wat daarop neerkom dat 'n sielkundige toets 'n gestandaardiseerde instrument is wat so ontwerp is dat een of meer aspekte van die totale persoonlikheid van 'n individu met behulp van steekproewe van verbale of nie-verbale response of ander gedragsuitinge gemeet kan word.

2.2 INDELING VAN TOETSE

Alvorens die funksies van toetse bespreek word, word 'n indeling van toetse gegee, aangesien verskillende tipes toetse opgestel word om

verskillende funksies te verrig. Daar is verskillende maniere om toetse in te deel, maar vir die doel van hierdie bespreking word twee groot groepe onderskei:

(a) Toetse vir maksimum taakverrigting

Met hierdie toetse word beoog om die beste wat 'n proefpersoon kan lewer, te bepaal. Hulle kan ook vermoëtoetse genoem word. Hieronder val toetse vir die bepaling van algemene intelligensie, aanleg, bekwaamheid en skolastiese prestasie.

(b) Toetse vir tipiese taakverrigting

Hierdie toetse word opgestel om die tipiese gedrag van 'n individu te bepaal, dit wil sê wat sy tipiese reaksie in 'n bepaalde situasie sal wees. Hieronder val toetse vir die meting van persoonlikheidstrekke soos gewoontes, belangstellings, houdings en temperament.

Toetse in hierdie twee breë groepe kan soms oorvleuel, aangesien daar nie altyd 'n duidelike skeidslyn tussen tipiese gedrag en vermoë getrek kan word nie.

2.3 TOETSE VIR MAKSIMUM TAAKVERRIGTING

Voorbeelde hiervan is die volgende:

(a) Intelligensietoetse

Sielkundiges is dit redelik eens dat wat 'n kind op eie inisiatief uit sy omgewing aanleer, 'n maatstaf van sy intelligensie is. Omdat een kind meer aanleer as 'n ander het sommige aanvaar dat die een met meer intelligensie as die ander gebore word. Of hierdie intelligensie nou aangebore word, of die gevolg is van omgewingsinvloede of die wisselwerking tussen die twee, is nie van kardinale belang nie. Daar word nie hier ingegaan op die vele en uiteenlopende definisies van intelligensie nie. Daar word volstaan met die feit dat die meeste definisies in hoofsaak drie kerngedagtes bevat, naamlik dat intelligensie te doen het met

- die hoëre psigiese funksies in die mens,
- die vermoë om te leer en
- die vermoë tot aanpassing by die omgewing - in die breedste sin beskou.

Die funksies van intelligensietoetse kan soos volg uiteengesit word:

(i) Bepaling van algemene intelligensie

Toetse vir algemene intelligensie word ontwerp om 'n kwantitatiewe beskrywing van 'n toetsling se intelligensie te gee. Daar moet op gewys word dat 'n intelligensiemeting nie die alfa en die omega van 'n persoon se vermoëns is nie en nie 'n beskrywing gee van alles waartoe die persoon in staat is nie. Die intelligensietoetse waaroor ons vandag beskik, is nie volmaak nie, maar nieteenstaande hulle tekortkominge word hulle met vrug in die onderwys gebruik as hulpmiddels om 'n oorsigtelike beeld van 'n leerling se vermoëns te bekom.

(ii) Individuele diagnose

Intelligensietoetse kan gebruik word om mense beter te verstaan. Voorbeelde:

(1) Vir onderpresteerders op skool kan dikwels bepaal word waar die oorsake lê as hulle nie met die skoolwerk kan tred hou nie. Aangesien die meeste intelligensietoetse verskillende fasette van intelligensie meet, kan die onderskeie sterk en swak punte in die intelligensie van die kind bepaal word. Lae skoolprestasie kan egter ook veroorsaak word deur onder andere swak gesigsvermoë of gehoor of wanaanpassing in die klas.

(2) Die lastige kind met 'n hoë intelligensie kan uitgeken word. So kan dalk verklaar word waarom 'n leerling maklik in die klas verveeld raak.

(3) 'n Onderwyser kan 'n insig kry in 'n leerling se motivering vir sy skoolwerk deur die kind se intelligensie met sy skoolprestasie te vergelyk.

(iii) Skoolorganisasie

Die intelligensietoets kan gebruik word as

(1) hulpmiddel in groot skole waar daar vir elke standerd 'n hele aantal klasse is en die leerlinge in relatief homogene groepe ingedeel moet word, en

(2) om aan die onderwyser 'n aanduiding te gee van wat van 'n klas as geheel verwag kan word sodat sy onderrigmetodes by omstandighede aangepas kan word.

(iv) Seleksie van leerlinge vir toelating

By sekere inrigtings, veral dié vir hoër onderwys en privaatskole waar daar meer aansoeke om toelating is as wat geakkommodeer kan word, kan die intelligensietoets 'n hulpmiddel wees by die seleksie van leerlinge sodat sover moontlik net dié toegelaat word wat die meeste sal baat by die onderrig wat aangebied word.

(v) Opvoedkundige leiding

Leerlinge wat aan die begin van hul hoërskoolloopbaan staan, is meesal nog onbewus van hul vermoëns en besluiteloos in verband met hul vakkeuse. In samewerking met die voorligteronderwyser is dit die taak van die hoof om aan hierdie leerlinge en hulle ouers die regte leiding in hierdie verband te gee. Resultate behaal in 'n intelligensietoets is een van die nuttige bronne van inligting waarop die hoof hierdie leiding kan baseer.

(vi) Beroepsleiding

Wanneer die leerling die einde van sy skoolloopbaan nader, tree die vraagstuk van 'n beroepskeuse al meer op die voorgrond en gaan die opvoedkundige leiding geleidelik oor in beroepsleiding. Intelligensietoetse kan ook hier 'n vername funksie vervul, hoewel hulle alleen nie as voldoende hulp-

middels beskou moet word om beroepsvoorligting op te baseer nie.

(vii) Navorsing

In navorsingsprojekte is 'n betroubare intelligensietoets soms van belang veral waar gelykwaardige groepe as kontrole=groepe geselekteer moet word.

(b) Aanlegtoetse

Waar intelligensie na 'n algemene vermoë verwys, het aanleg te doen met 'n spesifieke potensiële vermoë waaroor 'n individu op 'n sekere stadium beskik en wat hom in staat stel om sekere bekwaamhede te ontwikkel. 'n Basiese beginsel wat by aanlegtoetse geld, is dat elke proefpersoon in die populasie wat getoets word, naastenby dieselfde geleenthede moes gehad het om ervaring op te doen en om te leer. Die verskil in prestasie van persone word dan toegeskryf aan die verskil in hulle aanleg.

Terwyl die intelligensietoets die algemene vermoë van 'n proefpersoon in een getal uitdruk, kom die aanlegtoets gewoonlik in 'n battery toetse voor en is die eerste belangrike funksie van die toetsbattery om deur middel van 'n reeks tellings meer lig op die verskillende spesifieke vermoëns van die proefpersoon te werp. Die aanlegtoetsbattery verstrek dus meer inligting as die intelligensietoets.

Sekere aanlegtoetse kan gebruik word om sukses in bepaalde studierigtings te voorspel, maar daar is ook aanlegtoetse vir voorspelling van sukses in bepaalde beroepsrigtings soos ingenieurswese, musiek, regte, en so meer. In die praktyk kom dit soms voor dat 'n algemene intelligensietoets gebruik word as 'n siftingstoets by groot getalle toetslinge om eers 'n globale indruk van hulle vermoë te verkry en dat daarna dan aanlegtoetse toegepas word op daardie geselekteerde gevalle waar meer inligting nodig is aangaande hulle spesifieke aanlegte. Byvoorbeeld nadat 'n groep militêre rekrute hulle basiese opleiding voltooi het, kan hulle op grond van algemene vermoë in aanmerking kom vir opleiding as offisiere in die lugmag maar op grond van spesifieke aanlegte, soos uitgewys deur

aanlegtoetse, kan verdere seleksie plaasvind vir opleiding as vlieënier, navigator, kanonnier, ensovoorts.

(c) Bekwaamheidstoetse

'n Bekwaamheidstoets kan gesien word as 'n toets wat bepaal in welke mate 'n taak wat op sigself van belang is, byvoorbeeld om Duits te kan lees of klavier te kan speel, bemeester is. Bekwaamheidstoetse is gebaseer op ervaring en verworwe vaardighede maar die items is nie leerplangebonde nie, alhoewel kennis wat deur middel van algemene opvoedkundige agtergrond - ook skoolonderrig - bekom is, veronderstel word.

Die funksies van die bekwaamheidstoets is om

- die toetsling se bekwaamheid in die besondere vakgebied wat deur die toets gedek word, te bepaal,

- die toetsling se toekomstige sukses in die betrokke kennisgebied te voorspel en

- vir keuring en plasing van toetslinge gebruik te word, veral in beroepsverband waar 'n spesifieke leereenheid nie noodwendig voltooi hoef te wees nie.

(d) Prestasietoetse

Die prestasietoets word opgestel om na afhandeling van 'n sekere gedeelte van die leerplan in 'n vak toegepas te word en is volkome leerplangebonde. Vir elke vak sal 'n afsonderlike prestasietoets opgestel word. Die toets het dus betrekking op die huidige prestasie van die toetsling. Dit kan egter ook vir voorspelling gebruik word, maar dan neig dit meer in die rigting van 'n aanlegtoets. Onderrig en evaluering is onafskeidbaar aan mekaar verbonde en prestasietoetse word dikwels in skoolverband gebruik waar hulle 'n verskeidenheid van funksies vervul.

(i) Funksies met betrekking tot die leerling

(1) Terugvoering

Die prestasietoets vervul 'n belangrike funksie as daar, nadat die toets nagesien is, terugvoering na die leerling is. Deur klem te lê op daardie dele van die werk waarin die leerling swak gepresteer het in die toets, kan aan hom uitgewys word watter dele van die leerplan hy nog nie bemeester het nie en meer lig gewerp word op die doeltreffendheid van sy studiemetodes.

(2) Motivering

Algemene ervaring toon dat 'n leerling meer moeite doen met sy werk as daar op bepaalde tye 'n toets afgelê word en dat hy dan meer klem op daardie dele van die werk lê waarop vrae verwag word. Die wete dat 'n toets afgelê gaan word, het 'n heilsame uitwerking op die leerling se motivering en gee rigting aan sy studie. Baie leerlinge het egter geen spesiale aansporing tot studie nodig nie, maar vir 'n groot persentasie is die motivering wat deur die toets verskaf word net onmisbaar.

(3) Diagnose

Die prestasietoets kan ook soms gebruik word vir die diagnosering van die leerling se swakhede met die oog op remediërende onderrig. 'n Deeglike ondersoek na die oorsake van sy probleme behoort egter eers ingestel te word voordat 'n remediërende program opgestel word.

(ii) Funksies met betrekking tot die onderwyser

(1) Doeltreffendheid van onderrig

Die onderwyser behoort van tyd tot tyd na te gaan hoe suksesvol sy onderrigmetodes is en in hoeverre hy sy doelstellings met sy onderrig bereik het. Dit sal

onverstandig wees om maand na maand met onderrig voort te gaan sonder om gereeld 'n opname te maak van die suk= ses wat behaal is. Sommige metodes is meer effektief as ander om die beoogde doelstellings te bereik en daar= om moet die onderwyser weet met watter metodes hy kan voortgaan, watter metodes hy moet probeer verbeter en watter hy heeltemal moet laat vaar.

Die prestasietoets vervul in hierdie verband 'n funksie wat beswaarlik deur ander middels vervang kan word.

(2) Verkenning

Met behulp van die prestasietoets kan die onderwyser vooraf bepaal wat sy leerlinge reeds weet en wat nie. Op grond van die resultate wat met 'n eerste toepassing verkry word, kan hy verder beplan om aan die spesifieke behoeftes van die klas as geheel en ook van die indivi= duele leerlinge te voldoen.

(iii) Ander funksies

Elke skool het 'n plig, nie alleen teenoor die leerlinge nie, maar ook teenoor ouers, die gemeenskap en ander onderwysin= rigtings wat die leerling vir verdere studie mag opneem om bevredigende standaarde te handhaaf. Die prestasietoets het dus ook 'n baie belangrike funksie te vervul met die oog op

(1) bevordering van leerlinge na 'n hoër standerd

Die prestasietoets kan gebruik word vir die bepaling van die leerling se peil van kennis naamlik of dit van so 'n standaard is dat hy na 'n hoër standerd gepromoveer kan word. Dit is ook van nut om die standaard wat deur die leerlinge behaal is te vergelyk met dié van die normgroep.

(2) verslaglewering aan ouers

Van tyd tot tyd behoort die leerlinge se ouers ingelig te word oor die vordering van hulle kinders en ook van raad gedien te word in verband met die keuse van vakke en toekomstige loopbane. Saam met ander meetinstrumente soos intelligensie- en aanlegtoetse kan die prestasie=toets gebruik word om 'n beeld te vorm van die leerling se potensialiteite.

(3) toelating tot ander inrigtings of beroepe

Die prestasietoets kan in samehang met ander meetmid=dels gebruik word vir keuringdoeleindes veral by toe=lating tot inrigtings vir verdere studie of by die keuring van werkaansoeke.

(e) Leerdoelgerigte toetse

In elke skoolvak is daar sekere basiese beginsels, kennis en vaar=dighede wat deur alle leerlinge bemeester moet word en wat beskou kan word as die minimum kennis waaroor 'n leerling moet beskik al=vorens hy 'n volgende fase van onderrig in die vak met sukses sal kan aanpak.

Met leerdoelgerigte toetse word dan beoog om te bepaal of die leer=linge die basiese beginsels, kennis en vaardighede wat 'n betrokke leereenheid ten grondslag lê, onder die knie het. Afsnypunte is uiteraard gewoonlik hoog in hierdie toetse omdat van leerlinge verwag word om feitlik volpunte te behaal. In die opstelling van leerdoelgerigte toetse is dit nodig dat elke leeruitkoms in die fynste besonderhede ontleed en beskryf word. In elke leereenheid word bepaalde mikpunte gestel. Dit impliseer verder dat hersiening gedoen en/of remediërende onderrig gegee moet word na elke toetsing indien dit sou blyk dat die betrokke leerstof nie bemeester is nie.

Net soos in die geval van prestasietoetse is inhoudsgeldigheid 'n belangrike vereiste vir die items van leerdoelgerigte toetse.

2.4 TOETSE VIR TIPIESE TAAKVERRIGTING

2.4.1 Inleiding

Die toetse in paragraaf 2.3 genoem staan bekend as "kognitiewe" toetse. Daarteenoor kry ons die "nie-kognitiewe" toetse of persoonlikheidstoetse vir die evaluering van emosionele, sosiale en motiveringsaspekte van ge= drag. Die doel met hierdie tipe toetse is nie om te bepaal wat 'n indi= vidu kan doen of in staat is om te doen nie, maar wat hy onder tipies normale omstandighede wel doen. Behalwe tipiese persoonlikheidstrekke soos introversie-ekstraversie, aggressie, angs, en so meer, word be= langstellings, houdings ten opsigte van sekere vraagstukke, studiege= woontes, en so meer, ook as aspekte van 'n persoon se persoonlikheid beskou.

Daar word hier volstaan deur te meld dat die vernaamste metodes wat vir die evaluering van nie-kognitiewe eienskappe gebruik word die volgende is:

- vraelyste,
- observasiemetodes,
- projektiewe tegnieke en
- fisiologiese metings.

Aangesien daar aansienlike oorvleueling is by verskillende instrumente, word daar nie gepoog om die persoonlikheidstoetse volgens trekke wat gemeet word te klassifiseer nie en word die funksies wat hulle vervul in die algemeen en gesamentlik bespreek.

2.4.2 Funksies van persoonlikheidstoetse

a. Voorspelling

Indien bepaalde waarnemings in verband met 'n persoon se gedrag in sekere situasies gemaak word, word dikwels aanvaar dat hy in die toekoms in ander soortgelyke situasies min of meer dieselfde gedragspatrone sal openbaar. 'n Belangrike funksie van persoonlikheidstoetse is dus om inligting aan= gaande 'n individu se persoonlikheidstrekke te bekom en met behulp daarvan voorspellings in verband met toekomstige gedrag soos aanpassing en skool-,

akademiese en beroepsprestasies te maak. Voorspellings behoort egter nooit op grond van enkele toetse gemaak te word nie. Saam met aanleg= toetse vervul belangstellingsvraelyste, byvoorbeeld, 'n belangrike funk= sie op die gebied van beroeps- en opvoedkundige leiding. Prestasie is die gevolg van sowel belangstelling as aanleg en op grond van resultate in toetse in verband met albei kan 'n voorligter doeltreffender voor= spellings maak as wanneer hy net een tipe meetinstrument sou gebruik.

b. Beskrywing (kliniese diagnose)

Die studie van persoonlikheid behels onder meer twee vraagstukke, naam= lik

- watter persoonlikheidseienskappe openbaar 'n individu op 'n bepaalde tydstip en

- op welke wyse het die individu daardie bepaalde eienskappe ontwikkel, dit is, hoe het hy so geword.

Persoonlikheidstoetse het veral met die eerste vraagstuk te doen en die funksie wat die persoonlikheidstoetse in hierdie verband vervul is om inligting te verskaf oor eienskappe soos die volgende:

(i) Sosiale trekke: Die individu se verhouding tot sy medemens en hoe hy in die samelewing optree. Tipiese trekke is byvoorbeeld eerlikheid, samehorigheidsgevoel of gemeenskapsgevoel, aanpassing, dominasie, en so meer.

(ii) Motiewe: Die dinamika agter persoonlikheid en wat die "behoeftes" en "dryfvere" verteenwoordig wat agter sekere vorme van gedrag sit soos aggressie, prestasie en bepaalde behoeftes.

(iii) Persoonlike houdings of opinies: Wat mense van hulself of van andere dink, hoe hulle oor sekere vraagstukke voel, en so meer.

c. Terapie

Die resultate van 'n persoonlikheidstoets kan deur die voorligter met die toetsling bespreek word en die toets kan werk soos 'n spieël wat

alles wat hy in die toets oor homself gesê het, reflekteer. Die self= verslag bied informasie aan die voorligter wat vir die toetsling geïn= terpreteer kan word. Die toetsresultate kan 'n gesprekmedium word en dit op sigself is al 'n element van die psigoterapie wat kan bydra om die moeilikhede wat die leerling ondervind met die formulering van sy gedagtes en probleme te oorbrug. Die toetsresultate bevorder dus goeie rapport tussen voorligter en toetsling en dit weer baan die weg tot beter begrip en selfkennis - een van die vernaamste oogmerke by terapie.

d. Voorligting

Wanneer aan 'n individu raad en leiding gegee moet word in verband met vakkeuse, verdere studie of 'n beroepskeuse, kan daar nie maklik van te veel meetinstrumente gebruik gemaak word nie. Soveel inligting moont= lik moet ingesamel word sodat alle aspekte van die persoonlikheid deeg= lik ondersoek kan word. 'n Persoon kan byvoorbeeld die aanleg vir 'n bepaalde beroep hê, maar as hy nie die belangstelling daarvoor het nie, is dit waarskynlik dat hy later ongelukkig in sy beroep sal wees. Daarom is dit so belangrik dat daar nie net alleen na aanleg gekyk word wanneer voorligting in verband met 'n beroepskeuse gegee word nie. Naas aanleg- en bekwaamheidstoetse is persoonlikheidstoetse by voorligting dus van groot waarde.

LITERATUURLYS VIR HOOFSTUK 2

1 AIKEN, L.R. Jr. *Psychological and educational testing.* Boston, Allyn and Bacon, 1971.

2 ANASTASI, A. *Psychological testing.* New York, Macmillan, 1961.

3 BLOOM, B.S., HASTINGS, J.T. & MADAUS, G.F. *Handbook on formative and summative evaluation of student learning.* New York, McGraw-Hill, 1971.

4 BROWN, F.G. *Principles of educational and psychological testing.* Hinsdale (Illinois), The Dryden Press, 1970.

5 CRONBACH, L.J. *Essentials of psychological testing.* New York, Harper and Row, 1970.

6 DOWNIE, N.M. *Fundamentals of measurement.* New York, Oxford University Press, 1959.

7 FREEMAN, F.S. *Theory and practice of psychological testing.* Third ed. New York, Holt, Reinhart and Winston, ©1962.

8 GRONLUND, N.E. *Measurement and evaluation in teaching.* New York, Macmillan, 1965.

9 NUNNALLY, J.C. Jr. *Test and measurements assessment and prediction.* New York, McGraw-Hill, 1959.

10 SARGENT, W.E. *Teach yourself psychology.* London, English Universities Press, 1955.

11 THORNDIKE, R.L. & HAGEN, E. *Measurement and evaluation in psychology and education.* 3rd ed., New York, John Wiley and Sons, 1969.

HOOFSTUK 3
DIE ONTWERP VAN TOETSE

3.1 INLEIDING

Menslike eienskappe wat deur toetse gemeet word, is gewoonlik uiters kom= pleks en relatief moeilik definieerbaar. Dit gebeur ook dikwels dat en= kele definisies van sulke eienskappe nie deur alle kenners op die betrokke gebied aanvaar word nie. As voorbeeld kan genoem word dat daar nie eers 'n enkele algemeen aanvaarbare definisie bestaan vir intelligensie nie alhoewel redelik algemeen aanvaar word dat intelligensie redelik betrou= baar en geldig gemeet kan word.

3.2 KONSTRUKTE

Menslike eienskappe kan nie direk met toetse gemeet word nie. So byvoor= beeld is dit onmoontlik om alle intelligente gedrag (wat dit ook al pre= sies mag wees) in een enkele toets vir intelligensie te betrek. 'n Infe= rensiële sprong moet dus gemaak word wanneer 'n meting met 'n toets ge= interpreteer word as 'n meting van 'n relatief abstrakte eienskap. Ab= strakte menslike eienskappe word ook dikwels konstrukte genoem.

3.3 DEFINISIE VAN DIE KONSTRUK WAT GEMEET MOET WORD

Een van die belangrikste stappe in die opstelling van 'n toets is om die eienskap wat gemeet moet word so volledig as moontlik te definieer. Die definisie behoort so opgestel te word dat dit 'n metingstegniek sal sug= gereer. 'n Definisie wat aan hierdie vereiste voldoen word 'n operasionele definisie genoem.

3.4 FASETTE

Nadat 'n konstruk gedefinieer is, is dit nodig om die gedragsveld waarop die beplande toets betrekking het, so duidelik as moontlik af te baken. Dit is dikwels weer moontlik om die gedragsveld te verdeel volgens eien= skappe, wat fasette genoem word. Die identifisering van die belangrik= ste fasette van 'n konstruk behoort te berus op goedgevestigde teoretiese oorwegings. Elke faset van 'n konstruk kan weer in elemente verdeel word. As ons byvoorbeeld aan die konstruk geheue dink, kan onder andere die

volgende fasette van geheue geïdentifiseer word

- *tydperk* met drie elemente naamlik korttermyn, mediumtermyn en langter=myn,
- *inhoud* met vier elemente naamlik figure, simbole, taal en gedrag,
- *intensionaliteit* met twee elemente naamlik intensionele leer en nie-intensionele (insidentele) leer.

In die voorgaande voorbeeld kan die geheuekonstruk in 24 dit is (3 x 4 x 2) selle verdeel word. 'n Voorbeeld van een so 'n sel is dan *mediumtermyngeheue vir woorde wat intensioneel geleer word.*

3.5 INHOUDSGELDIGHEID

Die toetsopsteller moet in hierdie stadium besluit watter gewig aan elke sel in die konstruk vir die doel van meting toegeken moet word. Die ge=tal items in die toets afkomstig uit 'n sel is proporsioneel aan die gewig wat aan die sel toegeken word. Hierdie toekenning van gewigte hang ten nouste saam met die inhoudsgeldigheid van 'n toets (kyk hoofstuk 11). Vir 'n toets om inhoudsgeldigheid te besit, moet die gewigte sodanig toe=geken word dat vakkenners min of meer daarmee sal saamstem.

3.6 DIE PROBLEEM VAN DIMENSIONALITEIT

'n Kardinale aspek waaroor die toetsopsteller 'n besluit moet neem, is of die toets een enkele meting moet lewer (een ding meet) en of die afson=derlike selle in die konstruk of kombinasie van selle elk 'n afsonderlike meting moet oplewer (verskillende dinge meet). Die onderliggende pro=bleem wat so 'n besluit moeilik maak, is dat 'n enkele meting waarvan die uitslag deur 'n aantal faktore bepaal word, nie eenduidig interpreteer=baar is nie. Metings vir selle van konstrukte moet dus redelik hoog met mekaar korreleer voordat die selle saamgevoeg mag word om 'n enkele meting op te baseer.

Ongelukkig is dit nie altyd vooraf moontlik om die dimensionaliteit van die totale konstruk te bepaal nie. As daar tydens die beplanningstadium enigsins vermoed word dat die dimensionaliteit meervoudig is, is dit die beste om sodanig te beplan dat 'n afsonderlike toets opgestel word vir elke homogene (wat dimensionaliteit betref) groep selle in die oorspronk=

like breë konstruk. So byvoorbeeld sal vir die breë geheuekonstruk waarskynlik afsonderlike toetse opgestel moet word vir kort-, medium- en langtermyngeheue.

3.7 DIE RASIONAAL VAN 'N TOETS

Beskou nou 'n eendimensionele konstruk wat bevredigend gedefinieer is en waarvan die totale gedragsveld volgens 'n aantal fasette in redelik afgebakende selle verdeel kan word. Die gedragsveld in 'n enkele sel sal nog 'n oneindige aantal verskillende maar soortgelyke gedragsuitings verteenwoordig. Omdat dit onmoontlik is om die oneindige aantal gedragsuitings in die toets te betrek, word 'n sistematiese steekproef van die gedragsuitings deur middel van bepaalde tipes toetsitems in die toets betrek en *aanvaar* dat die toetsgedrag 'n getroue weergawe gee van die totale gedragsveld. Hierdie aannames vir elke sel van die totale gedragsveld tesame met die aannames oor die gewigte wat vir elke sel moet geld, word die rasionaal van die toets genoem.

3.8 DIE TOETSMODEL

Die besluite wat die toetsopsteller neem in verband met

- die onderliggende sielkundige/opvoedkundige teorie,
- definisie van die konstruk,
- fasette van die konstruk en
- die rasionaal van die toets

wat gebruik gaan word, word die toetsmodel genoem. Klaarblyklik verskaf die toetsmodel riglyne waarvolgens die items van die toets opgestel moet word.

3.9 DOEL MET DIE GEBRUIK VAN 'N TOETS

Die primêre doel met 'n toets is om een of ander konstruk te meet. Die sekondêre doelstellings van 'n toets het meer betrekking op die redes waarom 'n meting van 'n bepaalde konstruk nodig sal wees, met ander woorde, op die gebruike van die toets. In die vorige hoofstuk is 'n aantal gebruike vir 'n aantal tipes toetse genoem. Dit is noodsaaklik om voordat 'n toets ontwerp word, kennis te neem van die waarskynlike gebruike van

die toets. Die groep of groepe persone op wie die toetse toegepas kan word, is ook baie belangrik. Sowel die waarskynlike gebruike van 'n toets as die groep(e) persone vir wie die toets bedoel is, kan 'n wesentlike rol speel by die opstel van 'n toetsmodel, die bepaling van die optimum lengte van die toets en die keuse van geskikte stimulusmateriaal.

3.10 ANDER BELANGRIKE BESLUITE WAT IN DIE ONTWERPSTADIUM GENEEM MOET WORD

3.10.1 Skalingsmetode

Daar moet besluit word hoe punte vir elke item toegeken moet word. Die tipe items in 'n toets word in so 'n mate deur die rasionaal van die toets bepaal dat 'n nasienmetode gespesifiseer kan word sodra die rasionaal geformuleer is.

3.10.2 Normatiewe en ipsatiewe meting

'n Saak wat verband hou met die skalingsmetode is die keuse van 'n normatiewe of ipsatiewe metingstegniek.

'n Stel veranderlikes wat betrekking het op 'n persoon is normatief as die blote feit dat enigeen van die veranderlikes vir die persoon gemeet word geen invloed sal hê op die uitkoms van meting vir enigeen van die ander veranderlikes nie. Persone kan dus onafhanklik in rangorde geplaas word vir elkeen van die stel normatiewe veranderlikes.

'n Stel veranderlikes wat betrekking het op 'n persoon is ipsatief wanneer die som van die veranderlikes konstant is vir alle persone. In die geval van 'n ipsatiewe stel veranderlikes is dit *nie* moontlik om persone onafhanklik in rangorde te plaas vir elke veranderlike in die stel *nie*.

In die algemeen word gevind dat nie veel afleidings gemaak kan word wanneer twee metings X_1 en X_2 vir 'n veranderlike uit 'n stel ipsatiewe veranderlikes vir persone 1 en 2 respektiewelik bekend is nie. Die aard van eienskappe wat ipsatief gemeet kan word, is gewoonlik sodanig dat die relatiewe grootte van X_1 en X_2 nie betekenis het nie. In teenstelling hiermee kan sinvolle afleidings soms oor 'n enkele persoon gemaak word wanneer die metings A, B, C ... van elkeen van 'n stel ipsatiewe veranderlikes vir die persoon bekend is. Die relatiewe groottes A, B,

C ... kan wel betekenisvolle informasie oor die bepaalde persoon vir wie die metings gemaak is, gee.

Die volgende voorbeeld sal die aard en verskil van normatiewe en ipsatiewe meting illustreer:

Gestel die werklike bedrae wat twee persone A en B elkeen spandeer aan behuising, voedsel, vermaak, besparing en ander uitgawes is bekend. Hierdie stel van vyf veranderlikes kan as 'n normatiewe stel veranderlikes gesien word. As A R100 vir behuising spandeer het, maak dit sin om te sê dat hy meer spandeer het as B wat net R50 spandeer het.

Gestel egter dat *slegs* die persentasie van sy eie inkomste wat elk van twee persone A en B spandeer het aan behuising, voedsel, vermaak, besparing en ander uitgawes bekend is. Hierdie stel van vyf veranderlikes kan slegs as 'n ipsatiewe stel veranderlikes gesien word. As A 25 % van sy inkomste aan behuising spandeer het en B 50 % van sy inkomste kan nie veel oor hulle relatiewe spandering aan behuising gesê word nie. Aan die ander kant het die persentasies vir elke persoon wel betekenis. As A 25 % van sy inkomste aan behuising spandeer het en 20 % aan voedsel het dit betekenis om te sê dat hy 5 % meer *van sy inkomste* aan behuising spandeer het as aan voedsel. Let daarop dat die som van die veranderlikes in die ipsatiewe stel veranderlikes in hierdie voorbeeld vir elke persoon gelyk is aan 100 % terwyl die som van die veranderlikes in die normatiewe stel veranderlikes vir A gelyk is aan R400 en vir B gelyk aan R100.

Vir die huidige bespreking is dit voldoende om verder te meld dat:

(a) 'n Stel veranderlikes kan op 'n geïmpliseerde skaal varieer tussen suiwer normatief en suiwer ipsatief,

(b) daar reeds by die ontwerp van 'n verbandhoudende aantal toetse besin moet word of die metings normatief of ipsatief moet wees,

(c) daar teen gewaak moet word om ipsatiewe metings soos normatiewe metings te interpreteer,

(d) 'n stel normatiewe veranderlikes altyd na 'n stel ipsatiewe verander=

likes getransformeer kan word maar selde andersom,

(e) vanweë die lineêre verband tussen 'n stel ipsatiewe veranderlikes gewoonlik probleme ondervind word wanneer meerveranderlike statistiese verwerkings op die veranderlikes gedoen word en

(f) daar 'n saak voor uit te maak is dat die belangstellingsvelde wat deur 'n belangstellingsvraelys gemeet word 'n ipsatiewe stel veranderlikes behoort te wees.

3.10.3 Lengte van die toets

Die lengte van 'n toets (aantal items) bepaal in 'n groot mate dié betroubaarheid en dus ook die geldigheid van die toets. In die algemeen kan gesê word dat 'n toets so kort as moontlik moet wees (veral ook wat tyd aanbetref) met dien verstande dat die sekondêre doelstellings van die toets nog in 'n bevredigende mate bereik kan word. Selfs baie goeie toetse kan in onbruik raak as die totale toetstyd daarvoor uitermate lank is.

3.10.4 Bepaling van toetstyd

Die metode waarmee toetstyd bepaal moet word, behoort in die ontwerpstadium bepaal te word. Vir kragtoetse, dit wil sê toetse waar spoed nie 'n rol in die toetsmodel speel nie, word gewoonlik die tyd wat 90 % toetslinge in die normgroep neem om die toets te voltooi as toetstyd geneem. Vir spoedtoetse, dit wil sê toetse waar spoed wel 'n rol in die toetsmodel speel, kan 'n geskikte toetstyd bepaal word deur dié tydsbeperking wat die beste verhouding tussen die gemiddelde, standaardafwyking en strek van die toetstellings teweegbring as toetstyd te neem. 'n Verhouding van 1 tot 6 tussen die standaardafwyking en die strek van toetstellings word as geskik beskou.

3.10.5 Betroubaarheid en geldigheid

Reeds in die beplanningstadium behoort besluit te word hoe hierdie twee eienskappe geëvalueer moet word. Kyk asseblief hoofstukke 9 en 11 wat oor hierdie twee onderwerpe handel.

3.10.6 Normgroepe

Die besluit oor die groepe persone vir wie norms afsonderlik bereken moet word, hang saam met kennis wat reeds bestaan oor die verband van die onderhawige konstruk met steuringsveranderlikes (kyk hoofstukke 9 en 10) en die waarskynlike gebruike van die toets. Afgesien van bekende steuringsveranderlikes behoort informasie oor moontlike steuringsveranderlikes soos geslag, standerd, sosio-ekonomiese omstandigheede, geografiese gebied, tipe skool bygewoon, en so meer ingewin te word.

3.11 VOORBEELD VAN 'N TOETSMODEL VIR 'N PRESTASIETOETS

By die opstel van die spesifikasies vir 'n prestasietoets kan twee fasette van prestasie in die betrokke vak in gedagte gehou word, naamlik die inhoud van die leerstof en die manipulasie van die leerstof.

3.11.1 Die inhoud van die leerstof

Die toetsinhoud behoort verteenwoordigend van die sillabusinhoud van die betrokke vak te wees. Daarom is dit nodig om

- die sillabusinhoud in besonderhede te ontleed,
- dit in geskikte onderafdelings te verdeel en
- die regte gewig aan elke onderafdeling toe te ken.

Wanneer so 'n toets 'n wye toepassingsveld het is dit verder nodig dat die toetsspesifikasies eerstens aan 'n paneel deskundiges op die betrokke vakgebied voorgelê word.

Tweedens behoort 'n hele aantal handboeke in algemene gebruik geraadpleeg te word. Die gemiddelde aantal bladsye wat aan elke onderafdeling bestee word, kan byvoorbeeld gebruik word as 'n aanduiding van die gewig wat aan elke afdeling toegeken behoort te word.

3.11.2 Manipulasie van die leerstof

Die doel met onderrig is nie bloot om kennis van die leerstof aan die leerlinge oor te dra nie maar ook om sekere opvoedkundige doelstellings

na te streef. Daar is egter geen unieke patroon van opvoedkundige doelstellings wat op elke vak van toepassing gemaak kan word nie. Die indeling van doelstellings hieronder wat uit Ebel (1971) ontleen is, kan gebruik word:

- Begrip van terminologie
- Begrip van feite of beginsels
- Vermoë om te verduidelik of te illustreer
- Vermoë om berekeninge te doen
- Vermoë om resultate te voorspel
- Vermoë om toepaslike handeling voor te stel
- Vermoë om 'n evaluerende uitspraak te gee

'n Ander indeling van opvoedkundige doelstellings met die leerstof soos deur Bloom, Hastings en Madaus (1971) uiteengesit, is:

- Kennis
- Begrip
- Toepassing
- Ontleding
- Sintese
- Evaluasie of oordeel

3.11.3 Die tweerigtingtabel(model)

Die gewigte wat aan leerstofinhoud en opvoedkundige doelstellings toegeken word, kan in 'n tweerigtingtabel soos in tabel 3.1 aangedui word. Die leerstofinhoud word vertikaal gerangskik en die opvoedkundige doelstellings met die onderrig horisontaal. Die gewigte per sel word verteenwoordig deur die syfers in die blokkies waar die rye en kolomme mekaar kruis. Die gewigte dui ook die getal items aan wat vir die betrokke selle geskryf moet word. Hierdie tabel kan waarskynlik vir enige vak aangepas word.

TABEL 3.1

VAKSPESIFIKASIETABEL

Aspekte van onderrig / Leerstofinhoud	Begrip van		Vermoë om					Totaal
	Terminologie	Feit of beginsel	Te verduidelik of te illustreer	Te bereken	Resultate te voorspel	Handeling aan te beveel	Te evalueer	
1 a	1	3	2	1				7
b	1	3	2		2			8
c	1	2		1		1	1	6
2 a	1	5	2	1			1	10
b	2	4	3	1	1	1		12
c	1	4	1	2				8
3 a	1	4	2				1	8
b	2	3	3		1	1		10
4 a (i)	2	3	2				1	8
(ii)	1	2	4		2	1		10
b		2	1				1	4
c		2	2		1			5
d		2	1			1		4
Totaal	13	39	25	6	7	5	5	100

LITERATUURLYS VIR HOOFSTUK 3

1 BLOOM, B.S., HASTINGS, J.T. & MADAUS, G.F. *Handbook on formative and summative evaluation of student learning.* New York, McGraw-Hill Book Co., 1971.

2 CATTELL, R.B. & DREGER, R.M. *Handbook of modern personality theory.* Washington, Hemisphere Publishing Corporation, 1977.

3 EBEL, Robert L. *Essentials of educational measurement.* Englewood Cliffs, Prentice-Hall, 1971.

4 GRONLUND, N.E. *Measurement and evaluation in teaching.* New York, The Macmillan Co., 1965.

5 HICKS, L.E. Some properties of ipsative, normative and forced-choice normative measures. *Psychological Bulletin.* 1970, Vol. 74, No. 3, 167-184.

6 LINDQUIST, E.F. *Educational measurement.* Washington, American Council on Education, 1961.

7 TINKELMAN, S.N. Planning the objective test. In: THORNDIKE, R.L. (ed.) *Educational measurement* (2nd edition). Washington, American Council on Education, 1971.

8 VAUGHN, K.W. Planning the objective test. In: LINDQUIST, E.F. *Educational measurement.* Washington, American Council on Education, 1961.

HOOFSTUK 4
SKEPPING EN SKRYF VAN ITEMS

.1 INLEIDING

.1.1 'n Toets

'n Toets bestaan gewoonlik uit 'n reeks van verwante opdragte wat aan die toetsling gestel en deur hom uitgevoer moet word.

.1.2 'n Item

'n Item verwys na 'n enkele eenheid in 'n toets of subtoets en vorm die kleinste onderdeel (wat in geheel nog sin maak) waarin 'n toets verdeel kan word. 'n Toets kan uit 20, 30 tot 100 - miskien soms meer - items bestaan. Daar kan selfs so min as 10 tot 15 items in 'n toets wees. In die geval van objektiewe toetse bestaan die item weer uit:

- Die stam: Dit is die deel van die item waarin die vraag gevra, 'n probleem gestel of 'n opdrag gegee word.

- Keuses: Hierdie deel van sommige items bevat sowel die regte ant= woord as alternatiewe antwoorde, afleiers genoem, wat definitief ver= keerd is maar vir die proefpersoon wat nie weet waaroor dit gaan nie, tog so aantreklik lyk dat hy enigeen as 'n moontlike antwoord kan aan= vaar.

In sommige items kan vir elke keuse-antwoord 'n aantal punte toegeken word.

.1.3 Skep van items

Om items vir 'n objektiewe toets te skryf, kan gesien word as 'n kuns en slegs deur oefening en ervaring word 'n hoë standaard van "vakmanskap" bereik. Sommige items is eintlik 'n stukkie skeppingswerk op sigself en kan net soos 'n skildery of 'n roman nie volgens 'n vaste formule tot stand gebring word nie. Beginsels en wenke kan egter neergelê word maar dit is slegs die kreatiewe vermoë van die itemskrywer tesame met sy eie oordeel en vermoë tot selfkritiek wat uiteindelik 'n goeie item produseer.

4.1.4 Statistiese tegnieke

Statistiese tegnieke speel 'n belangrike rol by die konstruksie van toetse, veral by itemontleding. Gegewens wat met itemontleding ingewin word, kan wys op sekere gebreke in andersins goeie items. Die skrywer word dan in staat gestel om verbeteringe aan te bring.

4.2 ITEMTIPES

Objektiewe itemtipes val in twee breë groepe naamlik die waar die toetsing

- self die antwoord moet verskaf soos die kort antwoordvorm en
- 'n antwoord uit 'n gegewe aantal antwoorde moet selekteer, soos die waar-vals vorm, die veelkeusige antwoordtipe, afparing en klassifikasie.

4.2.1 Die kort-antwoord vorm

Hierdie vorm word nie baie gebruik nie maar het tog waarde in die sin dat as dit by 'n voorlopige uittoetsing van items gebruik word dit baie daartoe kan bydra om geskikte afleiers vir veelkeusige antwoordtipe-items te vind. Ter illustrasie die volgende paar voorbeelde:

(i) Die vraagvorm

a. Wie het Noord-Amerika ontdek? _____

b. By hoeveel grade Celcius het water sy grootste digtheid? ____

(ii) Voltooiing

a. "Die Profeet" is geskryf deur _____

b. 'n Insek het _____ voelhorings en _____pote.

(iii) Assosiasie

Skryf na elke dorp die provinsie waarin dit gelee is:

Lydenburg _____

Vryheid _____

Winburg _____

Aliwal Noord _____

4.2.2 Die Waar/Vals-vorm

'n Stelling of bewering word gegee en die toetsling moet aandui of dit korrek is al dan nie, deur die toepaslike woord (of soms 'n letter wat 'n verkorting is vir die woord) deur te haal. Soms kan ook gevra word om 'n verkeerde woord deur 'n korrekte een te vervang. Daar is verskil= lende vorms van hierdie tipe vraag:

(i) Waar/Vals

'n Bewering moet as reg of verkeerd aangedui word:

Wanneer 'n vloeistof verdamp neem dit hitte op uit die onmiddel= like omgewing. Waar/Vals

(ii) Ja/Nee

'n Antwoord op 'n direkte vraag moet met òf *Ja* òf *Nee* aangedui word:

Het 'n saag waarmee yster gesaag word groter tande as een waar= mee hout gesaag word? Ja/Nee

(iii) Regte antwoord verskaf

Die toetsling kry die opdrag om 'n verkeerde bewering reg te stel deur 'n onderstreepte woord met die regte woord te vervang:

Die laaste president van die Oranje-Vrystaat was Jan Brand.

A. Paul Kruger
B. F.W. Reitz
C. M.W. Pretorius
D. M.T. Steyn

Die keuses kan hier ook weggelaat word en die leerling moet self die regte naam in 'n oop spasie na die item inskryf. Dan neig dit meer na die kort-antwoord vorm.

(iv) Meer as een bewering

Hierdie vorm bestaan uit 'n onvolledige stam wat op meer as een manier voltooi kan word en elkeen moet as reg of verkeerd aange= dui word:

Die volume van 'n gasmassa

A. neem toe as temperatuur styg en drukking dieselfde bly. Waar/Vals
B. neem toe as drukking vermeerder en temperatuur verminder. Waar/Vals
C. kan konstant gehou word deur drukking te vermeerder en temperatuur te verminder. Waar/Vals

4.2.3 Veelkeusige antwoord-items

Hierdie tipe item is die gewildste vir gebruik in objektiewe toetsing en kan aangepas word om 'n wye veld van onderwerpe te dek. Verskillende vorms hiervan is moontlik. Slegs die belangrikstes word hier aangestip.

(i) Beste antwoord

Meer as een antwoord kan korrek wees maar die leerling moet die beste een uitsoek:

Kunsmis word in die grond ingewerk om

A. die grond te verbeter.
B. beter oeste te verseker.
C. die voggehalte van die grond te verhoog.
D. insekte uit te roei.

(ii) Korrekte antwoord

Hier kan daar net een antwoord as reg beskou word:

Wie was die eerste president van die Republiek van Suid-Afrika?

A. Genl. Louis Botha
B. Adv. C.R. Swart
C. Genl. J.C. Smuts
D. Adv. J.G. Strydom

(iii) Voltooiing

Hier moet 'n bewering voltooi word deur 'n woord of 'n sinsnede by te voeg:

Die belangrikste landbouproduk van die Vaaldriehoek is

A. Koring
B. Graansorghum
C. Mielies
D. Rys

(iv) Die negatiewe vraag

Wanneer 'n vraag meer as een ewe goeie antwoord kan hê, word 'n keuse ingesluit wat nie korrek is nie met die opdrag dat dit die een is wat gevind moet word.

Wie van die volgende was nie 'n eerste minister van die Unie van Suid-Afrika nie?

A. Genl. J.C. Smuts
B. Genl. J.B.M. Hertzog
C. Dr. H.F. Verwoerd
D. Adv. John Vorster

(v) Substitusie

Hierdie metode kan onder andere gebruik word om te toets of 'n leerling die taal korrek en effektief kan gebruik: Daar is baie mense wat dit lyk om laat te slaap.

Wat sal die beste woord(e) wees vir dié wat in bostaande sin onderstreep is?

A. lekker kry
B. dink
C. daarvan hou
D. voel

(vi) Onvolledige response

Wanneer die itemskrywer wil voorkom dat die korrekte antwoord te opsigtelik is kan hy van hierdie metode gebruik maak. Die toetsling word gevra om die eerste letter van 'n eenwoord-respons aan te dui:

Dui die eerste letter aan van die uitvoerproduk wat aan die smal kusstrook van Natal verbou word.

A. k
B. t
C. s
D. l

(vii) Gekombineerde respons

Die stam van die item word gevolg deur 'n aantal afleiers waarvan een of meer korrek kan wees en die toetsling moet dan aandui watter van hulle almal korrek is:

Die kusstrook van Natal is geskik vir die verbouing van suikerriet omdat

i. dit 'n winterreënvalstreek is.
ii. die grond baie vrugbaar is.
iii. dit na aan die see lê.
iv. die reënval laag is.

A. Slegs i
B. ii en iii
C. i en iv
D. Slegs iii

(viii) Afparing of klassifikasie

Hierdie metode kan gebruik word vir identifikasie van name, datums, assosiasies en so meer. In een kolom word 'n aantal onderwerpe ge= noem en in 'n tweede ander onderwerpe (of name) wat daarmee geas= sosieer word. Dit is beter om meer voorwerpe in een van die kolomme te hê omdat die laaste een in die eerste kolom dan nie vanselfspre= kend met die oorblywende een in die tweede kolom afgepaar sal word nie.

Skryf in die spasie voor die woord in die eerste kolom daardie let= ter wat voor die woord in die tweede kolom staan en met die in die eerste kolom geassosieer word.

_____	1. Stad	A. Oranje
_____	2. Berg	B. D.F. Malan
_____	3. Rivier	C. Transvaal
_____	4. Lughawe	D. Johannesburg
		E. Maluti

(ix) Sketse

Sketse kan gebruik word om 'n probleem te stel wat moeilik in woorde beskryf kan word of as 'n effektiewer manier om met die toetsling te kommunikeer. Die sketse kan in die vorm van 'n foto, 'n tekening, 'n diagram, 'n grafiek of 'n kaart wees om se= kere gegewens, situasies of selfs handelswyses duideliker voor

te stel. Dit dra ook daartoe by om die leesvermoë van die toetsling nie te sterk op die proef te stel nie of sekere nie-verbale vermoëns te toets. By items in verband met Geskiedenis, Aardrykskunde, meganiese aanleg, ruimtelike waarneming, en so meer, kan sketse baie nuttig gebruik word.

4.2.4 Die semantiese differensiaal en sy unieke soort item

Die semantiese differensiaal is 'n skalingsmetode wat hoofsaaklik vir die meting van houdings gebruik word. Hierdie metode is deur Charles Osgood geformuleer en verder ontwikkel deur Osgood en sy medewerkers (Osgood *et al.*, 1957). In die literatuur word soms verwys na *die* se=mantiese differensiaal asof dit 'n sekere toets is wat 'n spesifieke stel items bevat. Dit is egter nie die geval nie. Die semantiese differensiaal is 'n algemene metode om sekere tipes inligting in te samel, of anders gestel 'n veralgemeenbare metingstegniek wat aangepas moet word by die vereistes van 'n besondere situasie. Die besondere semantiese differensiaalskaal wat gebruik word is afhanklik van die aard van die eienskap wat gemeet word en die doel van die meting. Dit maak egter wel van 'n eie unieke soort item gebruik wat waarskyn=lik die mees kenmerkende van die metingstegniek is.

Volgens Snider en Osgood (1969) behoort daar vir elke byvoeglike naam=woord of uitdrukking ("phrase") 'n teenoorgestelde te wees, byvoorbeeld waar - vals, goed - sleg, sterk - swak. Hierdie aanname vorm die grond=slag van die metingstegniek.

'n Semantiese differensiaal bestaan uit 'n aantal beoordelingskaalitems ten aansien van een of ander konsep soos byvoorbeeld 'n persoon, voor=werp, situasie, gebeurtenis of groep mense. Elke beoordeling bestaan uit 'n aantal kategorieë of skaalindelings, gewoonlik vyf of sewe.

Voorbeeld:

Vader

sterk	—	—	—	—	—	—	—	swak
hard	—	—	—	—	—	—	—	sag
gelukkig	—	—	—	—	—	—	—	hartseer
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	

Soos in die voorbeeld waar die konsep "vader" beoordeel word, word elke beoordelingsitem (of semantiese skaal) gedefinieer deur polêre byvoeglike naamwoorde wat teenoorgesteld in betekenis is. Die posisies (1) tot (7) word by die instruksies byvoorbeeld soos volg aan die toetslinge gedefineer:

(1) Uiters (bv. sterk)
(2) Taamlik (bv. sterk)
(3) Bietjie (bv. sterk)
(4) Ewe ten opsigte van elke pool of nie die een of die ander
(5) Uiters (bv. swak)
(6) Taamlik (bv. swak)
(7) Bietjie (bv. swak)

Die toetsling word gevra om die betrokke konsep te beoordeel op die volledige stel items. Toetslinge word gewoonlik versoek om die items so vinnig en eerlik as moontlik te antwoord en om nie te lank na te dink oor enige besondere item nie.

Om 'n semantiese differensiaal te kwantifiseer moet gewigte aan elke antwoordposisie van elke item toegeken word. Hierdie gewigte word gewoonlik op die nasiensleutel aangedui teenoor die antwoordposisies van die items op die antwoordblad en met inagneming daarvan of 'n item 'n positiewe of negatiewe betekenis het ten opsigte van die konsep wat in die betrokke veld beoordeel word. Om die totaaltelling vir 'n veld te verkry word die individuele itemtellings bymekaar getel. Wanneer die nasiensleutel van 'n sekere toets wat van die semantiese differensiaal gebruik maak, oor die antwoordblad geplaas word, mag die volgende patroon van gewigte byvoorbeeld waargeneem word (vir 'n geval waar 5 skaalindelings voorkom):

15. ==5== ==4== ==3== ==2== ==1== 16. ==1== ==2== ==3== ==4== ==5==
17. ==1== ==2== ==3== ==4== ==5== 18. ==5== ==4== ==3== ==2== ==1==

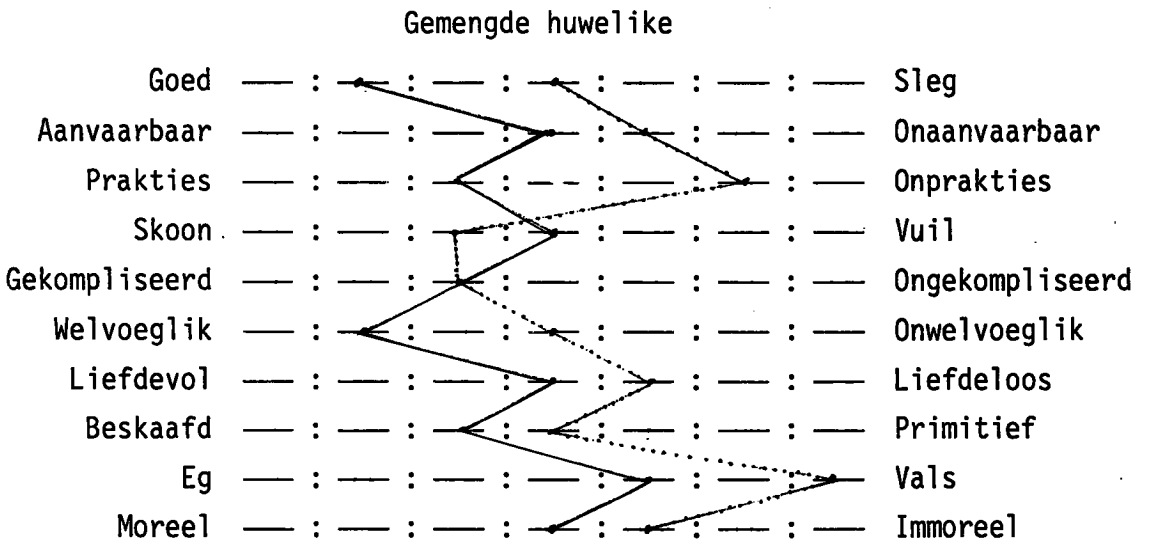
Die som van hierdie vier items se itemtellings is 16. Groepgemiddelde tellings vir elke item kan ook bereken word en in die vorm van 'n profiel voorgestel word (kyk figuur 4.1).

In Osgood en sy medewerkers se ondersoeke is die items vir 'n semantiese differensiaal gewoonlik geselekteer uit 'n lys van 50 pare polêre byvoeglike naamwoorde. In 'n poging om standaardskale daar te stel het sommige navorsers gekonsentreer op Osgood se klassieke lys van 50 woordpare, wat ook deur hulle aan faktorontledings onderwerp is. Hierdeur

kan vergelykings getref word tussen persone of groepe en is daar gehoop om veralgemeenbare houdingskale te kry. Vir baie navorsers was hierdie gestandaardiseerde lys egter onbuigsaam en ontoepaslik vir hulle spesifieke probleme. Dit is dus nodig om toepaslike woord- en uitdrukkinglyste vir 'n spesifieke doel saam te stel.

Voorbeeld:

Indien 'n navorser byvoorbeeld verskillende kultuurgroepe se houding ten opsigte van gemengde huwelike wil bepaal kan die semantiese differensiaal gebruik word. 'n Lys van toepaslike byvoeglike naamwoorde of uitdrukkings moet eerstens saamgestel word en daar moet verseker word dat die items toepaslik is en dat alle aspekte van die konsep (in dié geval gemengde huwelike) wel gedek word. Van die items van dié besondere semantiese differensiaal kan soos volg lyk:



FIGUUR 4.1: VOORBEELD VAN 'N SEMANTIESE DIFFERENSIAAL

Nadat die persone die beoordelingskale voltooi het, kan die gemiddelde tellings vir elke groep ten opsigte van elke item bereken word. Die profiele vir elke groep kan dan soos in figuur 4.1 getrek word. 'n Meer oorsigtelike beeld kan verkry word deur vir elke persoon 'n totaaltelling te bereken deur die itemtellings te sommeer. So 'n totaaltelling kan dan gesien word as 'n meting van die proefpersoon se houding teenoor die betrokke konsep. 'n Hoë telling kan byvoorbeeld dui op 'n positiewe houding en 'n lae telling op 'n negatiewe houding. Gemiddelde totaaltellings kan dan vir verteenwoordigende groepe proefpersone in

die betrokke kultuurgroepe bereken en met mekaar vergelyk word.

Sommige voordele van die semantiese differensiaal is dat dit 'n vinnige doeltreffende metode is, dat dit 'n aanduiding gee van die rigting en intensiteit van menings of houdings van persone. Verder is die tegniek maklik herhaalbaar en skakel dit baie van die probleme (soos byvoorbeeld dubbelsinnigheid en oorvleueling) uit wat by die formulering van vrae ontstaan.

Die Intra- en Interpersoonlike Verhoudingeskaal (IIPS) van die RGN maak gebruik van die semantiese differensiaal tipe item met 5 skaal indelings. Die IIPS bestaan uit vier velde (konsepte) naamlik self=beeld, moeder-kindverhouding, vader-kindverhouding en ideale self. Elke veld bestaan uit 30 teenoorgestelde woordpare in terme waarvan die leerling homself en sy ouers moet beskryf. *Uit sy beskrywing van homself na aanleiding van die 30 gegewe eienskappe (teenoorgestelde woordpare) kan daar bepaal word wat die leerling van homself dink. Het hy 'n positiewe of negatiewe siening van homself? Is sy eie=waarde hoog of laag? Is sy selfsiening realisties of onrealisties?* (Minnaar, 1975).

Volgens Finstuen (1977) is die semantiese differensiaal in byna al die vertakkinge van die sielkunde, naamlik sosiale, persoonlikheids, kliniese, voorligtings, fisiologiese en toegepaste sielkunde, toegepas. Dié tegniek is ook al op baie ander terreine aangewend. Hier word gedink aan taalstudies, kruis-kulturele studies, opvoedkunde, psigofisika, besigheid en in die kunste en politiek.

4.3 WENKE VIR DIE SKRYF VAN ITEMS

4.3.1 Algemeen

a. 'n Item behoort so duidelik, eenvoudig, bondig en korrek as moontlik gestel te word. Die diskriminasievermoë van 'n item word baie verswak as gevolg van onduidelike bewoording. Die moeilikheidsgraad van die item moet uit die probleem wat impliseer word, spruit en nie uit 'n troebel bewoording nie. Onnodige slaggate word vir die toetsling gestel deur onduidelikheid en dubbelsinnigheid. Hoe eenvoudiger die

taal hoe meer is die item 'n toets vir dit wat gemeet moet word.

b. Die bewoording behoort so gekies te word dat hulle regte betekenis geen twyfel laat nie. Ontoepaslike woordkeuse dra baie daartoe by dat die werklike bedoeling met die item onduidelik is.

c. Lomp en komplekse rangskikking van woorde behoort vermy te word. Die sinsbou moet eenvoudig wees. Lang lomp sinne kan liewer in twee of meer kort sinne gestel word.

d. Informasie wat nodig is om die vraag te beantwoord behoort in die stam gegee te word en nie aan veronderstelling oorgelaat te word nie. By= voorbeeld:

Die grootste skade word deur haelstorms aangerig aan

A. lewende hawe.
B. huise.
C. gesaaides.

Elkeen van hierdie keuses kan korrek wees afhangende van die gebied waar die storms voorkom, maar dit word geensins in die stam van die item vermeld nie.

e. Woorde of sinsnedes wat geen bydrae lewer om oor die regte antwoord te besluit nie, behoort gewoonlik nie in die stam ingesluit te word nie. Byvoorbeeld:

Mnr. X, wat 7 kinders gehad het, was in 19 ...

A. President.
B. Eerste Minister.
C. Minister van verdediging.
D. Speaker van die Volksraad.

Die sewe kinders wat hy gehad het dra hoegenaamd niks by tot die kennis van mnr. X se openbare lewe nie en moet weggelaat word.

f. Vermy onakkuraathede in enige deel van die stam. Dit werp 'n refleksie op die itemskrywer se kennis maar mag ook as 'n waarheid in die toetsling se gedagte vasgelê word. Vergelyk byvoorbeeld die volgende item:

Hoekom het Pres. Kruger in 1899 oorlog teen Engeland begeer?

A. Om Engelse uitbreiding teen te gaan.
B. Omdat hy 'n grief teen Engeland gehad het.
C. Om die uitlanders uit sy land te kry.
D. Omdat Engeland nie sy troepe van die landsgrense wou terug trek nie.

Feit is dat Kruger geensins die oorlog begeer het nie, maar dat hy deur omstandighede daartoe gedwing is. Die stam suggereer hier iets wat histories nie korrek is nie en dalk 'n verkeerde idee aan die toetsling kan oordra.

g. Die moeilikheidsgraad van 'n item behoort aangepas te word by die peil van die groep waarvoor dit geskryf word. Hoewel 'n itemskrywer nie maklik die moeilikheidsgraad van 'n item vooruit kan bepaal nie kan iemand wat bekend is met die algemene opvoedkundige peil van die toetsgroep met redelike sekerheid 'n skatting daarvan maak. Hy kan byvoorbeeld bepaalde feite wat hy seker is alle leerlinge behoort te weet, vermy. Die moeilikheidsgraad van 'n veelkeusige of paringsitem kan redelik beheer word deur die afleiers so homogeen moontlik te maak. As die afleiers heterogeen is, is die keuse van die regte antwoord vir die toetsling baie makliker. Vergelyk byvoorbeeld die twee stelle antwoorde op die volgende taalvraag:

Gee 'n sinoniem vir die onderstreepte woord:
Die arme bedelaar het daar maar <u>gehawend</u> uitgesien.

Swak:		Beter:	
A.	Mooi	A.	Verslae
B.	Verflenter	B.	Verflenter
C.	Lelik	C.	Moedeloos
D.	Gesond	D.	Terneergedruk

In die eerste geval sal die oningeligte dikwels makliker die regte keuse maak. Selfs met die verbetering is die item nie ideaal nie aangesien die korrekte antwoord 'n fisiese toestand verteenwoordig en

die afleiers verskillende geestestoestande.

h. Vermy wenke in die stam wat op die regte antwoord kan dui. Die diskri= minasievermoë van die item word daardeur aansienlik verswak en die item word ook baie makliker wanneer die stam wenke bevat wat op die regte ant= woord dui. Vergelyk byvoorbeeld die volgende:

Wat word op 'n diepseeduiker uitgeoefen as hy in die see afduik?

A. Energie
B. Wrywing
C. Druk
D. Gewig

Dis feitlik algemene kennis dat uitoefen met druk gepaard gaan.

i. Vermy geykte uitdrukkings of aanhalings uit 'n boek in òf die stam òf die keuses. Dit bevoordeel die toetsling met 'n goeie geheue waarskynlik onge= regverdiglik en verswak die diskriminasievermoë van die item. Byvoor= beeld:

Chaka word dikwels die swart ..... van Suid-Afrika genoem.

A. koning
B. Napoleon
C. kaptein
D. Hitler

Omdat hierdie uitdrukking dikwels in Geskiedenisboeke gebruik word sal die toetsling wat goed onthou sonder om verder te dink B kies.

4.3.2 Die kort-antwoordvorm

a. Die vraag moet een kort en saaklike korrekte antwoord hê en daar moet geensins die moontlikheid vir dubbelsinnigheid bestaan nie.

Swak: "Die Stem" is gekomponeer deur ...
Enigeen van die volgende antwoorde sou korrek kon wees:
'n Afrikaner, 'n Afrikaanse komponis, 'n predikant.
Beter: Die naam van die komponis van "Die Stem" is ...

b. Geykte uitdrukkings of aanhalings uit 'n boek met weglating van een of meer woorde moet liefs nie gebruik word nie.

Swak: President Brand het altyd gesê: "Alles sal ..."

c. Wanneer 'n sin voltooi moet word, moet daar nie meer as een woord weg=gelaat word nie, anders word die sin maklik onherkenbaar. Daarby moet die blanko spasie liefs aan die end of so na as moontlik aan die end van die sin staan en die spasie wat oopgelaat word moet groot genoeg wees vir die antwoord wat daarin geskryf moet word.

Swak: 'n ... het ... pote.
Beter: 'n Spinnekop het ... pote.

Behalwe dat die toetsling in die eerste geval nie sal weet wat aangaan nie, is die spasie hopeloos te klein om iets daarin te skryf.

4.3.3 Die Waar/Onwaar-vorm

a. Baseer die item op 'n stelling wat altyd waar of onwaar sal wees en nie slegs waar of onwaar na gelang van omstandighede nie.

Swak: Twee hoeke van 'n driehoek is saam gelyk aan 'n regtehoek.

Waar/Onwaar

Beter: In 'n reghoekige driehoek is die som van die twee skerphoeke gelyk aan 'n regtehoek.

Waar/Onwaar

In die eerste geval is die bedoeling seker dat dit onwaar moet wees, maar dit is nie altyd onwaar nie. Onder sekere omstandighede kan dit waar wees soos in die tweede voorbeeld genoem.

b. Woorde soos *altyd, nooit, gewoonlik, kan, soms, slegs, dikwels, geen,* is vir die intelligente leerling 'n leidraad om die waarheid, al dan nie, van 'n stelling te bepaal.

Swak: Die reënval in Nood-Wes-Kaapland is soms hoog.

Waar/Onwaar

Die woordjie *soms* vestig dadelik die aandag op 'n bepaalde tydstip toe daar miskien heelwat reën geval het en dus die item waar kan maak.

Beter: Die gemiddelde reënval in Noord-Wes-Kaapland is hoog.

Waar/Onwaar

c. Lang ingewikkelde sinne moet vermy word. Stel die bewering so kort as moontlik dat dit net een enkele begrip behels.

Swak: Sout los baie maklik in water op en as mens sout in water gegooi het om op te los kan dit weer herwin word deur die water te laat verdamp.

Waar/Onwaar

Beter: Die sout kan uit 'n soutoplossing herwin word deur middel van verdamping.

Waar/Onwaar

d. Vermy die gebruik van sinne uit handboeke of ander bronne as Waar/Onwaar-items. Wanneer sulke sinne alleen staan het dit selde dieselfde beteke=nis as wat dit in die oorspronklike konteks gehad het. Byvoorbeeld:

Die Tweede Wêreldoorlog is in Europa en in die Verre-Ooste gevoer.

Waar/Onwaar

In die verband waarin die sin in die oorspronklike teks gebruik is, mag dit waar gewees het, maar soos hier waar dit uit verband geneem is, is dit onwaar want Noord-Afrika en ander slagvelde of seeslae is nie genoem nie.

e. Nuttige kennis en begrippe van belang moet getoets word en nie onbenul=lige sake nie.

Swak: Genl. Smuts het 'n plaas naby Irene besit.

Waar/Onwaar

Beter: Genl. Smuts het 'n leidende aandeel in die stigting van die VVO gehad.

Waar/Onwaar

Kennis omtrent Genl. Smuts se privaat lewe is onbenullig gestel teenoor die rol wat hy in wêreldsake gespeel het.

f. Die items moet so gestel word dat die antwoorde Waar of Onwaar nie 'n ooglopende patroon vorm of die Waar en Onwaar-items nie op 'n gereëlde manier afwissel nie. Dit sal ook beter wees om 'n paar meer Onwaar-items as Waar-items in te sluit.

4.3.4 Veelkeusige antwoord-items

a. Die stam

(i) Die sentrale probleem moet in die stam gestel word of in die vorm van 'n vraag of as 'n onvoltooide bewering. Sommige idees kan beter in die vorm van onvoltooide bewerings na vore gebring word. Vir ander is die vraagvorm weer die geskiktes.

Swak: Natal ...

A. is 'n beboste gebied.
B. voer koffie van Suid-Amerika af in.
C. voer motors van Kanada af in.
D. is hoofsaaklik gekoloniseer deur koloniste uit Brittanje.

Beter: Die grootste gedeelte van Natal is gekoloniseer deur koloniste van ...

A. Brittanje.
B. Suid-Amerika.
C. Spanje.
D. Italië.

Die stam sou ook in vraagvorm gestel kon gewees het:
Van watter land het Natal die meeste koloniste getrek?

(ii) Sluit in die stam daardie woorde in wat andersins in elk van die keuses herhaal sou moes word.

Swak: Wat is 'n lengtegraad?

A. 'n Denkbeeldige lyn wat rondom die aarde van oos na wes strek.
B. 'n Denkbeeldige lyn wat twee lande skei.
C. 'n Denkbeeldige lyn wat die aardbol in die middel van oos na wes omsirkel.
D. 'n Denkbeeldige lyn wat twee oseane skei.
E. 'n Denkbeeldige lyn wat oor die oppervlakte van die aarde wat van die Noord- na die Suidpool strek.

Beter: 'n Lengtegraad is 'n denkbeeldige lyn wat ...

A. rondom die aarde van oos na wes strek.
B. twee lande skei.
C. die aardbol in die middel van oos na wes omsirkel.
D. twee oseane skei.
E. oor die oppervlakte van die aarde van die Noord- na die Suid= pool strek.

(iii) Indien moontlik vermy negatiewe bewerings en vrae in die stam. Die toetsling is in die reël gewoond daaraan om die korrekte respons uit te soek en mag deurmekaar raak as hy skielik 'n ver= keerde een moet vind.

Swak: Watter van die volgende is nie waar van 'n virus nie?

A. Dit bestaan uit baie groot lewendige selle.
B. Dit kan dit self reproduseer.
C. Dit leef slegs in plant en dierselle.
D. Dit kan siektes veroorsaak.

Beter: Watter een van die volgende is waar van 'n virus?

A. Dit bestaan uit verskillende lewendige selle.
B. Dit kan maklik met die blote oog gesien word.
C. Dit kan siektes veroorsaak.
D. Dit kom slegs in plantselle voor.

(iv) Wanneer 'n item oor 'n omstrede saak gaan is dit wenslik om die outoriteit aan te haal.

Swak: Die grootste Amerikaanse president was ...
Beter: Volgens X was die grootste Amerikaanse president ...

(v) As 'n onvoltooide stam gestel word moet die ontbrekende deel liefs aan die einde daarvan kom. Dit is minder verwarrend vir die toetsling en die leeswerk word verminder deurdat dit nie tweekeer gelees hoef te word nie.

Swak: ... is die vernaamste uitvoerproduk van die Republiek.
Beter: Die vernaamste uitvoerproduk van die Republiek is ...

(vi) Die bewoording van die stam moet nie indirekte wenke bevat wat maklik op een van die keuses kan dui nie. Byvoorbeeld:

Klein verskille tussen organismes van dieselfde soort staan bekend as ...

A. aanpassing.
B. oorerwing.
C. omgewing.
D. veranderings.

Die meervoudsvorme *verskille* en *veranderings* dui dadelik op 'n verband en gee 'n aanduiding van die korrekte antwoord.

b. Die keuses

(i) Al die keuses moet met die bewoording van die stam ooreenkom.

(1) Swak: In watter een van die volgende gevalle veroorsaak haelstorms die grootste skade?

A. Lewende have
B. Dakvensters
C. Gesaaides
D. Bosse

Die keuses wat hier verskaf word is nie *gevalle* nie. Die= selfde afleiers kan behou word maar die stam kan soos volg beter gestel word:

Die grootste verlies wat haelstorms die land as geheel ver= oorsaak is as gevolg van skade aan ...

(2) Swak: Watter proses is presies die teenoorgestelde van foto= sintese?

A. Spysvertering
B. Asemhaling
C. Assimilasie
D. Katabolisme

Hier word 'n vraag gevra wat eintlik geen antwoord het want geeneen van die prosesse is *presies* die teenoorgestelde van fotosintese nie.

(3) Swak: Hoekom het lewende organismes suurstof nodig?

A. Suiwering van bloed
B. Oksidering van afvalstowwe
C. Vrystelling van energie
D. Assimilasie van voedsel

In die stam word 'n rede gevra terwyl die keuses wat gegee word almal prosesse is. Die stam kan net so behou word en die afleiers in redes omskep word soos volg:

A. Om die bloed te suiwer.
B. Om afvalstowwe te oksideer.
C. Om energie vry te stel.
D. Om voedsel te assimileer.

(ii) Maak al die keuses so aantreklik moontlik sodat slegs daardie toetslinge wat werklik weet, die regte keuse sal doen. Hierdie wenk kan uitgevoer word deur die afleiers so homogeen moontlik te maak. Byvoorbeeld:

Die bloedvat wat geöksideerde bloed van die hart na die liggaam vervoer is die ...

A. monnikspier.
B. voorbrein.
C. knieskyfspier.
D. hoofslagaar.
E. rugwerwelkolom.

Hierdie item is ook 'n goeie voorbeeld van keuses wat nie met die bewoording in die stam ooreenstem nie. Selfs die swakste toetslinge sal dadelik weet dat net D die antwoord kan wees want al die ander keuses is nie bloedvate nie. Die keuses kan soos volg verbeter word:

A. okselslagaar.
B. longslagaar.
C. dyslagaar.
D. hoofslagaar.
E. nekslagaar.

(iii) Voorkom dat keuses mekaar oorvleuel. Byvoorbeeld:

Haelstorms veroorsaak die grootste skade aan ...

A. lewende hawe.
B. beeste.
C. skape.
D. gesaaides.
E. huise.

Hier oorvleuel A, B en C mekaar en nie een kan dus as die korrekte antwoord beskou word nie.

(iv) Die gebruik van "nie een van hierdie nie" as 'n laaste keuse moet baie oordeelkundig gebruik word en slegs wanneer een van die eerstes absoluut korrek is en daar nie 'n beste antwoord is nie. Dit is 'n nuttige keuse in wiskundige, spelling- en punktuasie-items. Byvoorbeeld:

Wat is die oppervlakte van 'n reghoekige driehoek waarvan die reghoeksye 30 mm en 40 mm is?

A. 70 mm^2
B. 1 200 mm^2
C. Nie een hiervan nie.

By hierdie tipe item hoef daar nie 'n korrekte keuse onder die eerstes te wees nie.

In gevalle soos die volgende is dié tipe keuse beslis misplaas.

Die eerste president van die Republiek van Suid-Afrika was ...

A. president Kruger.
B. president Diederichs.
C. president Vorster.
D. Nie een hiervan nie.

Die keuse wat gekies word moet in die oop spasie pas. Indien die laaste keuse geneem word lees die sin so: Die eerste pre= sident van die Republiek van Suid-Afrika was nie een van hier= die nie - en dit is onsinnig.

(v) Die keuses moet grammaties by die stam inpas en sover moontlik taalkundig parallel in vorm daarmee wees.

Swak: Die vader het sy seun onterf omdat ...

A. hy was te groot vir sy skoene.
B. hy was opstandig teen die ouerlike gesag.
C. sy lewe was losbandig.
D. hy wou vry wees.

Enigeen van hierdie keuses sal nie die sin vlot laat lees nie.

Beter: Die vader het sy seun onterf omdat hy ...

A. te groot vir sy skoene was.
B. opstandig teen die ouerlike gesag was.
C. 'n losbandige lewe gelei het.
D. vry wou wees.

(vi) Rangskik keuses in logiese volgorde as daar is, mits elke antwoord=posisie min of meer eweveel kere vir die regte antwoord in die toets as geheel gebruik word. As die keuses uit getalle bestaan behoort hulle normaalweg in stygende of dalende orde gerangskik te word, byvoorbeeld:

10, 15, 20, 25 of 25, 20, 15, 10

(vii) Spesifiek bepaalde woorde soos *altyd* of *nooit* moet liefs nie in enige van die keuse gebruik word nie. Die slimmer leerlinge kom gou agter dat daar maar min dinge is wat werklik *altyd* of *nooit* waar is.

c. Die skryf van afleiers

Dit is dikwels moeilik om geskikte afleiers te ontwerp. 'n Paar wenke kan help om hierdie probleem die hoof te bied.

(i) Omskryf die kategorie waarin al die keuses moet val. As in die stam gevra word hoe 'n elektriese yskas die inhoud daarvan verkoel dan kan die kategorie waarin al die keuses behoort te val al daar=die dinge wees wat verkoeling veroorsaak soos water, ys, lug wat sirkuleer, gas wat uitsit.

(ii) Dink aan dinge wat in verband staan met terme wat in die stam ge=noem word.

In die geval van die elektriese yskas kan die woord *elektries* byvoorbeeld lei tot afleiers wat handel oor *vloei van elektri=siteit deur digte gas*, *transformasie van hitte-energie deur elektromagnetiese krag*, en so meer.

(iii) As die vraag 'n kwantitatiewe antwoord vereis moet die keuses dui= delik verskillende punte op dieselfde skaal wees.

As die vraag byvoorbeeld vereis dat een of ander persentasie be= reken moet word dan kan die afleiers so gestel word.

A. 20 persent
B. 40 persent
C. 60 persent
D. 80 persent

(iv) As daar probleme is om geskikte afleiers te vind kan 'n ander benadering in die itemstam probeer word, byvoorbeeld 'n herdefi= niëring van die beginsel wat die item veronderstel is om te toets.

d. Punktuasie van die keuses

(i) As die stam van die item 'n onvolledige stelling is met die oop spasie aan die einde, dan is elke keuse (nie die hele reeks keu= ses nie) 'n moontlike voltooiing van die stelling en moet derhalwe met 'n klein letter begin en met 'n punt eindig. Byvoorbeeld:

'n Roos is 'n ...

A. blom.
B. groente.
C. vrug.
D. voël.
E. dier.

(ii) As die keuses woorde of sinsnedes is wat in 'n oop spasie in die stam ingevul moet word, dan moet elke keuse met 'n klein letter begin en daar is geen punt aan die einde daarvan nie. Byvoor= beeld:

Hulle het die laaste ... aan hom bewys.

A. eerbied
B. ontsag
C. respek
D. eer
E. dank

As die laaste keuse egter "Nie een van hierdie nie" is, dan moet dit met 'n hoofletter begin en met 'n punt eindig.

(iii) As die stam in die vorm van 'n vraag is en elke keuse is 'n volsin, dan behoort elke keuse met 'n hoofletter te begin en met 'n punt te eindig. Byvoorbeeld:

Wat beteken die idioom *die kat die bel aanbind.*

A. Roep die kat.
B. Verrig die moeilike taak.
C. Telefooneer die baas.
D. Lui die klokkie om die bediende te roep.

(iv) As die stam in die vorm van 'n vraag is, maar die keuses is woorde of sinsnedes (nie volledige sinne nie), dan begin elke keuse met 'n hoofletter, maar daar is geen leesteken aan die einde daarvan nie. Byvoorbeeld:

Watter een van die volgende houtsoorte wat gebruik word vir die vervaardiging van meubels word as 'n hardehoutsoort geklassifiseer?

A. Okkerneut
B. Wit denne
C. Rooi seder

As die keuses in die vorm van syfers is, insluitende gewone en desimale breuke, word geen leestekens gebruik nie.

4.3.5 Afparingsitems

a. Gebruik in die twee kolomme waarvan die items met mekaar afgepaar moet word slegs homogene items. Dit wil sê al die items moet tot dieselfde kategorie behoort. Hoe meer heterogeen die items is hoe makliker is dit om die items met mekaar af te paar.

In die onderstaande voorbeeld is heterogene items gebruik en die toetsling kan baie maklik die afparing doen sonder dat hy werklik die kennis het.

Swak: Paar elke item in die eerste kolom af met een in die tweede kolom deur die syfer voor die item in die tweede kolom in die oop spasie voor die betrokke item in die eerste kolom te skryf.

_____	A. Suid-Afrikaanse skrywer	1. Mondfluitjie
_____	B. 'n Snaarinstrument	2. Gé Korsten
_____	C. Afrikaanse sanger	3. Langenhoven
_____	D. Bekende hawestad	4. Viool
_____	E. Musiekinstrument	5. Kaapstad

b. Gebruik relatief kort lyste, liefs nie meer as vyf items nie, vir die eerste kolom. Dit is moeiliker om homogeniteit te behou met lang lyste as met kort lyste. Met korter lyste word minder tyd gebruik met die soek na die korrekte respons.

c. Die twee lyste moet nie soos in die voorbeeld hierbo ewe lank wees nie. As die eerste vier items korrek afgepaar is, spreek dit vanself dat die orige keuse in die tweede kolom by die laaste item hoort.

d. Rangskik die items in die eerste en die tweede kolomme so dat dit vir die toetsling gerieflik voorkom. As daar byvoorbeeld langer en kom=plekse sinne in voorkom, moet dit liefs in die linkerkantse kolom ge=skryf word. Die regterkantse kolom wat die response bevat, moet, in=dien moontlik, liefs uit net een woord bestaan.

e. Die instruksies vir die afparing moet baie duidelik meld op watter voorwaarde die afparing moet geskied. Onthou die toetsling se kennis moet getoets word, nie of hy ingewikkelde instruksies of die beginsel van afparing verstaan nie.

Swak: Paar elke dier in die linkerkantse kolom af met 'n naam in die reg= terkantse kolom. Skryf die betrokke letter in die oop spasie voor die dier in die eerste kolom.

_____	Leeu	
_____	Bok	A. Vleiseter
_____	Vink	B. Graseter
_____	Bees	C. Saadeter
_____	Aasvoël	

Beter: Waarvan leef die diere of voëls wat in die eerste kolom genoem word? Skryf 'n A in die oop spasie in die eerste kolom as jy dink hy is 'n vleiseter, 'n B as hy 'n graseter is en 'n C as hy 'n saadeter is. (Die kolomme en response bly soos hierbo.)

f. Indien daar 'n logiese volgorde is, alfabeties, numeries, chronologies, of andersins, moet sover moontlik die items in hierdie volgorde gerang= skik word. Dit skep die indruk van goeie beplanning, maar is tewens vir die leerling ook tydbesparend as hy sistematies te werk kan gaan. By= voorbeeld:

Skryf in die oop spasie voor die naam in die eerste kolom daardie letter in die tweede kolom wat aandui gedurende watter tydperk die persoon eerste minister van die Unie van Suid-Afrika was.

			A. 1910-1919
23.	_____	J.C. Smuts	B. 1924-1939
24.	_____	J.B.M. Hertzog	C. 1939-1948
25.	_____	J.G. Strydom	D. 1948-1954
			E. 1954-1958

Hierdie vraag kan, soos reeds genoem, ook net so as 'n veelkeusige item gebruik word. Die nommers voor die persone se name dui die nommer van 'n vraag in 'n toets aan en die moontlike keuse is A, B, ensovoorts.

4.3.6 Items in sielkundige toetse

Behalwe die wenke wat in paragrawe 4.3.1 tot 4.3.5 gegee is, volg hier ook nog 'n paar wenke spesifiek in verband met items in sielkundige toetse.

a. Redeneringstoetse

Maak dubbel seker dat daar net een reël is waarvolgens geredeneer kan word, anders kan dit baie maklik gebeur dat 'n item meer as een verdedigbare antwoord kan hê. Die slim leerlinge redeneer dalk anders as wat die itemskrywer in gedagte gehad het en word dan gepenaliseer vir 'n redenasie waarvoor hy eintlik volle krediet behoort te gekry het. Byvoorbeeld:

(i) Na analogie van Banesh Hoffman (1964) se kritiek kan die volgende voorbeeld genoem word:

Watter een van die volgende pas nie by die ander vier nie?

A. Voetbal
B. Korfbal
C. Biljart
D. Hokkie
E. Krieket

Die itemskrywer het bepaald net een aspek in gedagte gehad, maar die volgende is nie so vergesog nie en toon dat elke keuse as 'n moontlike antwoord meriete het:

(1) Voetbal: Die enigste woord wat 'n "v" bevat.
Die enigste spel waar 'n bal geskop mag word.
As 'n toetsling Rugby in gedagte het, die enigste spel wat nie met 'n ronde bal gespeel word nie.

(2) Korfbal: Die enigste woord wat 'n "f" bevat.
Die enigste spel waar die bal slegs met die hande gehanteer mag word.

(3) Biljart: Die enigste binnenshuise spel.
Enigste spel wat nie deur 'n span gespeel word nie.
Enigste spel wat op 'n tafel gespeel word.

(4) Hokkie: Enigste woord wat eindig op 'n klinker.
Enigste woord wat 'n "h" bevat.
Enigste woord met net 6 letters.

(5) Krieket: Enigste woord wat twee "e's" bevat.
Enigste spel waar 'n speler "uitgevang kan word".
Enigste spel waar die telling gebaseer word op die aantal kere wat 'n bepaalde distansie gehard=loop word.

(ii) Vergelyk ook die volgende voorbeeld in verband met getallerye:

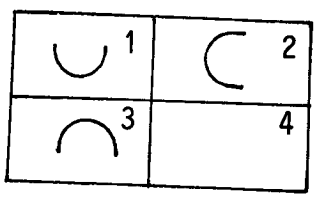
Watter een van die volgende getalle hoort nie by die ander vier nie?

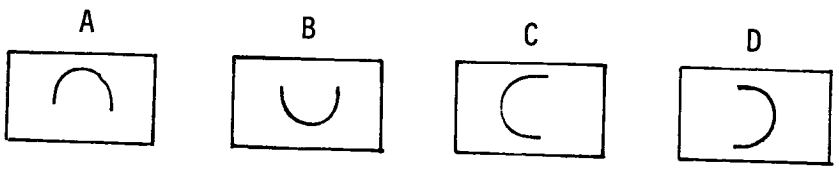
A. 9
B. 25
C. 49
D. 64
E. 83

Dit is duidelik dat die itemskrywer volledige vierkante in gedagte gehad het, maar die volgende getalle het ook meriete:

9: Dit is die enigste eensyfer-getal
64: Dit is die enigste ewegetal

(iii) 'n Patroonvoltooiingstoets





Daar is twee maniere om te redeneer ten einde die vierde blokkie in die boonste figuur te voltooi:

(1) Horisontaal word die figuurtjies kloksgewys geroteer, dus is D die regte antwoord.

(2) Vertikaal is die onderste figuurtjie 'n spieëlbeeld van die boonste een, dus is C die regte antwoord.

b. Kultuurgebonde items

Veral waar daar met verskillende kultuurgroepe gewerk word of met groepe in verskillende streke, moet gedurig gewaak word teen items wat as kul= tuurgebonde beskou kan word. Voorbeelde:

(i) 'n Vraag in verband met die noodsaaklikheid van hengellisensies sal 'n leerling in die binneland kan begryp maar vir die leerling aan die kus bestaan daar dalk nie so iets nie.

(ii) Name van persone of familie-aangeleenthede moet baie oordeelkun= dig benader word. As in 'n item byvoorbeeld 'n Hindoe-vader se kinders Moslemname het sal die Indiërleerling nie weet wat aan= gaan nie en nie sy volle aandag aan die item kan gee nie. Die Swart kind beskou dikwels sy vader se broer ook as sy vader en baie sal net nie die werklike betekenis van "oom" in die verband kan begryp nie.

(iii) Vrae in verband met die kleur van oë kan vir die Indiërleerling baie laat twyfel want hulle ken net bruin oë in hul eie rasse= groep. So maklik word die volgende rekenkundevraag gestel: Tien persent van die leerlinge in 'n klas het blou oë, 15 persent het grys oë, 25 persent het groen oë, die res het bruin oë. Hoe= veel het bruin oë as daar 40 leerlinge in die klas is?

(iv) Vermy liewer vrae waar woorde in verband met godsdiensgebruike en -opvattings in voorkom. Byvoorbeeld die woord preekstoel en loop= paadjie in 'n kerk kan vir die Moslem en die Hindoe vreemde be= grippe wees - hulle het net geen banke of preekstoele in hulle tempels nie. Hou in gedagte dat sekere diere, onder andere die olifant, vir hulle heilig is en geen leed aangedoen mag word nie. Vrae wat enigsins die moontlikheid bevat dat een of ander heilige dier doodgemaak of gevang moet word, moet liefs vermy word.

(v) Sosiale aangeleenthede behoort liefs vermy te word tensy daar 'n deeglike studie van die betrokke kultuurgroep se gewoontes gemaak word. Al die kultuurgroepe het byvoorbeeld nie dieselfde eetgewoontes, tafelgebruike, en so meer nie.

c. Taalgebruik

Wanneer 'n toets nie in 'n leerling se huistaal opgestel word nie, is die gulde reël om die taal so eenvoudig moontlik te hou.

d. Belangstellings- en persoonlikheidsvraelyste

(i) Items moenie sosiaal te aanvaarbaar of onaanvaarbaar wees nie. Items is soms nutteloos (in terme van statistiese vereistes) omdat leerlinge nie werklik in 'n bepaalde beroep belangstel nie, maar tog daaraan voorkeur gee omdat hulle ouers dit beoefen of in die betrokke sosiale kringe beweeg, of ook omdat dit die algemene opvatting is dat daar "baie geld in steek".

(ii) Skryf 'n item so kort moontlik en beperk dit tot net een begrip. Byvoorbeeld 'n item soos "Skryf van wetenskaplike artikels" behels twee begrippe, die taalkundige en die wetenskaplike.

4.3.7 Slot

Die voornemende itemskrywer word aangeraai om die goue reël te volg, naamlik "oefening maak die meester". Hoewel daar met vrug geput kan word uit die ervaring van andere en die beskikbare literatuur van waarde sal wees, kan die waarde van oefening en persoonlike praktiese betrokkenheid nie maklik oorskat word nie.

LITERATUURLYS VIR HOOFSTUK 4

1 ADKINS, DOROTHY C. *Test construction; development and interpretation of achievement tests.* Columbus, Merril, 1924.

2 BRINTON, J.E. Deriving an attitude scale from semantic differential data. *Public opinion quarterly.* 1961, Vol. 25, pp. 289-295.

3 EBEL, ROBERT L. *Essentials of educational measurement.* Englewood Cliffs, Prentice Hall, 1971.

4 EBEL, ROBERT L. Writing the test item. In: LINDQUIST, E.F. *Educational measurement.* Washington, American Council on Education, c1951.

5 FINSTUEN, K. Use of Osgood's semantic differential. *Psychological Reports.* 1977, Vol. 41, pp. 1219-1222.

6 GRONLUND, NORMAN E. *Measurement and evaluation in teaching.* New York, Macmillan, 1965.

7 HOFFMAN, B. *The Tyranny of testing.* New York, Cromwell-Collier and Macmillan, 1964.

8 HUGHES, G. Selecting scales to measure attitude change. *Journal of Marketing Research.* 1967, Vol. 4, pp. 85-87.

9 MINNAAR, G.G. *Handleiding vir die Intra- en Interpersoonlike Verhoudingeskaal.* Pretoria, RGN, 1975.

10 OSGOOD, C.E. The nature and measurement of meaning. *Psychological Bulletin.* 1952, Vol. 49, pp. 226-233.

11 OSGOOD, C.E., SUCI, G. & TANNENBAUM, P. *The measurement of meaning.* Urbana. University of Illinois Press, 1957.

12 SNIDER, J.G. & OSGOOD, C.E. (eds.) *Semantic differential technique: A sourcebook.* Chicago. Aldine, 1959.

13 VIDALI, J.J. Reliability of the semantic differential under practical conditions. *Psychological Reports.* 1976, Vol. 39, pp. 583-586.

14 WESMAN, A.C. Writing the test item. In: THORNDIKE, ROBERT L. *Educational measurement.* Washington, American Council on Education, 1971.

HOOFSTUK 5

TOETSPROGRAMME TYDENS STANDAARDISERING

5.1 VOORLEGGING AAN DESKUNDIGES

Nadat die items geskryf en in 'n toets of toetse saamgestel is en voordat enige daadwerklike toepassing op toetslinge kan plaasvind, is dit wenslik dat die konseptoets(e) voorgelê word aan ander navorsers wat op die betrokke terrein werk. Hierdie navorsers kan 'n belangrike rol speel in die regskaaf van items, wenke gee in verband met bewoording, die geskiktheid van afleiers en die geldigheid van die items beoordeel, en so meer.

Dit is ook wenslik om die konseptoetse aan potensiële gebruikers voor te lê. Meeste toetse, en veral skolastiese prestasietoetse, word deur onderwysdepartemente en opvoedkundige sielkundige klinieke gebruik. Dié organisasies beskik oor deskundiges of panele van deskundiges wat insiggewende kommentaar op die toetse kan lewer. Hulle sal ook kan oordeel of 'n toets aan sy doel sal beantwoord, al dan nie.

5.2 INSAMELING VAN GEGEWENS

5.2.1 Doel

Nadat die kommentaar wat van deskundiges ontvang is, ingewerk en die toetse finaal gereed gemaak is, moet dit op 'n verteenwoordigende steekproef toetslinge van die normpopulasie (die groep waarvoor dit opgestel is) toegepas word met die doel om gegewens in te samel op grond waarvan items verbeter kan word en die geskikste items vir die finale toets geselekteer kan word. Itemseleksie is egter nie noodwendig 'n eenmalige proses wat na 'n eksperimentele toepassing afgehandel word nie. Seleksie begin reeds met die opstel van die toetsmodel. Hoewel daar dan nog nie enige voltooide items is nie, word die grondslag tog reeds gelê waarop die itemskrywer te werk moet gaan om daardie items te skep wat aan die doel met die toets moet beantwoord.

5.2.2 Aard van gegewens

Die gegewens wat met een of meer toepassings ingesamel word, word gebruik om

(a) lig te werp op swak items en items met gebreke, soos 'n stam wat dubbelsinnig gestel is, meer as een korrekte antwoord onder die keuses, twyfel oor die korrekte antwoord, afleiers wat geen doel dien nie, en so meer;

(b) gemiddelde itemtellings of die moeilikheidsgraad van items te be= paal. Hierdie informasie is veral nodig om items so te selekteer dat 'n verdeling van moeilikheidswaardes wat aan die doel met die toets beantwoord, verkry word;

(c) die vermoë van items om tussen goeie en swak leerlinge te onderskei ten opsigte van die veranderlike wat getoets word, te bepaal. 'n Indeks, genoem die diskriminasiewaarde, kan vir elke item uit die gegewens bereken word;

(d) die beplande lengte van die toets te evalueer;

(e) die tydsduur van die toets te kontroleer - of die tydsbeperking voldoende is al dan nie;

(f) leemtes in die toetsaanwysings vir sowel die toetslinge as die toetsafnemers te ontdek en waar moontlik verbeteringe aan te bring;

(g) sydigheid van items vir sekere subgroepe toetslinge te ontdek;

(h) die betroubaarheid en geldigheid van die toets te evalueer.

5.3 WERKSPROGRAM

Dit is nodig dat 'n werksprogram opgestel word om aan te dui op watter datums die verskillende stadia in die standaardiseringsproses afgehan= del moet wees. So 'n program kan soos volg daar uit sien:

WERKSPROGRAM

	Stadium	Moet voltooi wees einde
1.	Skryf van items	Maart 1982
2.	Saamstel en tik van voor= lopige toetse	Augustus 1982
3.	Druk van voorlopige toetse	September 1982
4.	Eerste toepassing	November 1982
5.	Nasien en verwerking van resultate	Maart 1983
6.	Itemseleksie en saamstel van finale toetse	Junie 1983
7.	Druk van finale toetse	September 1983
8.	Toepassing vir normbepaling	November 1983
9.	Nasien en verwerking van resultate	Februarie 1984
10.	Druk van handleiding	Maart 1984
11.	Vrystelling van toets	April 1984

5.4 DIE VERSKILLENDE TOEPASSINGS

5.4.1 Die verkenningstudie

Voordat 'n toets met die oog op 'n volledige itemontleding toegepas word, is dit nuttig om eers 'n verkenningstudie te onderneem. Dit neem die vorm aan van 'n toepassing op 'n klein groepie toetslinge, sê so 10 tot 50, uit die normpopulasie vir wie die toets gestandaardiseer moet word. As die groep verteenwoordigend van die normpopulasie kan wees, soveel te beter, maar dit is nie 'n absolute vereiste nie. Die doel met so 'n verkenningstudie is nie soseer itemontleding nie, maar dit word onder= neem om te bepaal of

(a) die aanwysings vir die toepassing van die toets sonder leemtes is, en duidelik genoeg gestel is;

(b) toetslinge nie taalprobleme ondervind nie;

(c) daar items is wat nie mooi begryp word nie;

(d) die tydsbeperking soos voorlopig bepaal geskik is;

(e) die toets as geheel nie te moeilik of te maklik is nie. Op hierdie stadium kan 'n aantal items nog maklik deur ander vervang word;

(f) die toets of toetsprogram nie te lank en vermoeiend vir die bepaalde toetslinge is nie.

Dit is gewens dat so 'n program deur die toetsopsteller òf self toegepas òf bygewoon word as 'n ander toetsafnemer gebruik word. Baie kan geleer word uit persoonlike gesprekke met 'n toetsafnemer of van die toetslinge self na so 'n toepassing.

Dit mag nie altyd nodig wees om so 'n verkenningstudie te onderneem nie, veral nie wanneer die toetsopsteller reeds wye ervaring het nie. Wanneer 'n toets wat vir een kultuurgroep gestandaardiseer is, vir 'n ander kultuurgroep aangepas moet word, is so 'n verkenningstudie waarskynlik noodsaaklik. Die informasie wat dan ingewin word, sal meer betrekking hê op taalprobleme, kultuurgebonde items, geldigheid van die items vir die betrokke groep, en so meer.

Dit is gebruiklik om ongeveer 1½ keer soveel items as wat in die finale toets moet wees, in die voorlopige toets op te neem sodat die minder geskiktes na 'n itemontledingtoepassing geëlimineer kan word en daar nog voldoende items vir die samestelling van die finale toets beskikbaar sal wees.

5.4.2 Toepassing vir itemontleding

Nadat die meeste probleme met die verkenningstudie uitgestryk is, moet die toets op 'n verteenwoordigende steekproef van 350 of meer toetslinge vir elke normpopulasie toegepas word.

Die toepassing vir itemontleding het ten doel om statistiese gegewens in verband met elke item asook van die toets as geheel te verkry. Die doel hiermee is om itemontleding te kan doen. Dit behels die ontdekking van leemtes in die stamme van die items, die evaluering van die aantreklikheid van die afleiers of die nuttelosheid van sommige en die evaluering van die moeilikheids- en diskriminasiewaardes van die items.

Op grond van dié ontleding word die geskikste items geselekteer vir samestelling van die finale toets(e). Daar kan ook op suiwer statistiese gronde 'n redelike goeie beraming van die betroubaarheidskoëffisiënt van die finale toets gemaak word en deur die seleksie van die geskikste items op 'n bepaalde wyse kan 'n gewenste betroubaarheidskoëffisiënt by benadering vooruit bepaal word. Ervaring het getoon dat hierdie vooruitbeplande koëffisiënt weinig verskil van diè wat met die finale toepassing verkry word.

Indien die finale toets afsonderlik op verskillende groepe gestandaardiseer moet word, byvoorbeeld ouderdoms- of standerdgroepe, dan moet die steekproef so getrek word dat dit ook verteenwoordigend is van elke afsonderlike groep, soos sê 12-jariges, 14-jariges, 16-jariges of leerlinge in standerd 4, standerd 6 en standerd 8. Itemontleding moet dan op elke groep afsonderlik gedoen word. Hoewel daar heelwat probleme aan verbonde kan wees, moet die itemseleksie verkieslik so gedoen word dat dieselfde toets vir al die groepe opgestel word.

Dit kan gebeur dat met so 'n toepassing so baie gebreke in items ontdek word - veral in die geval van 'n totaal nuwe toets - dat heelwat veranderinge aangebring of nuwe items bygeskryf moet word en 'n tweede toepassing vir itemontleding nodig sal wees.

5.4.3 Toepassing vir die bepaling van norms

Nadat die items vir die finale toets geselekteer en die toets saamgestel is, is die volgende fase in die standaardisering van die toets die toepassing vir die bepaling van norms wat ten doel het

(a) die berekening van norms vir die universum as geheel of vir subgroepe afsonderlik soos dit die geval mag wees;

(b) om die betroubaarheid en die geldigheid van die toets te evalueer. Vir laasgenoemde is dit ook nodig om die toetsresultate met 'n buitekriterium te korreleer. (Kyk hoofstuk 11 in verband met geldigheid.);

(c) om weer die aanwysings vir toetsafnemers asook die toetslinge te kontroleer. In hierdie stadium behoort daar egter nie meer wysigings in die toets self aangebring te word nie.

Die toets in sy finale vorm word vir elke normpopulasie op 'n verteen=woordigende steekproef proefpersone toegepas. Daar word aanbeveel dat die getal proefpersone wat nou betrek word ongeveer 400 (selfs meer) vir elke normpopulasie moet wees.

5.4.4 Die handleiding

Die handleiding van 'n toets is 'n dokument waarsonder die toets nie vry=gestel behoort te word nie. Dit bevat die nodige informasie in verband met die ontstaan van die toets, die standaardisering daarvan en aanwy=sings vir die gebruik van die toets asook die normtabelle.

5.5 ADMINISTRASIE VAN TOETSPROGRAMME VIR STANDAARDISERING

5.5.1 Aanwysings vir die toepassing

Die omstandighede waaronder die toetsprogramme vir standaardisering plaasvind, behoort so na as moontlik ooreen te kom met dié waaraan die finale toets gedurende praktiese gebruik onderhewig sal wees. Omge=wingsfaktore is 'n aansienlike bron van foutvariansie by sielkundige en opvoedkundige meetinstrumente. Die aanwysings vir die toepassing van 'n toets kan die betroubaarheid en geldigheid daarvan aansienlik beïn=vloed. Die aanwysings kan egter nie op grond van statistiese ontledings saamgestel word nie en behoort dus vanaf die ontstaan van die toets parallel met die hele standaardiseringsproses ontplooi te word. Hoewel een van die doelstellings met 'n toetstoepassing tydens standaardisering ook die kontrolering van die toetsaanwysings is, behoort alle aanwysings vooraf sorgvuldig uitgewerk te word en so na as moontlik ooreen te kom met dié vir die finale toets.

5.5.2 Die trek van 'n steekproef

In die reël is dit nie moontlik om 'n toets op die hele normpopulasie toe te pas nie. Dit is dus nodig dat vir elke eksperimentele toepas=sing 'n steekproef uit elke groep persone vir wie die toets opgestel word, getrek word.

Slegs wanneer die steekproef verteenwoordigend is, kan betroubare ge=gewens in verband met moeilikheids- en diskriminasiewaardes van items, funksionering van afleiers, betroubaarheid en geldigheid van die toetse

bekom word. Daar word aanbeveel dat die hulp van 'n statistikus inge=roep word vir die trek van die steekproewe.

5.5.3 Toetsafnemers

In baie gevalle is dit haas onmoontlik vir 'n toetsopsteller om self die volledige toetsprogram te behartig en moet daar byvoorbeeld van onderwysers by skole, skoolsielkundiges of ander hulp gebruik gemaak word. Om te verseker dat al die toetsafnemers 'n standaardprosedure tydens die toepassing volg, is dit nodig dat die instruksies wat aan elkeen verskaf word, stap-vir-stap in besonderhede duidelik uiteenge=sit word. In bepaalde gevalle mag dit nodig wees om toetsafnemers eers vooraf op te lei om die toetse te kan toepas.

Elke toetsafnemer behoort versoek te word om verslag te doen oor die verloop van die toepassing en enige onderbreking, onreëlmatigheid, af=wykings van die aanwysings (met rede), of enige ander probleem wat op=geduik het, te noem. Hierdie verslae kan die toetsopsteller in staat stel om die aanwysings te hersien sodat latere toepassings glad kan verloop. Uit persoonlike onderhoude met toetsafnemers en selfs 'n groepie van die toetslinge kan heelwat geleer word.

5.5.4 Antwoorde op alle items

Een van die doelstellings met toetstoepassing vir itemontleding is om betroubare gegewens in verband met elke individuele item in te win en daarom is dit gewens as elke item - ook die laastes - deur nagenoeg al die toetslinge in die steekproef beantwoord word.

a. Kragtoetse

By hierdie tipe toetse val die klem meer op die vermoë van 'n toetsling om 'n vraag te kan doen as op die spoed waarmee hy dit doen. Daarom is dit nodig om voldoende tyd toe te staan sodat die laaste items in die toets deur nagenoeg al die toetslinge beantwoord kan word. Aangesien daar altyd toetslinge sal wees wat stadig werk, moet daar tog in 'n mate 'n tydsbeperking wees. Sommige sal nooit klaar kry nie en onbeperkte tyd kan nie toegestaan word nie. As die toetsing beëindig word, sê, wanneer ongeveer 90 % van die toetslinge die toets voltooi het, sal daar genoeg wees wat die laaste items bereik het om betroubare ontle=

dings daarop te kan uitvoer. As die toetsafnemers dan volgens opdrag in die aanwysings rapporteer hoe lank dit die 90 % toetslinge tydens die toepassing vir itemontleding geneem het om die toets te voltooi, kan van hierdie gegewens bepaal word watter tydsduur daar vir die finale toets toegestaan moet word.

b. Spoedtoetse

'n Probleem ontstaan wanneer die tydsbeperking vir 'n spoedtoets sodanig is dat maar min toetslinge die laaste items bereik. Dit het geen sin om items wat later in 'n spoedtoets opgeneem moet word sonder enige tyds= beperking toe te pas nie. Die omstandighede sal so verskil dat die ge= gewens net nie bruikbaar sal wees nie. 'n Praktiese metode om die ge= skikste tyd vir 'n spoedtoets te bepaal is om vooraf die toets op 'n aan= tal gelykwaardige groepe toetslinge toe te pas maar met verskillende tydsbeperkings vir elke groep. Die toetstyd wat die beste gemiddelde en standaardafwyking vir die toets lewer, is dan die geskikste tydsbe= perking vir die betrokke toets.

5.5.5 Surplusitems

Dit is wenslik dat meer items as wat uiteindelik nodig is vir die finale toets, vir itemontleding toegepas word - sodat daar voldoende beskikbare items vir itemseleksie sal wees. Daar is geen vaste reël oor hoeveel addisionele items geskryf moet word nie, maar in die reël word een en 'n half keer soveel as wat uiteindelik nodig is, as voldoende beskou. Die aantal addisionele items hang van verskeie faktore af soos

(a) die bedrewenheid van die itemskrywer,
(b) die moeilikheidsgraad van die items,
(c) die tipe items, byvoorbeeld redenering, geheue, woordeskat, feite= kennis, en so meer,
(d) of dit 'n nuwe toets is of 'n parallelle vorm van 'n reeds bestaande een,
(e) die aantal kategorieë items in die toets. 'n Taalgebruiktoets kan byvoorbeeld bestaan uit kategorieë vir punktuasie, hoofletters, woordorde, selfstandige naamwoorde, werkwoorde, en so meer.

Dit gebeur dikwels dat daar meer geskikte items as wat uiteindelik nodig is, die itemseleksie oorleef. Sulke items kan tesame met die

nodige itemstatistiek bewaar word vir latere gebruik, soos wanneer 'n parallelle vorm van die toets opgestel moet word.

5.6 REëLINGS NA AFHANDELING VAN DIE TOETSPROGRAM

Met die terugontvangs van die toetsmateriaal is dit nodig om te kontroleer of al die materiaal teruggekom het. Duplikate van lyste wat met die versending na die toetsafnemers opgestel is, kan hier handig te pas kom.

Die bestudering van verslae van toetsafnemers oor die toepassing kan nuttige gegewens oplewer. 'n Vlugtige ondersoek van die antwoordblaaie kan inligting in verband met leemtes in die aanwysings, wat die toetslinge se resultate nadelig kan beïnvloed, oplewer. Voorbeelde:

(a) Blanko antwoordblaaie of te veel onbeantwoorde items kan dui op

(i) gebrekkige aanwysings soos om nie onderaan 'n bladsy te noem dat die toetsling moet omblaai en op die volgende bladsy moet aangaan nie,

(ii) te min aandag aan vooroefeninge en/of

(iii) te min maklike items aan die begin van 'n toets.

(b) Krapwerk in die toetsboekies of selfs op die antwoordblaaie kan dui op die moontlike wenslikheid om papier vir kladwerk te verskaf.

5.7 DIE NASIEN VAN ANTWOORDE

Toetslinge se antwoorde kan in die toetsboekies ingeskryf of op 'n aparte antwoordblad aangetoon word. Indien 'n optiese leser nie beskikbaar is vir die nasien van die antwoordblad nie, sal dit met die hand nagesien moet word. Dit, sowel as die nasien van die antwoorde in 'n toetsboekie kan 'n tydrowende proses wees.

5.7.1 Handnasien

Vir elke bladsy van die toetsboekie kan 'n nasienmasker of -sleutel ontwerp word sodat die antwoorde duidelik deur die openinge daarin sigbaar is. Die antwoordblaaie behoort vooraf deeglik gekontroleer te word vir dubbelresponsies en waar dit voorkom moet al die merkies vir die betrokke item met byvoorbeeld 'n rooi pen doodgetrek word. Sulke items verdien gewoonlik geen punte nie. Deur die tellings vir items bymekaar te tel word die toetsling se totaaltelling vir die toets verkry. Dit kan dan op 'n gerieflike plek op die voorblad van die toetsboekie of boaan die antwoordblad neergeskryf word.

5.7.2 Nasien met 'n optiese leser

Masjinale nasien is slegs moontlik wanneer 'n aparte antwoordblad wat spesiaal vir die doel ontwerp is, gebruik word. Elke antwoordblad sal vooraf eers deeglik gekontroleer moet word vir onnodige potloodmerke en te dowwe merkies. Ook sal die nodige kodes op die antwoordblaaie aangebring moet word en antwoordblaaie wat te veel beskadig is om deur die masjien opgeneem te word, verwyder moet word.

5.8 VERWERKING VAN RESULTATE

As die verwerking van resultate met behulp van 'n rekenaar geskied, is daar geen noemenswaardige verwerkingsprobleme nie. In hoofstuk 6 word 'n metode van itemontleding wat met behulp van 'n rekenaar gedoen kan word, beskryf.

LITERATUURLYS VIR HOOFSTUK 5

1 CONRAD, HERBERT S. The experimental tryout of test materials. In: LINDQUIST, E.F. *Educational measurement.* Washington, American Council on Education, 1951.

2 HENRYSSON, STEN. Gathering, Analyzing, and using data on test items. In: THORNDIKE, ROBERT L. *Educational measurement.* Washington, American Council on Education, 1971.

HOOFSTUK 6
ITEMONTLEDING EN -SELEKSIE

6.1 INLEIDING

Itemontleding en -seleksie is twee prosesse wat onderskei behoort te word. Die eersgenoemde behels die insameling en berekening van soveel gegewens as moontlik in verband met die items van 'n toets, hetsy dit goeie of swak items is, en word in 'n sekere stadium in die standaardiseringsproses onderneem. Itemseleksie aan die ander kant is nie 'n spesifieke proses wat noodwendig op 'n bepaalde tydstip in die konstruksie van 'n toets uitgevoer word nie, maar kan as 'n deurlopende proses beskou word wat reeds begin met die opstel van 'n rasionaal en 'n model vir die toets en sy hoogtepunt bereik met die statistiese interpretasie van die inligting wat met itemontleding ingesamel is en die keuring wat daarop volg. Selfs dan nog kan itemseleksie nie as finaal afgehandel beskou word nie, want na die toepassing vir normbepaling kan daar weer itemontleding gedoen word om te kontroleer of die seleksie op grond van die statistiese gegewens van die eerste toepassing korrek was, al dan nie.

6.2 DOELSTELLINGS VAN ITEMONTLEDING EN -SELEKSIE

6.2.1 Itemontleding

Die doel met itemontleding is om

(a) objektiewe inligting in verband met items te verkry waardeur die subjektiewe oordeel van die toetsopsteller in verband met die keuse van items gekontroleer kan word;

(b) inligting in verband met items in te win sodat die geskikste items geïdentifiseer kan word vir 'n toets wat bepaalde eienskappe ten opsigte van moeilikheidsgraad, betroubaarheid en geldigheid sal besit, en

(c) gebreke in items te ontdek wat reggestel kan word.

6.2.2 Itemseleksie

Met itemseleksie word beoog om

(a) slegs die mees geskikte items vir 'n toets te selekteer en

(b) 'n toets sodanig saam te stel dat dit bepaalde eienskappe van moeilikheidsgraad, betroubaarheid en geldigheid sal besit.

6.3 DIE GEBRUIK VAN ITEMONTLEDINGSRESULTATE

Met behulp van itemontleding kan

(a) tè maklike of tè moeilike items geïdentifiseer en geëlimineer word,

(b) bepaal word watter items die beste tussen goeie en swak leerlinge diskrimineer,

(c) bepaal word waar in 'n item vir gebreke gesoek moet word indien dit nie behoorlik funksioneer nie en watter items verbeter kan word of verwerp moet word en

(d) die kuns van itemskrywing bevorder word deur middel van die kennis wat opgedoen word in verband met persone se reaksies op items en toetse en van die verskillende tipes afleiers.

6.4 BEPERKINGE VAN ITEMONTLEDING

Omdat die skryf van items en die vooraf redigering daarvan miskien eerder 'n kuns as 'n tegniek is, word dit soms verwaarloos ten gunste van die makliker redigeringsmetodes wat met behulp van 'n rekenaar gedoen kan word. Die goeie item kom van die skrywer af en nie van 'n rekenaar nie. Die regte tipe items vir 'n toets word nie slegs deur itemontledingsmetodes verkry nie, maar word veral deur 'n deeglike beplanning en besinning vooraf bepaal. Itemontleding is 'n hulptegniek en kan nie die oorspronklikheid, moeite en vaardigheid of die subjektiewe oordeel van die toetsopsteller vervang nie. Wanneer 'n item op logiese gronde gereg-verdig kan word, dan behoort na gelang die aard van die toets, die subjektiewe opinie deurslaggewend te wees al is die statistiese resultate relatief ongunstig. As egter op logiese gronde daar 'n gebrek in 'n item bespeur word, behoort die item verwerp (of verbeter) te word ten spyte van moontlike gunstige statistiese resultate.

Daar behoort in gedagte gehou te word dat die resultate wat met item-ontleding verkry word onder andere beïnvloed word deur die hele toets waarvan die item 'n deel uitmaak, die wyse waarop dit toegepas word en die groep waarop dit toegepas word.

6.5 ITEMSTATISTIEKE

6.5.1 Die moeilikheidswaarde (p)

a. Definisie

Die moeilikheidswaarde (p) van 'n item word gedefinieer as die gemiddelde itemtelling van die item oor al die toetslinge in die groep.

Dus

$$p_i = \frac{1}{N_t} \sum_{j=1}^{N_t} X_{ij}$$

waar p_i = moeilikheidswaarde van item i,

N_t = getal proefpersone in die steekproef

en X_{ij} = itemtelling van item i vir proefpersoon j.

Wanneer die itemtelling of 0 of 1 kan wees, is die moeilikheidswaarde gelyk aan die verhouding proefpersone wat die item korrek het tót al die proefpersone.

In laasgenoemde geval is die formule

$$p_i = \frac{N_r}{N_t}$$

waar p_i = moeilikheidswaarde van die item,

N_r = getal proefpersone wat die item reg het

en N_t = getal proefpersone in die steekproef.

Hoe hoër die indeks, hoe makliker is die item. Conrad (1945) druk die verhouding uit as 'n persentasie en noem dit dan die maklikheidswaarde van die item.

In plaas van N_t in die noemer kan ook N_a gebruik word (die getal proef= persone wat die item aangedurf het) wanneer die proefpersone nie vol= doende tyd gehad het om die toets te voltooi nie. Wanneer baie proef= persone nie die laaste items bereik het nie kan soms 'n verkeerde beeld van die werklike moeilikheidswaarde gekry word. Omdat meesal slegs die slimste leerlinge die laaste items bereik en dit ook regkry kan N_a die waarde van p effens hoër laat voorkom. Waar N_t gebruik word, kan p weer te laag voorkom omdat van die leerlinge wat nie die laaste items bereik het nie, dit moontlik nog reg kon gehad het as hulle dit wel bereik het. Volgens Conrad sal die gebruik van N_a egter die beste skatting van p gee.

b. <u>Gebruik van die moeilikheidswaarde</u>

Dis nie gewens om bloot volgens moeilikheidswaardes die items vir 'n toets te selekteer nie, omdat daar beter metodes vir seleksie is.

Die indeks p is egter tog van belang:

(i) Na itemseleksie kan die gemiddelde van die nuwe toets bepaal word deur $\sum_{i=1}^{k} p_i$ te bereken, waar k die getal items in die nuwe toets is.

(ii) Dit beïnvloed die vorm van die verspreidingskurwe ten opsigte van die totale toetspunte.

(iii) Met behulp van p kan die standaardafwyking S_i van 'n reg/verkeerd-item (binêre item) bereken word, naamlik $S_i = \sqrt{p_i q_i}$ waar $q_i = 1 - p_i$. Dit lei weer na die variansie, naamlik $p_i q_i$ van die item. Hoe groter dié variansie, hoe beter kan die item potensieel tussen proefpersone onderskei. Die ideale moeilikheidswaarde vir reg/verkeerd-items is 0,5 want dit gee die grootste moontlike item= variansie, naamlik 0,25. Maar 'n p-waarde van 0,5 is nie noodwen= dig 'n waarborg vir 'n hoë item-totaaltelling-korrelasie nie. Ver= der neig raaiery by kognitiewe toetse ook om p-waardes hoër te laat voorkom as wat hulle werklik is, met die gevolg dat 'n item wat die beste diskrimineer se p-waarde ietwat hoër as 0,5 behoort te wees.

c. Korreksie vir raai

Guilford (1966) gee 'n formule waarvolgens die verkreë p-waarde aangesuiwer kan word om die effek van raai te elimineer en gee ook 'n gerieflike tabel met behulp waarvan die aangesuiwerde waarde direk opgesoek kan word, na gelang daar 2, 3, 4 of 5 verskillende afleiers is. Horst (1933) reken dat die persentasie proefpersone wat die antwoord op 'n item reg raai min of meer ooreenkom met die persentasie wat die mees populêre afleier kies. Die korreksie vir raai is ook maar 'n skatting. Die werklike persentasie proefpersone wat die regte antwoord weet sal nooit presies bepaal kan word nie. Om raaiwerk uit te skakel word aan die hand gedoen dat

(i) alle afleiers so aantreklik moontlik gemaak word,
(ii) daar voldoende tyd toegestaan word vir dink en
(iii) proefpersone vooraf teen blindelings raai gewaarsku word.

6.5.2 Die diskriminasiewaarde (r_{it})

a. Die definisie

Die diskriminasiewaarde van 'n item word gedefineer as die produkmoment-korrelasie tussen die itemtelling en die totaaltelling van die toets. Waar items 0 of 1 tel is dit die beste om die puntbiseriaalkorrelasie tussen die item en die totaaltelling van die toets waarvan die item 'n deel is, te bereken met behulp van die formule

$$r_{it} = \frac{\bar{X}_r - \bar{X}_t}{S_t} \cdot \sqrt{\frac{p}{q}}$$

waar r_{it} = die korrelasie van die item met die totaaltelling,

$\bar{X}_r$ = die gemiddelde van die toets van almal wat die item reg het,

$\bar{X}_t$ = die gemiddelde van die toets vir die hele groep,

S_t = die standaardafwyking van die totaaltellings van die hele groep en

p en q reeds vroeër gedefinieer is.

b. Betekenis

Die diskriminasiewaarde is een van die belangrikste indekse in verband met 'n item en dui die mate aan waarin 'n item onderskei (diskrimineer) tussen persone wat baie van die eienskap wat gemeet word, besit en persone wat min van die onderhawige eienskap besit.

Versigtigheid behoort aan die dag gelê te word met die interpretasie van die diskriminasie-indeks. Gestel twee items wat twee effens verskillende vermoëns meet, het presies dieselfde moeilikheidswaarde en diskriminasiewaarde ten opsigte van elke betrokke vermoë. As hulle saam in een toets gebruik word, sal die een wat die vermoë meet wat swaarste in die toets weeg 'n hoër diskriminasiewaarde hê as die item wat 'n vermoë meet wat nie soveel weeg nie.

Byvoorbeeld: 'n Leestoets bevat 100 items waarvan 90 oor feitekennis handel en 10 oor afleidings. Die items oor afleidings sal oor die algemeen dan laer diskriminasiewaardes hê as die wat op feitekennis gebaseer is en gevolglik maklik wegval as vir die finale toets net 50 items geselekteer moet word. Waar daar voldoende items is, is dit dus raadsaam om 'n item ook te korreleer met die totaaltelling van die afdeling waartoe dit behoort, in plaas van slegs met die totaaltelling van die hele toets.

c. Korreksie vir vergrote korrelasie

Wanneer 'n item met 'n totaaltelling van 'n toets waarvan dit deel is gekorreleer word, neig die korrelasie om hoër te wees as wat dit behoort te wees, veral wanneer die toets nie uit baie items bestaan nie. Die ideaal is om verskillende totaaltellings vir 'n toets te bereken waar elke keer 'n ander item uitgesluit is en dan die betrokke items met daardie totaal te korreleer. Guilford (1966) gee 'n formule wat gebruik kan word om die diskriminasiewaarde aan te suiwer, maar as 'n toets meer as 50 items bevat is die korreksie ook nie eintlik nodig nie.

6.5.3 Die itemkriteriumkorrelasie (r_{ik})

Items kan ook met 'n buite-kriterium gekorreleer word om te bepaal of die items dieselfde vermoë meet as die betrokke kriterium.

6.5.4 Ander indekse

Uit die standaardafwyking, die diskriminasiewaarde en die geldigheids= waarde van 'n item word twee nuttige indekse verkry:

a. Gullikson se produk ($r_{it}\ s_i$)

Die produk van die diskriminasiewaarde r_{it} en die standaardafwyking S_i van 'n item gee 'n indeks wat as Gulliksen se produk bekend staan. Hoe groter hierdie produk is hoe meer geskik is die betrokke item.

b. Die itemgeldigheidsproduk ($r_{ik}\ S_i$)

Hierdie indeks word verkry deur die item-kriteriumkorrelasie r_{ik} met die itemstandaardafwyking s_i te vermenigvuldig.

Hierdie twee indekse is van belang wanneer daar na finale seleksie meer inligting in verband met die finale toets in sy geheel, nodig is.

6.5.5 Itemstatistieke vir subgroepe proefpersone

a. Die proefpersone wat 'n item oorgeslaan het

Hierdie subgroep is dié wat geen respons op 'n bepaalde item gegee het nie, maar wel op items daarna.

Vyf nuttige statistieke in verband met dié subgroep kan vir elke item bereken word, naamlik

(i) die aantal proefpersone wat die item oorgeslaan het,
(ii) die persentasie wat hulle uitmaak van die subgroep wat die item aangedurf het,
(iii) die persentasie wat hulle van die totale groep uitmaak,
(iv) die gemiddelde van die hele toets vir dié subgroep en
(v) die afwyking van hulle gemiddelde in die hele toets vanaf die gemiddelde van die hele groep.

Aan die begin van 'n kognitiewe toets is die persentasie proefpersone wat geen respons op 'n item kan gee nie gewoonlik klein. Die gemiddelde pres= tasie in die toets dui aan of dit 'n goeie of swak subgroep is wat nie

die item gedoen het nie. As die aantal persone wat 'n item oorgeslaan het en die gemiddelde toetstelling vir hierdie groep redelik hoog is, kan dit 'n aanduiding wees dat die item moontlik te vroeg in die toets geplaas is.

b. Die proefpersone wat nie 'n item bereik het nie

Dit is daardie subgroep wat nie 'n respons op 'n item gegee het nie en ook geen respons op enige item daarna nie.

Dieselfde inligting as by die vorige subgroep kan ook hier vir elke item bereken word, sodat afgelei kan word watter tipe proefpersone nie in staat was om die toets te voltooi nie.

c. Informasie in verband met die subgroepe wat die verskillende itemresponse gegee het (afleiers gekies het)

Dieselfde tipe inligting soos reeds hierbo genoem, kan vir elke item afsonderlik bereken word vir die subgroepe wat die alternatiewe itemresponse gekies het.

Indien 'n item se response korrek geskaal is (die puntetoekenning vir response korrek is) behoort daar 'n positiewe korrelasie tussen die itemtelling en totaaltelling vir die toets te wees. Dit beteken dat hoe groter die itemtelling vir 'n bepaalde itemrespons is hoe groter moet die gemiddelde toetstelling wees van die subgroepie toetslinge wat die betrokke itemrespons (afleier) gekies het. Indien empiries gevind word dat hierdie voorwaarde nie geld nie, is dit feitlik seker dat daar iewers 'n tekortkoming in die item is. So 'n tekortkoming kan te wyte wees aan swak bewoorde afleiers, dubbelsinnighede in die stam en/of afleiers, 'n swak skalingstegniek, 'n item wat nie in die betrokke veld tuisbehoort nie, onverwagte kultuurinvloede, foutiewe inligting of veronderstellings in die stam en/of afleiers, en so meer.

6.6 INFORMASIE IN VERBAND MET DIE TOETS AS GEHEEL

Benewens die reedsgenoemde statistieke is dit nuttig om ook die betroubaarheidskoëffisiënt en die standaardmetingsfout van die betrokke toets, asook die gemiddelde en die standaardafwyking van die toets en elke kriterium wat vir die berekening van diskriminasiewaardes gebruik is, te bereken.

6.7 WENKE VIR ITEMSELEKSIE

a. Die itemstatistieke kan bestudeer word om te bepaal of daar items is wat nie goed funksioneer nie. As 'n groep wat 'n verkeerde antwoord gegee het se gemiddelde in die toets hoër is as dié van die groep wat die regte antwoord gegee het, dan skort daar iewers iets met die item. Waar 'n verkeerde antwoord deur 'n groot aantal proefpersone gekies word, is dit miskien 'n baie goeie afleier op voorwaarde dat die gemiddelde van die subgroep wat hierdie antwoord gegee het, laer is as dié van die subgroep wat die korrekte antwoord gegee het.

b. 'n Binêre item ('n item waarvan die telling slegs 0 of 1 kan wees) in 'n kognitiewe toets funksioneer in die reël bevredigend as die toetsgemid=deldes van subgroepe toetslinge in die volgende volgorde van hoog na laag gevind word:

(i) dié wat die regte antwoord gegee het,
(ii) dié wat nie daarin kon slaag om 'n regte antwoord te gee nie,
(iii) dié wat die item oorgeslaan het en
(iv) dié wat die item nie bereik het nie.

Waar raaiery egter ook 'n rol speel, kan die orde soms omgeruil wees. Dit is egter belangrik dat die gemiddelde toetstelling van die groep wat die regte antwoord gegee het hoër as die ander groepe se gemiddel=de toetstellings moet wees.

c. Die bogenoemde gemiddeldes gee in die reël 'n aanduiding van waar daar 'n tekortkoming in die item kan wees. Die hele item, stam sowel as af=leiers, kan dan ondersoek word om 'n verklaring te probeer vind. Mis=kien kan die fout herstel en die item in 'n baie nuttige item omskep word.

d. Wat aanleg-, bekwaamheid- en intelligensietoetse betref sal 'n item swak wees as:

(i) 'n Alternatiewe antwoord op logiese gronde moontlik is - al gee slegs een proefpersoon daardie antwoord. Die gemiddelde van die subgroep wat daardie antwoord gegee het, sal in die meeste gevalle die aandag daarop vestig. Die slim leerling dink aan so 'n antwoord en dit is onbillik teenoor hom om die item so te behou.

(ii) 'n Alternatiewe antwoord wat nie op bekende logiese gronde verdedigbaar is nie tog deur 'n subgroep leerlinge wat 'n hoër gemiddelde telling in die toets as enige ander subgroep behaal, gekies word.

Wat skolastiese prestasietoetse betref, is hierdie reël nie so streng van toepassing nie, want ander faktore kan miskien 'n rol speel, byvoorbeeld swak onderrig, die sillabus wat nog nie deurgewerk is nie, en so meer.

e. Normaalweg is dit voldoende om items te selekteer op grond van die diskriminasiewaardes. By 'n prestasietoets is dit egter raadsaam om items te korreleer met die totaaltelling van die betrokke afdeling van die sillabus waartoe die item behoort en dan op grond van daardie korrelasie slegs die items vir die betrokke afdeling te selekteer. Items met 'n diskriminasiewaarde van laer as 0,25 behoort slegs by wyse van uitsondering ingesluit te word. By prestasietoetse is die itemontleding in 'n mate ondergeskik aan die inhoudsgeldigheid. Daar behoort moeite gedoen te word om te verseker dat die items inhoudsgeldig is voordat dit aan itemontleding onderwerp word. Die ontleding voorsien addisionele informasie maar die beslissing om 'n item in te sluit of te verwerp bly onderworpe aan die oordeel van die toetsopsteller. So kan byvoorbeeld 'n paar maklike items aan die begin van 'n toets ingesluit word om die swak leerlinge nie uit die staanspoor te ontmoedig nie, en ook miskien 'n paar moeilike items aan die einde ingevoeg word ter wille van die skrander leerling.

f. Aangesien die beste diskriminasiewaardes gewoonlik verkry word met items wat 'n moeilikheidswaarde in die omgewing van 0,5 het, sal die kans goed wees om 'n simmetriese verdeling van totaaltellings te verkry. Indien die gewenste vorm nie verkry word nie, kan dit reggestel word deur moeiliker items in plek van makliker items in te voeg, of omgekeerd, na gelang die toets te maklik of te moeilik is - of die verdeling negatief of positief skeef is. Diskriminasiewaardes van die betrokke items moet egter in gedagte gehou word.

g. Na die items geselekteer is en sekere items dus weggeval het, kan skattings gemaak word van die statistiese eienskappe van die finale toets. Die volgende word aan die hand gedoen:

(i) Die gemiddelde

Die gemiddelde $\bar{X}$ van die nuwe toets kan geskat word deur die moeilikheidswaardes van al die geselekteerde items bymekaar te tel.

Dus $\bar{X} = \sum_{i=1}^{k} P_i$ (k = die getal items in die nuwe toets.)

(ii) Die standaardafwyking

Die standaardafwyking S_x word geskat deur 'n blote optelsommetjie te maak: Tel al die geselekteerde items se Gulliksonprodukte bymekaar.

Dus $S_x = \sum_{i=1}^{k} r_{it} s_i$.

'n Verklaring vir die simbole word by (iii) hieronder gevind.

(iii) Die betroubaarheidskoëffisiënt

'n Aanduiding van die K-R 20 betroubaarheidskoëffisiënt r_{tt} van die nuwe toets kan met behulp van die volgende formule gevind word:

$$r_{tt} = \frac{k}{k-1} \cdot \left\{ 1 - \frac{\Sigma S_i^2}{(\Sigma r_{it} S_i)^2} \right\}$$

waar k = aantal items in die finale toets,

S_i = standaardafwyking van item i van die finale toets en

r_{it} = diskriminasiewaarde van item i soos vroeër in die langer toets bereken is

(iv) Die seleksieproses kan herhaal word totdat die verlangde resultate verkry word. Die betroubaarheid kan dus gekontroleer en moontlik opgeskuif word totdat die toetsopsteller daarmee tevrede is.

h. Slotopmerking

Na die toepassing van die toets vir normbepaling kan daar weer 'n item=ontleding gedoen word om te bepaal of die items nou na verwagting funk=sioneer. Die gemiddelde, die standaardafwyking en die betroubaarheids=koëffisiënt van die toets in die finale toepassing kan ook vergelyk word met die statistieke wat volgens die metodes van paragraaf g. bereken is, om 'n aanduiding te kry van hoe geslaagd die itemseleksie was.

LITERATUURLYS VIR HOOFSTUK 6

1 AHMAND & GLOCK. *Evaluating pupil growth. Principles of test measurement.* Boston, Allyn and Bacon, 1967.

2 ANSTEY, E. "The d-method of item-analysis." *The British Journal of Psychology (Statistical Section).* 1948, 167-177.

3 CONRAD, H.S. "Characteristics and uses of item-analysis data." *Psychological monographs*, 62(295), 1945, 1-48.

4 DAVIS, FREDERICK B. "Item selection techniques." In: LINDQUIST, E.F. (ed.) *Educational measurement.* 266-328. Washington, American Council on Education, 1963.

5 EBEL, ROBERT L. *Measuring educational achievement.* Englewood Cliffs, New-Jersey, Prentice-Hall, 1965.

6 GULLIKSON, HAROLD. *Theory of mental tests.* New York, John Wiley and Sons, 1962.

7 GUILFORD, J.P. *Psychometric methods.* New York, McGraw-Hill, 1966.

8 GUILFORD, J.P. *Fundamental statistics in psychology and education*, 3rd ed., New York, McGraw-Hill, 1956.

9 GUTTMAN, L. Measurement as structural theory. *Psychometrica.* 1971, 36, 329-347.

10 HORST, P. "The difficulty of a multiple choice test item." *Journal of Educational Psychology.* 1933, 229-244.

11 LORD, F.M. & NOVICK, M.R. *Statistical theories of mental test scores.* Reading, Muss., Addison-Wesley, 1968.

12 NUNNALLY, JUM C. *Educational measurement and evaluation.* New York, McGraw-Hill, 1964.

13 NUNNALLY, JUM C. *Psychometric theory.* New York, McGraw-Hill, 1967.

14 NUNNALLY, JUM C. *Introduction to psychological measurement.* New York, McGraw-Hill, 1970.

HOOFSTUK 7
NORMBEPALING

7.1 INLEIDING

Die skalingstegnieke wat by die meeste psigometriese toetse gebruik word, lewer metings wat tussen die ordinale- en intervalvlak van meting lê. In hoofstuk 1 is gemeld dat enige transformasie van die onderliggende skaal waarin ordinale metings gemaak is (die roupuntskaal), toelaatbaar is so= lank die transformasie nie die verkreë rangordes van toetslinge versteur nie.

In die algemeen is die betekenis of relatiewe waarde van punte (toets= tellings) op 'n roupuntskaal nie eksak en duidelik interpreteerbaar nie. Daar kan darem wel gesê word dat 'n hoër punt meer van die onderhawige eienskap impliseer as 'n laer punt. As die gemiddelde en standaardafwy= king van die roupunte beskikbaar is, kan effens meer gegewens uit 'n en= kele roupunt afgelei word.

Aangesien die verdeling van roupunte vir die normpopulasie byna nooit in 'n wiskundige formule gegee kan word nie, kan nie veel van die addisionele inligting van die gemiddelde en die standaardafwyking van die roupunte afgelei word nie. Daar is dus 'n behoefte daaraan om toetspunte met een van die talle toelaatbare tipes transformasies na so 'n skaal te trans= formeer dat die toetsinterpreteerder makliker sal weet hoe "lyk" die toetslinge wat 'n bepaalde punt op die nuwe skaal het. So 'n nuwe skaal word 'n normskaal genoem.

7.2 NORMSKALE

Normskale word so gekies dat

(a) die verdeling van normpunte in die normpopulasie (dit is enige goedgedefineerde populasie toetslinge) deur 'n gerieflike wiskun= dige formule beskryf kan word,

(b) of so dat die normpuntskaal skynbaar op dieselfde skaal lê as 'n ander "bekende" veranderlike wat vir elke lid van die normpopula= sie gedefinieer is.

Die statistiese eienskappe van die normpunte in die normgroep voorsien dus 'n standaard vir die evaluering van enige roupunt nadat die roupunt na 'n normpunt getransformeer is.

7.3 DEFINISIE VAN NORMBEPALING

Kortliks kan gesê word dat normbepaling die proses is waarin die transformasietabel gevind word waarmee roupunte van 'n bepaalde toets vir elke lid van die normpopulasie na normpunte getransformeer kan word. Hierdie *transformasietabel* word dikwels 'n *normtabel* genoem en soms ook losweg *norms*. Normbepaling kan ook gesien word as 'n standaardiserings- of ykproses om 'n algemeen aanvaarbare standaard waarteen 'n roupunt geëvalueer kan word, te verkry.

In die proses van normbepaling is dit nodig om vir elke lid van die normpopulasie 'n roupunttelling te verkry. Aangesien dit byna altyd onmoontlik is, kan volstaan word deur die roupunte vir elke lid van 'n verteenwoordigende steekproef persone uit die normpopulasie te verkry. So 'n verteenwoordigende groep persone word die *normgroep* genoem.

7.4 DIE KEUSE VAN NORMPOPULASIES

'n Noodsaaklike aanname by alle psigometriese toetsing, maar een wat nie te dikwels eksplisiet gestel word nie, is dat metings met dieselfde instrument vir twee persone slegs vergelykbaar kan wees indien die twee persone ten minste *potensieel* dieselfde lewenservarings kon gehad het. In ander woorde kan gesê word dat dié deel van die twee persone se kultuuragtergrond wat hulle toetstellings betekenisvol kan bepaal naastenby eenders behoort te wees. Die woord *kultuur* is hier in die wye betekenis van die woord gebruik. 'n Verskil in toetstellings kan dan geïnterpreteer word as 'n verskil in die eienskap wat die toets veronderstel is om te meet en nie bloot 'n kultuurverskil tussen die twee toetslinge nie.

Die aanname van gelyke potensiële lewenservarings kan vir 'n groep persone ondersoek word deur te kyk of sekere kultuurverwante veranderlikes soos ouderdom, geslag, etniese groep, skoolstanderd, en so meer betekenisvolle bydraes lewer tot die variansie van toetstellings in die groep. Indien so 'n kultuurverwante veranderlike wel 'n betekenisvolle bydrae tot die variansie van toetstellings lewer, is dit waarskynlik 'n aanduiding

dat die groep te heterogeen vir 'n normgroep is. Algemeen gesproke is dit wenslik om normpopulasies sodanig te kies dat kultuurverwante veranderlikes 'n relatiewe klein bydrae sal lewer tot die variansie van toetstellings binne die normpopulasie. In hoofstuk 9 word kultuurverwante veranderlikes wat 'n betekenisvolle bydrae tot toetsvariansie kan lewer, *steuringsveranderlikes* genoem. 'n Voorbeeld van 'n steuringsveranderlike word in dié hoofstuk bespreek. Die probleem van steuringsveranderlikes is verwant aan die probleem van veralgemeenbaarheid wat in hoofstuk 10 bespreek word.

7.5 TIPES NORMSKALE

7.5.1 Normskale wat op die normaalverdeling gebaseer is

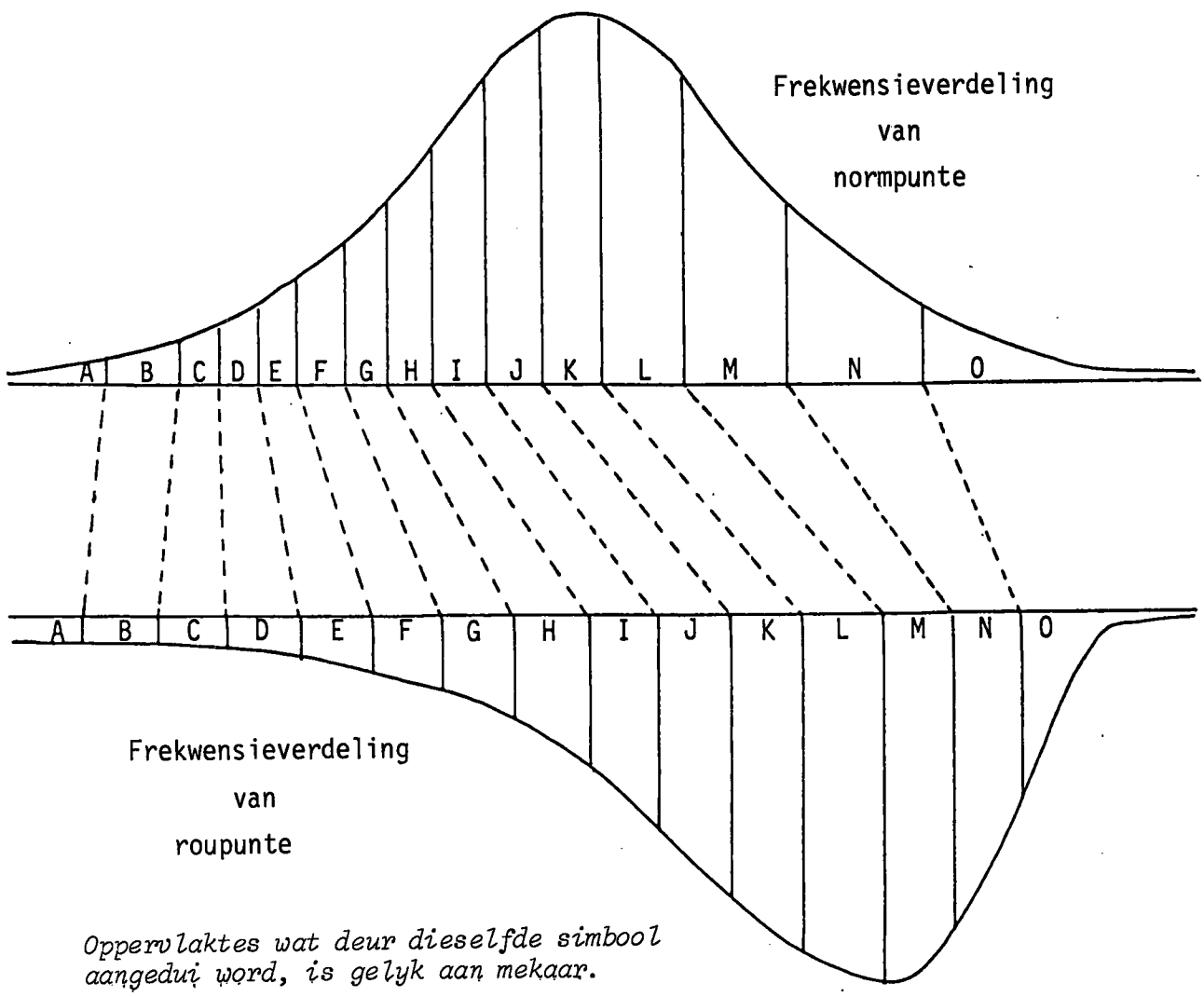
In hierdie tipe normskaal is die transformasie van roupunte na normpunte sodanig dat die *normpunte* in die normpopulasie normaal sal verdeel met 'n gekose gemiddelde en standaardafwyking. Die volgende skale word redelik algemeen gebruik:

Skaal	Omvang	Gemiddelde	Standaardafwyking
Stanegeskaal	1 tot 9	5	2
Stienskaal	1 tot 10	5,5	2
T-skaal	$\pm$ 20 tot $\pm$ 80	50	10
IK-skaal	$\pm$ 55 tot $\pm$ 145	100	15
Stavyfskaal	1 tot 5	3	1

Aangesien gewoonlik veronderstel word dat die roupunte op 'n kontinue skaal lê, volg dat die transformasie tussen roupunte en normpunte ook kontinu behoort te wees. Normpunte in bogenoemde skale word egter altyd en roupunte gewoonlik in heelgetalle uitgedruk. In die normskaal word normpunte dus afgerond tot die naaste heelgetal. Afgesien van die afronding van die normpunte na heelgetalle, word normpunte in sommige skale soos die stanege-, stien- en stavyfskaal verder tot 'n vaste minimum en maksimum beperk. Vir hierdie skale word die normaalverdeling dus op sy eindpunte effens saamgedruk. Dit is om hierdie rede dat die werklike standaardafwyking van die stanegeskaal 1,96 is en nie 2 soos in die definisie van die skaal gebruik is nie. Wanneer die rouputverdeling ver van 'n normaalverdeling afwyk, is dit nie moontlik om 'n transformasie van roupunte na normpunte te vind nie hoofsaaklik vanweë die feit dat slegs die heelgetalle van albei skale

gebruik word. In sulke gevalle is dit nie wenslik om 'n normskaal wat op die normaalverdeling gebaseer is, te gebruik nie.

Dit is interessant om by hierdie tipe normskaal te let op die aard van die transformasie van roupunte na normpunte. Wanneer die roupuntverdeling in die normpopulasie presies normaal is, sal die kontinue transformasie 'n lineêre transformasie wees. Die roupunte verdeel egter gewoonlik nie presies normaal in die normpopulasie nie. In sulke gevalle sal daar 'n nie-lineêre transformasie van roupunte na normpunte wees. In figuur 7.1 word 'n voorbeeld van so 'n nie-lineêre transformasie tussen roupunte en normpunte voorgestel.



FIGUUR 7.1: 'N GRAFIESE VOORSTELLING VAN DIE NORMALISERING VAN 'N VERDELING VAN TOETSPUNTE

7.5.2 Die persentielrangskaal

In die persentielrangskaal word die persentielrang wat by 'n gegewe roupunt (gewoonlik word aanvaar dat die roupunt op 'n kontinue skaal lê) hoort gedefinieer as die persentasie persone in die normpopulasie wat 'n laer roupunt as die gegewe roupunt behaal het. Die persentielrangskaal strek dus van 0 tot by 100. In die praktyk word persentielrange gewoonlik tot die naaste heelgetal benader.

7.5.3 Die ouderdomskaal (toetsouderdom)

In die ouderdomnormskaal word 'n normpunt wat by 'n gegewe roupunt hoort die toetsouderdom van die roupunt genoem. Die toetsouderdom van 'n gegewe roupunt word gedefinieer as die ouderdom van die *homogene* ouderdomsgroep uit die normpopulasie wat 'n *gemiddelde roupunt* gelyk aan die gegewe roupunt het. Wanneer norms op 'n ouderdomskaal bereken word, is dit 'n vereiste dat die normpopulasie relatief heterogeen ten opsigte van ouderdom moet wees.

7.5.4 Die standerdskaal (toetsstanderd)

Hierdie skaal is soortgelyk as die ouderdomskaal. Die toetsstanderd van 'n gegewe roupunt word gedefinieer as die homogene standerdgroep (elke standerd kan in 4 kwartale of selfs in 12 maande verdeel word) uit die normpopulasie wat 'n gemiddelde roupunt gelyk aan die gegewe roupunt het. Die normpopulasie moet relatief heterogeen ten opsigte van skoolstanderd wees.

7.6 STATISTIESE VERWERKINGS MET NORMPUNTE

Statistiese hipotesetoetsing berus dikwels op die aanname dat die betrokke veranderlikes normaal verdeel of ten minste naastenby normaal verdeel. Aangesien punte op die persentielrangskaal 'n reghoekige frekwensieverdeling het en punte op die ouderdom- en standerdskaal waarskynlik ook naastenby reghoekige verdelings sal hê, behoort versigtigheid aan die dag gelê te word wanneer statistiese verwerkings met die oog op hipotesetoetsing met punte op hierdie skale gedoen word.

7.7 DIE VERBAND TUSSEN DIE PERSENTIELRANGSKAAL EN DIE NORMSKALE WAT OP DIE NORMAALVERDELING GEBASEER IS

Aangesien daar gewoonlik veronderstel word dat die eienskappe wat ons met toetse meet op 'n kontinue skaal lê, sal 'n heelgetal normpunt X op 'n skaal met gemiddelde $\bar{X}$ en 'n standaardafwyking S enige normpunt in die interval $(X - \frac{1}{2}, x + \frac{1}{2})$ verteenwoordig. Hierdie interval sal ook 'n sekere persentielrangomvang verteenwoordig. Hierdie persentielrangomvang kan bepaal word soos in die volgende voorbeeld verduidelik word:

Gestel X = 3 op die stanegeskaal.
Dan is $\bar{X}$ = 5 en S = 2,00
Die onderste grens van 'n stanege van 3 = 2,5.
Die boonste grens van 'n stanege van 3 = 3,5.
Die z-waarde van die onderste grens = (2,5 - 5)/2,00 = -1,25.

Die z-waarde van die boonste grens = (3,5 - 5)/2,00 = -0,75.

Die oppervlakte onder die normaalkromme links van die onderste grens (z = -1,25) = 0,1056(10,56 %).

Die oppervlakte onder die normaalkromme links van die boonste grens (z = -0,75) = 0,2266(22,66 %).

Die persentielrangomvang van 'n stanege van 3 is dus van 10,56 tot 22,66.

In tabel 7.1 word die boonste persentielranggrens vir elke stanegepunt gegee.

TABEL 7.1

BOONSTE PERSENTIELRANGGRENSE VAN ELKE STANEGEPUNT

Stanege	Persentielrang van die boonste grens	
	Tot 2 desimale syfers	Benader tot die naaste heelgetal
1	4,01	4
2	10,56	11
3	22,66	23
4	40,13	40
5	59,87	60
6	77,34	77
7	89,44	89
8	95,99	96
9	100,00	100

.8 VOORBEELD VAN DIE BEPALING VAN DIE TRANSFORMASIETABEL VAN ROUPUNTE NA PERSENTIELRANGE

In tabel 7.2 word die frekwensieverdeling van die roupunte in 'n reken= toets vir 450 leerlinge gegee.

TABEL 7.2

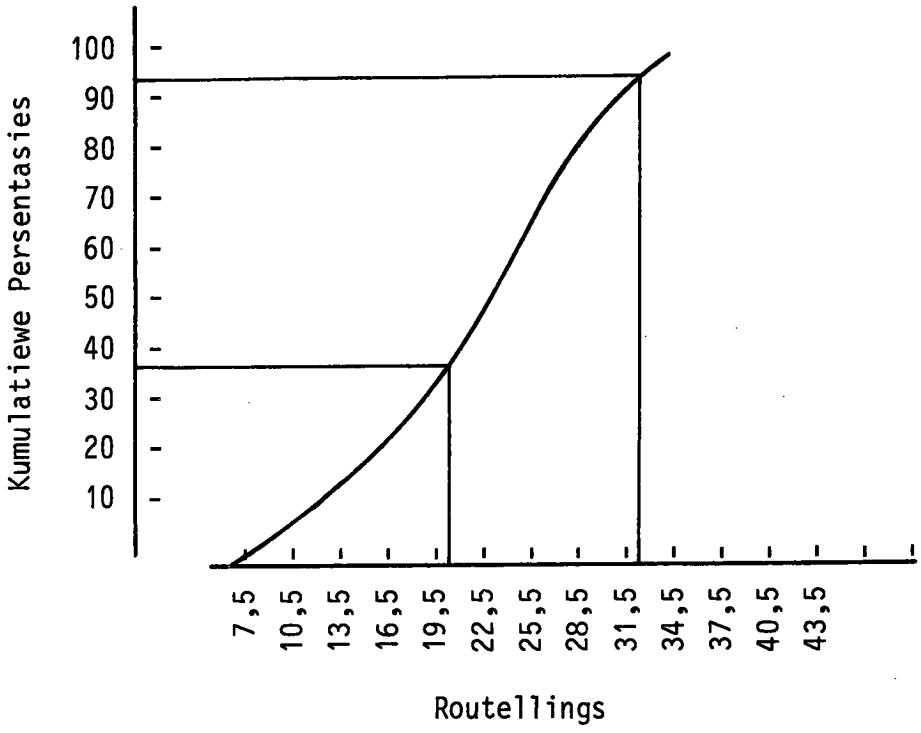
FREKWENSIEVERDELING VAN DIE ROUTELLINGS VAN 450 LEERLINGE IN 'N REKEN= TOETS

Roupunte klasinterval	Klasinterval= eindpunte	Frekwensies	Kumulatiewe frekwensies	Persentasie kumulatiewe frekwensie (benader)
2-4	4,5	0	0	0
5-7	7,5	10	10	2,2
8-10	10,5	20	30	6,7
11-13	13,5	30	60	13,3
14-16	16,5	40	100	22,2
17-19	19,5	55	155	34,4
20-22	22,5	70	225	50,0
23-25	25,5	75	300	66,7
26-28	28,5	55	355	78,9
29-31	31,5	35	390	86,7
32-34	34,5	20	410	91,1
35-37	37,5	25	435	96,7
38-40	40,5	10	445	98,9
41-43	43,5	5	450	100,0

Die metode vir normbepaling behels dat die kumulatiewe persentasie= grafiek vir die toetspunte getrek word.

Om die kumulatiewe persentasiekromme te trek, word die kumulatiewe persentasies langs die Y-as en die roupunte langs die X-as voorgestel. Vir elke interval word die kumulatiewe persentasie teenoor die *eind= punt* van die betrokke interval op die grafiekpapier uitgestip (kyk figuur 7.2). Die verskillende punte word dan verbind om die kumula= tiewe persentasiekromme te gee. Indien daar onreëlmatighede in die kromme voorkom, kan die kromme aangepas of "gladgemaak" word deur 'n

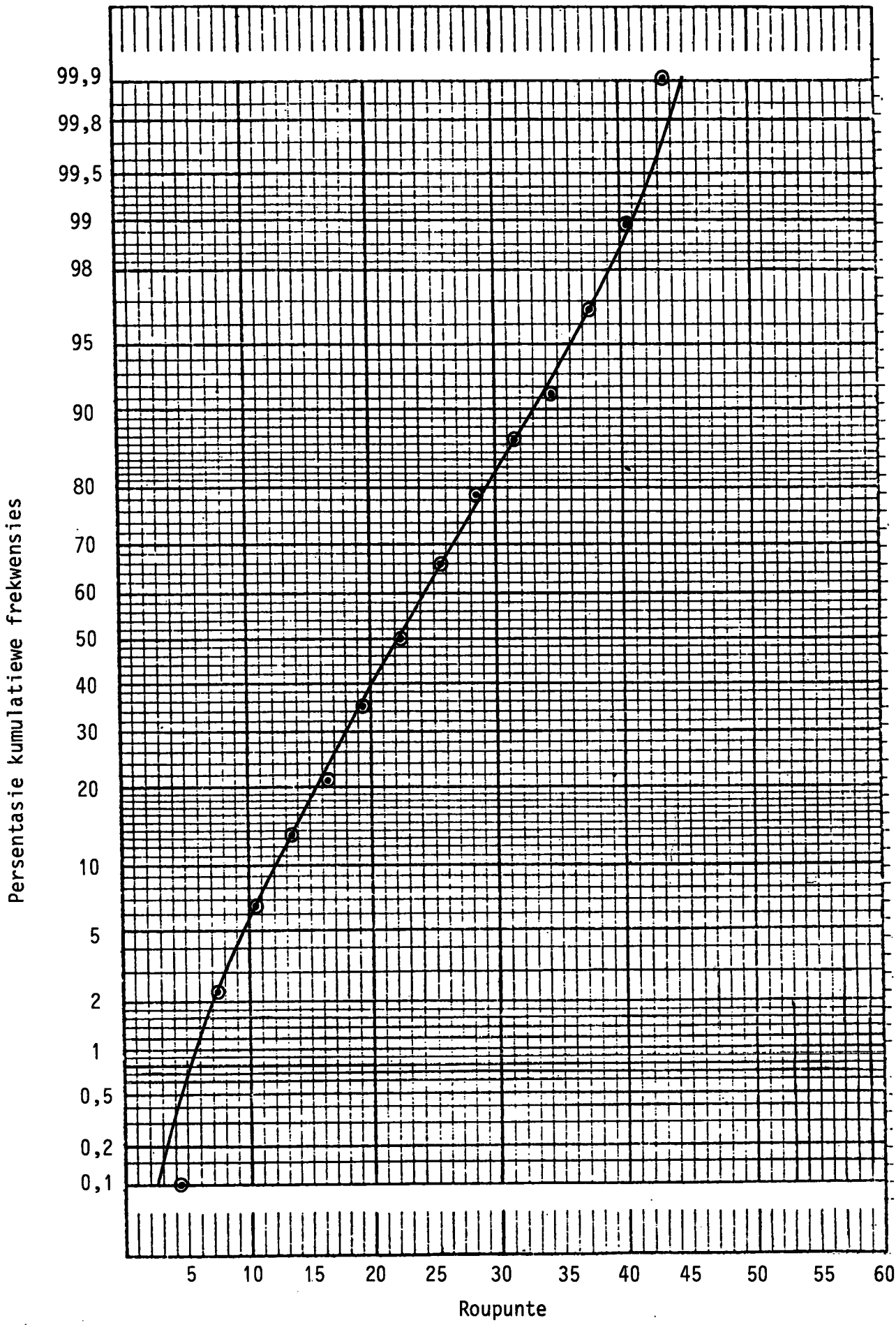
benaderde kromme tussen die punte deur te trek, op so 'n manier dat naastenby eweveel punte bo en onder die lyn val.



FIGUUR 7.2: KUMULATIEWE PERSENTASIEKROMME VAN FREKWENSIEVERDELING IN TABEL 7.2

Met behulp van die kumulatiewe persentasiekromme kan vir elke roupunt 'n persentielrang afgelees word. Vir die routellings 20 en 33 is die persentielrange 37 en 91 onderskeidelik.

Dit is beter om van normaalwaarskynlikheidspapier gebruik te maak om die normgrafiek op te teken (kyk figuur 7.3). Wanneer die kumulatiewe proporsies van 'n *normaalverdeling* op normaalwaarskynlikheidspapier uitgestip word, vorm die grafiek 'n reguit lyn, terwyl dit op gewone grafiekpapier die bekende S-vorm (ogief) aanneem. Die grafiek sal vir *enige* normaalverdeling van roupunte 'n reguit lyn vorm en naastenby 'n reguit lyn wees vir enige verdeling wat min van die normale afwyk. Onreëlmatighede in die verdeling wat as gevolg van kansfaktore kon ingesluip het, sal waarskynlik verwyder word deur die kromme glad te stryk. Die kromme hoef egter nie noodwendig 'n reguit lyn te wees nie. Die normgrafiek in figuur 7.3 is weer eens gebaseer op die gegewens uit



FIGUUR 7.3: KUMULATIEWE PERSENTASIEKROMME VAN FREKWENSIEVERDELING IN TABEL 7.2 OP NORMAALWAARSKYNLIKHEIDSPAPIER

tabel 7.2. Die resultate wat verkry word van die grafieke in figure 7.2 en 7.3 is dieselfde. Vergelyk byvoorbeeld die persentielrange vir routellings 20 en 33. Op grond van die normgrafiek (figuur 8.3) kan die normpunte afgelees en tabel 7.3 opgestel word.

TABEL 7.3

PERSENTIELRANGNORMS VIR DIE REKENTOETS

Roupunt	Persentielrang	Roupunt	Persentielrang
44 - 50	100		
40 - 43	99	22	49
39	98	21	42
38	98	20	39
37	97	19	32
36	96	18	29
35	94	17	24
34	93	16	20
33	91	15	17
32	89	14	14
31	86	13	11
30	83	12	10
29	80	11	7
28	78	10	5
27	72	9	4
26	69	8	3
25	62	6 - 7	2
24	59	2 - 5	1
23	52	0 - 1	0

7.9 VOORBEELD VAN DIE BEPALING VAN DIE TRANSFORMASIETABEL VAN ROUPUNTE NA STANEGES

Dit is gerieflik om vir hierdie voorbeeld ook die gegewens in tabel 7.2 te gebruik. Daar word aanbeveel dat 'n kumulatiewe persentasiekromme soos in figuur 7.3 getrek word sodat onreëlmatighede in die kromme uitgestryk kan word. In die algemeen behoort slegs geringe aanpassings in die uitstrykproses gemaak te word.

Hierna kan die roupunte wat ooreenstem met die boonste persentielrang= grens (kyk tabel 7.1) afgelees word en tot die naaste heelgetal benader word en 'n tabel soos tabel 7.4 kan dan opgestel word.

TABEL 7.4

STANEGENORMS VIR DIE REKENTOETS

Stanege	Roupunte
1	0 - 9
2	10 - 13
3	14 - 17
4	18 - 20
5	21 - 24
6	25 - 29
7	30 - 32
8	33 - 36
9	37+

LITERATUURLYS VIR HOOFSTUK 7

1 ANASTASI, ANNE. *Psychological testing.* 2nd ed. London, Macmillan, 1961.

2 GRONLUND, N.E. *Measurement and evaluation in teaching.* New York, Macmillan, 1965.

3 GUILFORD, J.P. *Fundamental statistics in psychology and education.* New York, McGraw-Hill, 1956.

4 LYMAN, HOWARD B. *Test scores and what they mean.* Englewood-Cliffs, New Jersey, Prentice-Hall, 1963.

5 REMMERS, H.H. & GAGE, N.L. *A practical introduction to measurement and evaluation.* New York, Harper and Brothers, 1943.

HOOFSTUK 8
DIE KLASSIEKE TOETSTEORIEMODEL

8.1 INLEIDING

Die toetse wat gebruik word in sielkundige meting gee 'n kwantitatiewe beskrywing van 'n persoon in terme van die hoeveelheid van die psigologiese eienskap wat hy besit. Indien dieselfde toets verskeie kere op dieselfde persoon toegepas word, word feitlik sonder uitsondering gevind dat die tellings varieer van een herhaling na 'n ander. Die telling wat 'n persoon in 'n bepaalde toepassing behaal, kan dan nie die ware hoeveelheid van die eienskap wat die persoon besit, weerspieël nie.

Die klassieke toetsteoriemodel verskaf 'n wyse om hierdie verskynsels in sielkundige meting te beskryf en te verklaar. Die klassieke toetsteorie gaan uit van die veronderstelling dat die toetstelling wat 'n persoon op 'n toets behaal of die waargenome telling, uit 'n ware komponent en 'n komponent wat toegeskryf kan word aan 'n sogenaamde fouttelling of metingsfout bestaan.

Die klassieke toetsteorie is vir die eerste keer vroeg hierdie eeu in die werk van Spearman (Stanley, 1971, p. 372) gebruik, maar die eerste gedetailleerde oorsig is deur Gulliksen (1950) aangebied. Lord en Novick (1974) gee 'n volledige uiteensetting van die klassieke toetsteorie.

8.2 WARE TELLINGS EN FOUTTELLINGS

8.2.1 Herhaalde toepassing van 'n toets

Die klassieke toetsteoriemodel kan die beste beskryf word aan die hand van herhaalde toepassings van dieselfde toets op dieselfde persoon onder die aanname dat die tellings wat behaal word onafhanklik van mekaar is. Die aanname van onafhanklikheid van die toetstellings impliseer dat die eienskap wat gemeet word konstant sal bly oor die herhaalde toepassings en dat die telling wat op enige toepassing verkry word, nie beïnvloed word deur die tellings op enige vorige toepassings nie. In die praktyk weet ons egter dat baie sielkundige eienskappe gedurende

toetsing self veranderinge ondergaan en dat toetslinge se response as gevolg van die uitwerking van oefening en moegheid kan verander sodat dit eintlik nie moontlik is om volkome onafhanklike herhaalde toepassings te verkry nie. Die begrip van onafhanklike herhaalde toepassings bly egter 'n nuttige een en verhinder ons nie om byvoorbeeld die eienskappe van waargenome tellings te ondersoek en afleidings te maak nie.

Gestel 'n toets g word toegepas op persoon a op r verskillende geleenthede, dan kan die metings of waargenome tellings wat verkry word aangedui word deur

$$x_{ga1}, x_{ga2}, \ldots, x_{gar}$$

of kortweg as

$$x_{gak}, \; k = 1, 2, \ldots, r$$

Hierdie tellings word beskou as die waardes wat die waargenome tellingvariant X_{ga} aanneem en wat wissel oor herhaalde toepassings van toets g op persoon a. Die verdeling van X_g oor herhaalde metings is nie dieselfde vir alle persone nie, maar wissel van persoon tot persoon.

8.2.2 Definisie van 'n ware telling

Die *ware telling*, t_{ga}, van 'n persoon a in toets g word gedefinieer as

die gemiddelde of verwagte waarde van die waargenome tellings X_{ga} *oor 'n oneindige aantal herhalings.*

$$t_{ga} = \mathop{\mathcal{E}}_{k}(X_{ga})$$

waar $\mathop{\mathcal{E}}_{k}$ die verwagte waarde oor k herhalings van die toets aandui.

Die ware telling is nie direk meetbaar nie, maar soos blyk uit die voorgemelde definisie, bestaan daar 'n wiskundige verband tussen die ware telling en die waargenome telling wat direk meetbaar is. Uit die definisie volg ook dat die ware telling van 'n persoon in 'n be-

paalde toets 'n konstante is.

Hierdie definisie van 'n ware telling verwys egter nie na enige fundamentele of korrekte hoeveelheid van die eienskap wat gemeet word nie. Byvoorbeeld 'n toets kan foutiewelik en aanhoudend 'n te hoë telling vir 'n persoon gee en die ware telling van die persoon in hierdie toets sal nog steeds omskryf word as die gemiddelde van die tellings oor 'n oneindige aantal herhalings daarvan.

8.2.3 Definisie van 'n fouttelling

Terwyl die ware telling van 'n bepaalde persoon 'n konstante numeriese waarde het, wissel die fouttelling van een herhaling na die ander en word aangedui deur e_{gak}, $k = 1, 2, \ldots, r$. Hierdie fouttellings is die waardes wat die variant E_{ga} aanneem.

Die *fouttellingvariant*, E_{ga}, word omskryf as

> *die verskil tussen die waargenome tellingvariant* X_{ga} *en die konstante ware telling* t_{ga} *vir persoon a:*
>
> $$E_{ga} = X_{ga} - t_{ga}$$

Die fouttelling staan ook bekend as die metingsfout.

Indien die aanname gemaak word dat die eienskap wat gemeet word konstant oor alle herhalings bly, word aanvaar dat die fouttellings of metingsfoute die gevolg is van alle onwenslike *toevallige invloede* op die metingsproses, soos byvoorbeeld veranderinge in die toestand van die individu en ongekontroleerde omgewingsveranderinge.

Dit is belangrik om daarop te let dat as gevolg van die wyse waarop ware en fouttellings hier gedefinieer word, word die effek van alle konstante onwenslike invloede deel van die ware telling en nie deel van die fouttelling nie.

Uit voorgaande definisies van ware en fouttellings volg direk dat

$$X_{ga} = t_{ga} + E_{ga}$$

wat impliseer dat

die waargenome telling van 'n persoon in enige herhaalde toepassing van 'n toets, gelyk is aan die som van sy ware telling in die toets en sy fouttelling vir die bepaalde toepassing van die toets.

8.2.4 Afleidings uit die definisies van ware en fouttellings

Eerstens volg uit die definisies van ware en fouttellings dat

die verwagte waarde of gemiddelde van die fouttellings vir enige gegewe persoon oor herhalings nul is:

$$\mathop{\mathcal{E}}_{k}(E_{ga}) = 0$$

Volgens die definisie van 'n fouttelling is

$$E_{ga} = X_{ga} - t_{ga}$$

Dus
$$\begin{aligned}\mathop{\mathcal{E}}_{k}(E_{ga}) &= \mathop{\mathcal{E}}_{k}(X_{ga} - t_{ga}) \\ &= \mathop{\mathcal{E}}_{k}(X_{ga}) - \mathop{\mathcal{E}}_{k}(t_{ga}) \\ &= t_{ga} - t_{ga} \\ &= 0\end{aligned}$$

Bogenoemde afleiding beteken net dat die positiewe en negatiewe afwykings van die waargenome tellings vanaf die ware telling mekaar oor 'n onbeperkte getal herhalings uit kanselleer.

Tweedens, kan afgelei word dat

die fouttellingvariansie gelyk is aan die waargenome tellingvariansie vir enige gegewe persoon.

Met ander woorde $\underset{k}{\text{Var}}\,(E_{ga}) = \underset{k}{\text{Var}}\,(X_{ga})$

Beskou die vergelyking

$$\underset{k}{\text{Var}}\,(X_{ga}) = \underset{k}{\text{Var}}\,(t_{ga} + E_{ga})$$

$$= \underset{k}{\text{Var}}\,(t_{ga}) + \underset{k}{\text{Var}}\,(E_{ga}) + 2\underset{k}{\text{Kov}}\,(t_{ga}, E_{ga})$$

Aangesien t_{ga} 'n konstante is en die variansie van 'n konstante waarde sowel as die kovariansie met 'n veranderlike altyd nul is, is

$$\underset{k}{\text{Var}}\,(X_{ga}) = \underset{k}{\text{Var}}\,(E_{ga})$$

Dit is dus duidelik dat die variansie van die waargenome telling slegs toegeskryf kan word aan die variansie van die fouttelling.

Aangesien die kovariansie van 'n konstante en 'n veranderlike altyd nul is en gevolglik 'n korrelasie van nul gee, volg derdens dat

die ware telling en die fouttelling van elke persoon ongekorreleerd is.

$$\underset{k}{\text{Kov}}\,(t_{ga}, E_{ga}) = 0$$

8.3 WARE TELLINGS EN FOUTTELLINGS OOR PERSONE

Tot dusver is daar net oorweging geskenk aan die variansie in tellings oor herhaalde toepassings van 'n toets op 'n enkele persoon (intra-individueel). Aangesien daar in die toetsteorie hoofsaaklik klem gelê word op interindividuele verskille, word vervolgens gelet op die variasie in toetstellings van verskillende persone op 'n enkele toets.

Gestel 'n toets g word herhaaldelik op elk van N persone wat ewekansig uit 'n populasie P gekies is, toegepas. Die tellings wat op die wyse verkry word kan diagrammaties soos in tabel 1 voorgestel word. In tabel 1 stel elke ry die waargenome tellings oor herhaalde toepassings van 'n toets op 'n enkele persoon voor. Die bespreking in paragrawe

TABEL 1

DIAGRAMMATIESE VOORSTELLING VAN TELLINGS VERKRY UIT HERHAALDE TOEPASSING VAN 'N TOETS g OP ELK VAN DIE N PERSONE

Persone	Waargenome tellingvariant X_g Herhalings					Ware tellingvariant
	1	2		k	r	T_g
1	X_{g11}	X_{g12}	...	X_{g1k}	X_{g1r}	$\mathcal{E}_k(X_{g1}) = t_{g1}$
2	X_{g21}	X_{g22}	...	X_{g2k}	X_{g2r}	$\mathcal{E}_k(X_{g2}) = t_{g2}$
3	X_{g31}	X_{g32}	...	X_{g3k}	X_{g3r}	$\mathcal{E}_k(X_{g3}) = t_{g3}$
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
a	X_{ga1}	X_{ga2}	...	X_{gak}	X_{gar}	$\mathcal{E}_k(X_{ga}) = t_{ga}$
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
N	X_{gN1}	X_{gN2}	...	X_{gNk}	X_{gNr}	$\mathcal{E}_k(X_{gN}) = t_{gN}$

8.2.2 en 8.2.3 het gehandel oor die gemiddelde en die variansie van die tellings in so 'n ry. Byvoorbeeld die gemiddelde van die waargenome tel= lings in ry a gee die ware telling (t_{ga}) van persoon a. Die ware tel= lings van die verskillende persone verskyn in die regterkantste kolom van die tabel.

Verder, kan elke waargenome telling in die tabel in twee komponente verdeel word, naamlik 'n ware tellingkomponent en 'n fouttellingkompo= nent, byvoorbeeld

$$x_{gak} = t_{ga} + e_{gak}$$

Die kolomme van tabel 1 dui die waargenome tellings aan wat verskil= lende persone in 'n enkele toepassing van 'n toets behaal het. Hierdie waargenome tellings wat die N persone in een toepassing van 'n toets g behaal het, kan beskou word as die waardes wat die variant X_g aanneem.

In 'n vorige paragraaf is aangetoon dat die ware telling t_{ga} vir 'n be= sondere persoon a 'n konstante is wat nie oor herhaalde toepassings varieer nie. Die ware telling varieer egter van een persoon na 'n ander en die ware tellings

$$t_{ga},\ a = 1, 2, \ldots, N$$

kan beskou word as waardes wat die variant T_g aanneem en wat van persoon na persoon varieer maar nie oor herhaalde metings vir 'n persoon nie.

Net soos daar aanvaar word dat die waargenome tellingvariant X_g oor metings op verskillende persone varieer, word aanvaar dat die fouttel= ling E_g oor metings op verskillende persone ook sal varieer. Uit die definisie van die fouttelling volg dan

$$E_g = X_g - T_g$$

Hierdie vergelyking impliseer dat

$$X_g = T_g + E_g$$

waar X_g 'n waargenome telling van 'n toets, T_g 'n onwaarneembare ware telling en E_g 'n onwaarneembare fouttelling voorstel. Laasgenoemde vergelyking is een van die vyf basiese vergelykings van die klassieke toetsteoriemodel.

Om die oorblywende 4 basiese vergelykings van die model af te lei, word twee aannames vereis: Die eerste aanname is dat

die variansie van die waargenome tellings oor persone eindig is,

dit is

$$\underset{a}{\text{Var}}\,(X_g) < \infty.$$

Hierdie word nie as 'n streng aanname beskou nie en word in die praktyk geregverdig aangesien die verdelings van toetstellings eindige grense en gevolglik eindige momente van elke orde het. Hierdie aanname impliseer ook dat die verwagte waardes oor persone van die waargenome, ware en fouttellings asook die variansies van hierdie tellings eindig is.

Die tweede aanname wat gemaak word,

is dat vir elke persoon a in die populasie alle pare metings X_{ga} en X_{ha} op twee toetse g en h onafhanklik van mekaar is.

Dit beteken dat die pare metings ongekorreleerd is, dit wil sê

$$\underset{k}{\text{Kov}}\,(X_{ga}, X_{ha}) = 0 \text{ vir alle persone a}$$

Om die verwagte waarde, variansie en kovariansie van tellings oor persone te onderskei van dié oor herhaalde toepassings, word onderskeidelik gebruik gemaak van die volgende notasies:

$$\mu\,(X_g) = \underset{a}{\mathcal{E}}(X_g)$$

$$\sigma^2(X_g) = \underset{a}{\text{Var}}\,(X_g)$$

$$\sigma\,(X_g, X_h) = \underset{a}{\text{Kov}}\,(X_g, X_h)$$

Die tweede vergelyking van die klassieke toetsteorie volg uit die stelling dat

die verwagte waarde van die fouttelling oor persone nul is.

Dit word soos volg bewys: Omdat die verwagte waarde oor persone ook die verwagte waarde oor herhaalde toepassings insluit, word die ver= wagte waarde van die fouttelling oor persone deur middel van 'n dub= bele verwagtingswaarde uitgedruk, naamlik

$$\mathop{\mathcal{E}}_{a}\{\mathop{\mathcal{E}}_{k}(E_{ga})\}$$

Aangesien die verwagte waarde van die fouttellings vir 'n bepaalde per= soon oor herhaalde toepassings nul is, volg dat

$$\mu(E_g) = \mathop{\mathcal{E}}_{a}\{\mathop{\mathcal{E}}_{k}(E_{ga})\} = \mathop{\mathcal{E}}_{a}(0) = 0$$

Die volgende stellings lewer die oorblywende fundamentele vergelykings van die klassieke toetsteorie:

Die ware telling en die fouttelling oor persone is ongekorreleerd

$$\sigma\,(T_g, E_g) = 0$$

Die ware telling op een meting (toets g) en die fouttelling op 'n tweede meting (toets h) oor persone is ongekorreleerd.

$$\sigma\,(T_g, E_h) = 0$$

Die fouttellings op twee verskillende metings, toetse g en h, oor persone is ongekorreleerd.

$$\sigma\,(E_g, E_h) = 0$$

Laasgenoemde drie vergelykings kan met behulp van stellings wat in Mood et al. (1974, pp. 158-159) voorkom, afgelei word. As gevolg van die hoeveelheid Wiskunde betrokke by hierdie stellings word die bewyse van die vergelykings nie hier gegee nie.

Die aanname oor die onafhanklikheid van pare metings oor persone is slegs noodsaaklik vir die bewys van $\sigma\,(E_g, E_h) = 0$. Vir die bewys van

die ander vergelykings is die aanname van eindige variansies en verwagte waardes van waargenome, ware en fouttellings nodig.

Opsommend kan gestel word dat die volgende 5 vergelykings saam die klas= sieke toetsteoriemodel uitmaak:

$$X_g = T_g + E_g$$

$$\mu (E_g) = 0$$

$$\sigma (T_g, E_g) = 0$$

$$\sigma (T_g, E_h) = 0 \quad g \neq h$$

$$\sigma (E_g, E_h) = 0 \quad g \neq h$$

8.4 IMPLIKASIES VAN DIE KLASSIEKE TOETSTEORIEMODEL

Aan die hand van bogenoemde vyf vergelykings kan verskeie belangrike vergelykings wat in die psigometrika gebruik word, afgelei word. Hier= die vergelykings dui die verband aan tussen die momente van die verde= lings van die onwaarneembare ware en fouttellings en die momente van die waarneembare tellings.

8.4.1 Verwagte waardes, variansies en korrelasies

Uit die definisie van die waargenome telling $X_g = T_g + E_g$ en die verge= lyking $\mu (E_g) = 0$, volg dat

die verwagte waarde van die waargenome tellings oor persone gelyk is aan die verwagte waarde van die ware tellings,

dit is

$$\mu (X_g) = \mu (T_g)$$

Die definisie van die waargenome telling stel $X_g = T_g + E_g$. Nou volg $\mu (X_g) = \mu (T_g + E_g) = \mu (T_g) + \mu (E_g) = \mu (T_g) + 0 = \mu (T_g)$

Verder, is

die kovariansie tussen die waargenome tellings en ware tellings gelyk aan die variansie van die ware tellings,

dit is

$$\sigma\ (X_g,\ T_g) = \sigma^2\ (T_g)\ .$$

Deur gebruik te maak van die verwantskap tussen X_g en T_g, kan $\sigma\ (X_g,\ T_g)$ geskryf word as

$$\sigma\{(T_g + E_g),\ T_g\} = \sigma\ (T_g,\ T_g) + \sigma\ (T_g,\ E_g)^*,$$

maar T_g en E_g is ongekorreleerd $\{\sigma\ (T_g,\ E_g) = 0\}$ en $\sigma\ (T_g,\ T_g) = \sigma^2(T_g)$. Dus

$$\sigma\ (X_g,\ T_g) = \sigma^2\ (T_g)$$

Die algebraiese manipulasie van die variansie van die waargenome tellings gee

$$\sigma^2\ (X_g) = \sigma^2\ (T_g + E_g)$$

$$= \sigma^2\ (T_g) + \sigma^2\ (E_g) + 2\sigma\ (T_g,\ E_g)^{**}$$

Maar T_g en E_g is ongekorreleerd. Dus $\sigma\ (T_g,\ E_g) = 0$.

Daarom $\sigma^2(X_g) = \sigma^2\ (T_g) + \sigma^2\ (E_g)$

---

* Kov $\{(X + Y),\ Z\}$ = Kov $(X,\ Z)$ + Kov $(Y,\ Z)$

** Var $(X + Y)$ = Var (X) + 2Kov $(X,\ Y)$ + Var (Y)

Kyk bladsy 179 van Mood, A.M. et al. Introduction to the theory of statistics, 3rd Edition, Tokyo, McGraw-Hill Kogakusha, 1974, vir die afleidings van die twee vergelykings.

Hieruit volg die stelling dat

die waargenome tellingvariansie gelyk is aan die som van die ware tellingvariansie en die fouttellingvariansie.

Die waargenome tellingvariansie oor persone, kan dus opgebreek word as die som van twee variansiekomponente. Een komponent van hierdie waargenome variansie sal die gevolg wees van die verskille tussen die ware tellings van verskillende individue, dit is die ware tellingvariansie $\sigma^2 (T_g)$. Hierdie variansie verteenwoordig dus interindividuele varieerbaarheid in die eienskap wat gemeet word. Die oorblywende deel van die waargenome tellingvariansie, $\sigma^2 (E_g)$, reflekteer die varieerbaarheid in die tellings van die individuele persone oor herhaalde toepassings van dieselfde toets. Dit dui dus intra-individuele varieerbaarheid aan.

8.4.2 Betroubaarheid

Gestel dat elke persoon se waargenome tellings op elke herhaalde toepassing van 'n bepaalde toets dieselfde bly. Elke persoon se fouttellings sal dus gelyk aan nul wees en sy waargenome tellings gelyk wees aan sy ware telling. Dit impliseer dat alle waargenome variansie in so 'n situasie uit die variansie van ware tellings bestaan, aangesien die fouttellingvariansie gelyk aan nul is. Die varieerbaarheid van die waargenome tellings sal dus slegs die verskille tussen die ware tellings van die persone reflekteer. Sodanige metings sal perfek betroubaar wees.

Omdat metings in die sielkunde nooit perfek betroubaar is nie, sal die waargenome tellingvariansie altyd in 'n mindere of meerdere mate fouttellingvariansie bevat. Indien die waargenome tellingvariansie slegs uit fouttellingvariansie bestaan, sal sodanige metings geheel en al onbetroubaar wees. Hoe groter die bydrae van die ware tellingvariansie tot die waargenome tellingvariansie, hoe hoër sal die betroubaarheid van die toets wees.

Die betroubaarheidskoëffisiënt van 'n meting word gedefinieer as die proporsie van die waargenome tellingvariansie wat uit ware tellingvariansie bestaan, dit is die verhouding van die ware telling=

variansie tot die waargenome tellingvariansie, $\frac{\sigma^2 (T_g)}{\sigma^2 (X_g)}$.

'n Betroubaarheidskoëffisiënt van 0,92 beteken dat 92 persent van die waargenome tellingvariansie toe te skryf is aan ware tellingvariansie en slegs 8 persent aan fouttellingvariansie.

Die betroubaarheidskoëffisiënt kan egter ook uitgedruk word as

die gekwadreerde korrelasiekoëffisiënt tussen waargenome en ware tellings.

Dit kan soos volg aangetoon word:

Vermenigvuldig beide die teller en noemer van bostaande verhou= ding met $\sigma^2 (T_g)$ en aangesien $\sigma (X_g, T_g) = \sigma^2 (T_g)$ volg nou

$$\frac{\{\sigma^2 (T_g)\}^2}{\sigma^2 (X_g)\, \sigma^2 (T_g)} = \frac{\{\sigma (X_g, T_g)\}^2}{\{\sigma (X_g)\, \sigma (T_g)\}^2}$$

$$= \left\{\frac{\sigma (X_g, T_g)}{\sigma (X_g)\, \sigma (T_g)}\right\}^2$$

Per definisie word die korrelasie tussen die waargenome telling en ware telling gegee deur

$$\rho (X_g, T_g) = \frac{\sigma (X_g, T_g)}{\sigma (X_g)\, \sigma (T_g)} .$$

Dus volg dat

$$\frac{\sigma^2 (T_g)}{\sigma^2 (X_g)} = \rho^2 (X_g, T_g)$$

Deur gebruik te maak van die vergelyking $\sigma^2 (X_g) = \sigma^2 (T_g) + \sigma^2 (E_g)$, kan laasgenoemde vergelyking ook geskryf word as

$$\rho^2 (X_g, T_g) = \frac{\sigma^2 (X_g) - \sigma^2 (E_g)}{\sigma^2 (X_g)}$$

$$= 1 - \frac{\sigma^2 (E_g)}{\sigma^2 (X_g)} .$$

Uit die vergelyking $\sigma^2 (X_g) = \sigma^2 (T_g) + \sigma^2 (E_g)$ blyk dat die ware tel= lingvariansie nie die waargenome tellingvariansie kan oorskry nie. Die verhouding

$$\frac{\sigma^2 (T_g)}{\sigma^2 (X_g)}$$

impliseer dat die betroubaarheidskoëffisiënt $\rho^2 (X_g, T_g)$ nooit groter as een kan wees nie. Aangesien die fouttellingvariansie ook nie die waargenome tellingvariansie kan oorskry nie, sal

$$1 - \frac{\sigma^2 (E_g)}{\sigma^2 (X_g)}$$

nooit negatief wees nie. Gevolglik word die grense van die betroubaar= heidskoëffisiënt gegee deur

$$0{,}00 \leq \rho^2 (X_g, T_g) \leq 1{,}00$$

8.4.3 Parallelle metings

In die vorige paragraaf is 'n definisie van die betroubaarheidskoëffisiënt geformuleer aan die hand van herhaalde toepassings van dieselfde toets. In hierdie paragraaf sal aangetoon word dat die betroubaarheidskoëffi= siënt ook in terme van parallelle toetse omskryf kan word.

Parallelle toetse is toetse wat dieselfde attribuut of eienskap meet. Statisties is toetse parallel indien

die ware tellings en fouttellingvariansies van die toetse gelyk is vir elke persoon in die populasie.

Uit hierdie definisie kan afgelei word dat

1. die verwagte waardes van die waargenome tellings van die parallelle toetse gelyk is;

2. die variansies van die waargenome tellings van die parallelle toetse gelyk is;

3. die kovariansies van die waargenome tellings van die parallelle toetse gelyk is;

4. die kovariansies van die waargenome tellings van die parallelle toetse gelyk is aan die variansie van die ware telling van enigeen van die toetse;

5. die korrelasies tussen die parallelle toetse gelyk is en

6. die korrelasies tussen die parallelle toetse en enige willekeurige afsonderlike toetse gelyk is.

Die bewyse vir bogenoemde afleidings kan in Lord en Novick (C1974, pp. 47-50 en pp. 58-59) gevind word.

Die korrelasiekoëffisiënt tussen twee parallelle metings X_g en X'_g word per definisie gegee deur

$$\rho\,(X_g, X'_g) = \frac{\sigma\,(X_g, X'_g)}{\sigma\,(X_g)\,\sigma\,(X'_g)}$$

maar afleiding 4 hierbo stel

$$\sigma\,(X_g, X'_g) = \sigma^2\,(T_g) = \sigma^2\,(T'_g)$$

en afleiding 2 stel

$$\sigma^2\,(X_g) = \sigma^2\,(X'_g).$$

Gevolglik sal

$$\rho\,(X_g, X'_g) = \frac{\sigma^2\,(T_g)}{\sigma^2\,(X_g)} \quad \text{of} \quad \frac{\sigma^2\,(T'_g)}{\sigma^2\,(X'_g)}$$

$$= \rho^2\,(X_g, T_g) \quad \text{of} \quad \rho^2\,(X'_g, T'_g).$$

Hierdie resultaat gee 'n verband tussen die betroubaarheidskoëffisiënt en die korrelasie van parallelle metings, naamlik

die korrelasie tussen twee parallelle metings is gelyk aan die betroubaarheidskoëffisiënt van enigeen van die metings.

Die vergelyking

$$\rho^2\,(X_g, T_g) = \frac{\sigma^2\,(T_g)}{\sigma^2\,(X_g)}$$

gee die betroubaarheidskoëffisiënt in terme van onwaarneembare groothede terwyl die linkerkantste term van vergelyking

$$\rho\,(X_g, X'_g) = \frac{\sigma^2\,(T_g)}{\sigma^2\,(X_g)}$$

die betroubaarheidskoëffisiënt uitdruk in 'n waarneembare grootheid.

8.4.4 Ware tellingvariansie en fouttellingvariansie

Die variansies van ware en fouttellings kan ook in waarneembare groothede uitgedruk word. Op grond van die definisie van die betroubaarheidskoëf= fisiënt as die verhouding van die ware tellingvariansie tot die waarge= nome tellingvariansie

$$\rho\,(X_g, X'_g) = \frac{\sigma^2\,(T_g)}{\sigma^2\,(X_g)}$$

kan die onwaarneembare ware tellingvariansie geskryf word as

$$\sigma^2 (T_g) = \sigma^2 (X_g)\, \sigma (X_g, X'_g)$$

Al die terme aan die regterkant van hierdie vergelyking is waarneembare groothede.

Op grond van die vergelykings

$$\sigma^2 (X_g) = \sigma^2 (T_g) + \sigma^2 (E_g) \text{ en } \sigma^2 (T_g) = \sigma^2 (X_g)\, \rho (X_g, X'_g)$$

kan die variansie van die onwaarneembare fouttellings eweneens in terme van waarneembare groothede uitgedruk word:

$$\sigma^2 (X_g) = \sigma^2 (X_g)\, \rho (X_g, X'_g) + \sigma^2 (E_g)$$

dit wil sê $\sigma^2 (E_g) = \sigma^2 (X_g) - \sigma^2 (X_g)\, \rho (X_g, X'_g)$

$$= \sigma^2 (X_g) \{1 - \rho (X_g, X'_g)\}$$

Die vierkantswortel van die fouttellingvariansie, naamlik

$$\sigma (E_g) = \sigma (X_g) \sqrt{1 - \rho (X_g, X'_g)}$$

staan bekend as die standaardmetingsfout. Die standaardmetingsfout is dus die standaardafwyking van die fouttellings, $X_g - T_g$, wat die afwyking aandui waarmee die waargenome telling van die ware telling verskil.

8.4.5 Geldigheid

Die klassieke model het ook belangrike implikasies vir die geldigheidskoëffisiënte van toetse, dit wil sê vir die korrelasiekoëffisiënte tussen die toets- en kriteriumtellings. Toets- en kriteriumtellings is waargenome tellings wat aan metingsfoute onderhewig is en gevolglik is die verkreë korrelasies tussen die waargenome tellings kleiner as dié tussen die ooreenstemmende ware tellings. Vervolgens word enkele verhoudings ondersoek tussen die korrelasie tussen twee waargenome tellings, X en Y, en die korrelasie tussen die ooreenstemmende ware tellings, T_x en T_y.

Met die reeds genoemde aannames van die klassieke toetsmodel kan aangetoon word dat

die korrelasiekoëffisiënt tussen die tellings op 'n toets en 'n kriterium gelyk is aan die produk van die korrelasiekoëffisiënt tussen die ware tellings van die twee veranderlikes en die vierkantswortels van die betroubaarheidskoëffisiënte van die veranderlikes,

dit wil sê:

$$\rho\ (X, Y) = \rho\ (T_x, T_y) \sqrt{\rho\ (X, X')\ \ \rho\ (Y, Y')}$$

Hierdie vergelyking toon dat indien die toetsbetroubaarheid, $\rho\ (X, X')$, laag is, die geldigheidkoëffisiënt van die toets met betrekking tot enige kriterium laag sal wees.

Laasgenoemde vergelyking kan ook in 'n vorm geskryf word om aan te toon dat die korrelasie tussen die ware tellings uitgedruk kan word as 'n funksie van die betroubaarheidskoëffisiënte en die korrelasie tussen die waargenome toets- en kriteriumtellings. Hierdie vergelyking verskaf die korrelasiekoëffisiënt wat verkry sou gewees het indien die toets- en die kriteriumtellings volkome betroubaar was, dit wil sê indien die korrelasie tussen hulle ware tellings in plaas van die waargenome tellings bereken sou gewees het. Hierdie vorm van die vergelyking is die volgende:

$$\rho\ (T_x, T_y) = \frac{\rho\ (X, Y)}{\sqrt{\rho\ (X, X')}\ \sqrt{\rho\ (Y, Y')}}$$

Laasgenoemde vergelyking staan bekend as die korreksie vir *attenuasie* of die korreksie vir *verswakking*. Die begrip verswakking verwys na die verkreë korrelasie wat verswak is as gevolg van die onbetroubaarheid van die waargenome toets- en kriteriumtellings. Met behulp van bogenoemde vergelyking kan 'n korreksie vir hierdie verswakking aangebring word en die korrelasie tussen metingsfoutvrye toets- en kriteriumtellings bepaal word, dit wil sê tussen die ware toets- en kriteriumtellings. 'n Woord van waarskuwing is egter hier nodig. In die praktyk word talle aannames gemaak wanneer betroubaarheidskoëffisiënte bereken, of beter gestel, geskat word. Hierdie skattings is dikwels onderskattings en gevolglik is daar groot gevaar dat die korrelasiekoëffisiënt, $\rho\ (T_x, T_y)$, tussen die ware tellings oorskat kan word.

Uit die attenuasievergelyking kan redelik maklik ingesien word dat

as die betroubaarheidskoëffisiënt van die kriterium kleiner is as die betroubaarheidskoëffisiënt van die toets

sal

$$\rho (X, X') \geq |\rho (X, Y)|$$

met ander woorde die absolute waarde van die geldigheidskoëffisiënt van die toets vir dié bepaalde kriterium is kleiner as die betroubaarheidskoëffisiënt van die toets. Die volgende vergelyking geld egter in alle omstandighede:

$$\sqrt{\rho (X, X')} \geq |\rho (X, Y)|$$

met ander woorde die absolute waarde van die geldigheidskoëffisiënt van 'n toets vir enige kriterium is kleiner as die vierkantswortel van die betroubaarheidskoëffisiënt van die toets.

Die klassieke toetsteorie kan, ten spyte van sekere tekortkominge, gesien word as een van die mees bruikbare en gebruikte wiskundige modelle in die psigologie. Die teorie word met relatief min aannames afgelei en verder blyk dit dikwels dat dit bruikbaar bly selfs wanneer sekere aannames van die model nie ten volle geld nie.

LITERATUURLYS VIR HOOFSTUK 8

1 BROWNE, M.W. *Sielkunde (Natuurwetenskap).* Module PSY 302. Pretoria, Universiteit van Suid-Afrika, 1975.

2 DE LEEUW, J. *Algemene psychodiagnostiek II: Testtheorie.* Amsterdam, Swets en Zeitlinger, 1978.

3 GULLIKSEN, H. *Theory of mental tests.* New York, John Wiley and Sons, Inc., 1950.

4 HUYSAMEN, G.K. *Psychological test theory.* Durbanville, Uitgewery Boschendal, 1980.

5 LEMKE, E. & WIERSMA, W. *Principles of psychological measurement.* Chicago, Rand McNally College Publishing Company, 1976.

6 LORD, F.M. & NOVICK, M.R. *Statistical theories of mental test scores.* Massachusetts, Addison-Wesley Publishing Company, C1974.

7 MOOD, A.M., GRAYBILL, F.A. & BOESS, D.C. *Introduction to the theory of statistics.* 3rd Edition, Tokyo, McGraw-Hill Kogakusha, 1974.

8 STANLEY, J.C. Reliability. In: Thorndike, R.L. ed. *Educational measurement.* 2nd Edition. Washington, D.C., American Council on Education, 1971.

HOOFSTUK 9
BETROUBAARHEID

9.1 BEGRIPSOMSKRYWING

Die betroubaarheid van 'n meetinstrument verwys na die mate waarmee ongewenste faktore enige meting met die instrument sal beïnvloed. Hoe hoër die betroubaarheid is hoe kleiner is die invloed van sulke faktore en omgekeerd. Met ander woorde, daar kan gesê word dat die betroubaarheid van 'n meetinstrument sal bepaal hoe konsekwent dit van een geleentheid na 'n ander sal meet, op voorwaarde dat die eienskap wat gemeet word nie verander nie. Wanneer ons die betroubaarheid van 'n psigologiese toets wil kwantifiseer, vind ons egter dat dit slegs gedoen kan word deur 'n indeks, wat die betroubaarheidskoëffisiënt genoem word, wat 'n funksie is van die toets self, maar ook 'n funksie is van 'n bepaalde groep persone vir wie die toets gebruik kan word.

9.2 DIE KLASSIEKE TOETSMODEL

In sy poging om dinge te verstaan, maak die wetenskaplike dikwels gebruik van 'n model of modelle om verskynsels en die verband tussen verskillende verskynsels te verklaar. Wanneer 'n model goed in sy doel slaag, is sommige mense ongelukkig geneig om die model met absolute waarheid te verwar. 'n Mens is ook soms geneig om te vergeet dat 'n model veral in die geesteswetenskappe gewoonlik 'n baie onvolmaakte benadering tot werklike waarnemings gee. 'n Goeie voorbeeld in hierdie verband is persoonlikheidsteorie en die bestaan van redelik stabiele persoonlikheidstrekke. Die meeste mense sal seker saamstem dat die teorie van persoonlikheidstrekke geregverdig is, maar daar is min besonder betroubare navorsingsresultate om dit te steun. Dit is net 'n ander manier om te sê dat ons op grond van metings van persoonlikheidstrekke die toekomstige gedrag van mense nie besonder goed kan voorspel nie.

Voorgenoemde gebreke in ons modelle is geopper om net weer te waarskuteen 'n te rigiede siening van 'n model soos die klassieke toetsmodel wat die basis vorm van ons bespreking van betroubaarheid.

In hoofstuk 8 word die klassieke toetsmodel redelik volledig bespreek. In dié model word afgelei dat enige meting X van 'n meetinstrument ver= deel kan word in 'n ware komponent T en 'n foutkomponent E. Dus: $X = T + E$.

In paragraaf 8.4.1 word aangetoon dat die totale variansie $\sigma^2(X)$ (soms geskryf as σ_X^2) van die veranderlike in twee onafhanklike komponente verdeel word, soos volg:

$$\sigma^2(X) = \sigma^2(T) + \sigma^2(E)$$

waar $\sigma^2(T)$ = variansie van die ware komponent (ware variansie)
en $\sigma^2(E)$ = variansie van die foutkomponent (foutvariansie)

Die betroubaarheidskoëffisiënt ρ_{XX} van die meetinstrument word soos volg gedefinieer:

$$\rho_{XX} = \frac{\sigma^2(X) - \sigma^2(E)}{\sigma^2(X)}$$

$$= \frac{\sigma^2(T)}{\sigma^2(X)}$$

$$= \frac{\sigma^2(T)}{\sigma^2(T) + \sigma^2(E)}$$

$$= \frac{\text{ware variansie}}{\text{totale variansie}}$$

T en E is nie direk waarneembaar nie. Die natuurlike vraag is dan hoe $\sigma^2(T)$ en $\sigma^2(E)$ gevind kan word. In die paragrawe wat volg sal aange= dui word hoe skattings van hierdie hoeveelhede onder sekere verdere aannames gemaak kan word.

9.3 TOETS-HERTOETSBETROUBAARHEID

Gestel 'n toets word twee keer onder identiese omstandighede op elke lid van 'n universum toegepas. Laat $X_1 = T_1 + E_1$ die toetstellings met die eerste geleentheid en $X_2 = T_2 + E_2$ die toetstellings met die tweede geleentheid voorstel. Laat $\rho(X_1, X_2)$ die korrelasiekoëffi=

siënt van X_1 met X_2 oor persone voorstel en $\sigma(T_1)$ die standaardafwyking van T_1. Dan kan uit die formule van die produkmoment-korrelasiekoëffi= siënt afgelei word dat

$$\rho(X_1, X_2) = \rho(T_1 + E_1, T_2 + E_2)$$

$$= \frac{\rho(T_1, T_2)\sigma(T_1)\sigma(T_2) + \rho(T_1, E_2)\sigma(T_1)\sigma(E_2) + \rho(E_1, T_2)\sigma(E_1)\sigma(T_2) + \rho(E_1, E_2)\sigma(E_1)\sigma(E_2)}{\sigma^2(T_1) + \sigma^2(E_1) + 2\rho(T_1, E_1)\sigma(T_1)\sigma(E_1)^{\frac{1}{2}} \times \sigma^2(T_2) + \sigma^2(E_2) + 2\rho(T_2, E_2)\sigma(T_2)\sigma(E_2)^{\frac{1}{2}}}$$

As die aannames van die klassieke toetsmodel geld, volg egter dat

$$\rho(T_1, T_2) = 1$$

$$\rho(T_1, E_1) = \rho(T_1, E_2) = 0$$

$$\rho(T_2, E_1) = \rho(T_2, E_2) = \rho(E_1, E_2) = 0$$

$$\sigma(T_1) = \sigma(T_2)$$

en $\sigma(E_1) = \sigma(E_2)$.

Dus word

$$\rho(X_1, X_2) = \frac{\sigma^2(T_1)}{\sigma^2(T_1) + \sigma^2(E_1)}$$

= betroubaarheidskoëffisiënt van die toets (aangedui deur ρ_{XX})

Die volgende belangrike stelling kan dus gemaak word: As die klas= sieke toetsmodel geld, is die korrelasiekoëffisiënt oor persone tus= sen 'n eerste meting en 'n tweede meting van 'n meetinstrument gelyk aan die betroubaarheidskoëffisiënt van die meetinstrument. Hierdie stel= ling is in ooreenstemming met die intuïtiewe begrip dat 'n meetinstru= ment betroubaar is as dit konsekwent van geleentheid tot geleentheid meet.

Op die oog af lyk dit nou of ons 'n eenvoudige manier het om die be=troubaarheid van 'n meetinstrument te evalueer. Neem bloot twee me=tings vir elke persoon en korreleer die eerste met die tweede meting. Hierdie eenvoudige metode werk egter slegs as aan die vereistes van die klassieke toetsmodel voldoen word. Dit impliseer dat

(a) die ware tellings van persone nie van die eerste tot die tweede meting mag verander nie en
(b) ware- en fouttellings ongekorreleerd moet wees. Daar mag dus nie geheue- of oefeningseffekte tussen die eerste en tweede metings wees nie.

Slegs onder ideale omstandighede kan aanvaar word dat die vereistes van die klassieke toetsmodel wel geld. Die rol van geheue- en oefe=ningsfaktore kan verminder word deur die tydperk tussen die eerste en tweede toepassing te verleng. Die gepaardgaande gevaar bestaan dat die ware tellings van die eienskap wat gemeet word dan vir die proefpersone kan verander. 'n Gedeeltelike oplossing van die probleem lê daarin dat die tydperk tussen die twee toepassings van die toets sodanig moet wees dat die gesamentlike rol van al die probleemfaktore 'n minimum sal wees. Die navorser moet op grond van sy kennis van die eienskap wat gemeet word en die aard van die meetinstrument, op die geskikste tydperk tussen eerste en tweede meting besluit.

As 'n voorbeeld van die oorwegings by die bepaling van die toets-her=toetsbetroubaarheidskoëffisiënt van 'n toets, kan 'n verbale redenerings=toets ('n kragtoets) vir 6-jarige kinders genoem word. Indien hierdie toets by twee geleenthede kort na mekaar (binne weke) toegepas word, kan sommige kinders van hul antwoorde uit die eerste toepassing van die toets onthou. Met die tweede toepassing kan hulle dan sonder om weer te redeneer weer dieselfde antwoorde (wat reg of verkeerd kan wees) gee. In so 'n geval sal die korrelasiekoëffisiënt van die punte van die eerste en tweede toepassing kunsmatig hoog wees. As die toets egter na 'n relatiewe lang periode (sê langer as 4 maande) op dieselfde leerlinge toegepas word, kan differensiële ontwikkeling van verbale vermoë by die kinders hulle relatiewe posisies wat betref die onder=hawige vermoë aansienlik verander het. Die gevolglike lae korrelasie=koëffisiënt van die punte van die eerste en tweede toepassing van die toets sal dan nie 'n korrekte weergawe van die betroubaarheidskoëffi=

siënt van die toets wees nie. 'n Toets-hertoetstydperk van ongeveer 4 tot 8 weke behoort in hierdie geval bevredigende resultate te lewer.

Oor die algemeen kan beweer word dat as die betroubaarheidskoëffisiënt deur die toets-hertoetsmetode bereken word en dit laag is, is die betrokke meetmiddel 'n swak meetmiddel. As die betroubaarheidskoëffisiënt soos bereken deur die toets-hertoetsmetode hoog is, kan dit slegs aanvaar word as aan die vereistes van die klassieke toetsmodel voldoen word.

Met die ontwikkeling van die formule

$$\rho_{XX} = \rho(X_1, X_2)$$

is geen aannames gemaak oor die faktoriale samestelling van die ware komponent van 'n meting nie. Die toets-hertoetsmetode is dus geskik om die betroubaarheid van 'n toets wat meer as een onderliggende persoonlikheidstrek meet, te skat. Geen aannames is ook gemaak dat 'n toets saamgestel is uit 'n aantal items nie, of dat die toets 'n kragtoets of 'n spoedtoets moet wees nie. Die metode kan dus geskik wees om die betroubaarheid van spoed- en kragtoetse te skat, mits aan die basiese aannames van die klassieke toetsmodel voldoen word. In die geval van spoedtoetse kan egter dikwels verwag word dat sowel die ware komponent T as die fout komponent E funksies van tyd sal wees. Ten spyte hiervan is die metode steeds die eenvoudigste en waarskynlik die beste om die betroubaarheid van spoedtoetse te skat. 'n Hele aantal ander metodes is gelukkig vir kragtoetse beskikbaar.

9.4 DIE KUDER-RICHARDSONFORMULES

Kuder en Richardson het reeds in 1937 'n aantal formules afgelei wat gebruik kan word om die betroubaarheid van 'n meetinstrument te skat. Hierdie formules bly steeds die mees algemeen bruikbare formules vir dié doel. Die formules is genommer en hoe hoër die nommer van 'n formule word, hoe strenger is die vereistes waaraan die metings moet voldoen vir die formule om te geld. Die volgende beperkings geld vir al die Kuder-Richardsonformules, behalwe waar dit eksplisiet anders gestel word:

(a) Die meetinstrument moet 'n toets wees wat bestaan uit 'n aantal items wat elk tellings van 0 of 1 kan hê.

(b) Die toets moet 'n kragtoets wees.

(c) Die toetstellings en itemtellings moet aan die vereistes van die klassieke toetsmodel voldoen.

Laat nou

$$X = x_1 + x_2 + x_3 + \ldots + x_n$$

$$Y = y_1 + y_2 + y_3 + \ldots + y_n$$

waar X = totale toetstelling met 'n eerste meting

x_i = itemtelling van item i met 'n eerste meting

Y = totale toetstelling met 'n tweede meting

y_i = itemtelling van item i met 'n tweede meting

n = aantal items in die toets.

Met behulp van die gewone formule vir die produkmoment-korrelasiekoëf= fisiënt kan die volgende formule afgelei word:

ρ_{XX} = betroubaarheidskoëffisiënt van die toets

$= \rho(X, Y)$

$= \rho(\Sigma x, \Sigma y)$

$$= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \rho(x_i, y_j)\sigma(x_i)\sigma(y_j)}{\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} \rho(x_i, x_j)\sigma(x_i)\sigma(x_j)\right\}^{\frac{1}{2}} \times \left\{\sum_{i=1}^{n}\sum_{j=1}^{n} \rho(y_i, y_j)\sigma(y_i)\sigma(y_j)\right\}^{\frac{1}{2}}}.$$

Met behulp van verdere algebraïese verwerkings en die aannames tot dus= ver gemaak, kan aangetoon word dat:

$$\rho(x_i, x_j) = \rho(y_i, y_j),$$

$$\sigma(x_j) = \sigma(y_j)$$

en $\rho(x_i, x_j) = \rho(x_i, y_j)$ vir $i \neq j$.

Aangesien itemtellings slegs 0 of 1 kan wees, is

$$\sigma^2(x_i) = p_i q_i = \text{variansie van item i}$$

waar p_i = moeilikheidswaarde van item i
en $q_i = 1 - p_i$.

Verder is $\sigma_x^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma(x_i)\sigma(x_j)\rho(x_i, x_j)$

= variansie van die saamgestelde telling X.

Hieruit word die Kuder-Richardsonformule 3 gevind:

$$\rho_{xx} = \frac{\sigma_x^2 - \sum_{i=1}^{n} p_i q_i + \sum_{i=1}^{n} \rho_{ii} p_i q_i}{\sigma_x^2} \qquad \text{K-R 3}$$

waar $\rho_{ii} = \rho(x_i, y_i)$

= betroubaarheidskoëffisiënt van item i.

Die K-R-formule 3 is nie besonder bruikbaar nie aangesien die itembetroubaarhede ρ_{ii} nie uit eksperimentele data bereken kan word nie. Die itembetroubaarhede kan wel geskat word deur byvoorbeeld die hoogste korrelasie van 'n item met alle ander items as die itembetroubaarheid te neem. Wanneer 'n faktorontleding van die items gedoen is, kan die itemkommunaliteite as skattings van die itembetroubaarhede geneem word.

Die Kuder-Richardsonformule 8 kan uit die K-R-formule 3 afgelei word as die volgende addisionele aanname gemaak word: Dit interkorrelasiematriks van items van die toets het 'n rang van 1. Dit beteken eenvoudig dat elke item dieselfde eienskap moet meet, of anders gestel, 'n faktorontleding van die items moet slegs een faktor oplewer en elke item moet hoog op dié faktor laai. Die formule is soos volg:

$$\rho_{xx} = \frac{\sigma_x^2 - \Sigma p_i q_i}{2\sigma_x^2} + \sqrt{\frac{\Sigma \rho_{it}^2 p_i q_i}{\sigma_x^2} + \left(\frac{\sigma_t^2 - \Sigma p_i q_i}{2\sigma_x^2}\right)^2} \qquad \text{K-R 8}$$

waar σ_x^2 = variansie van die toets

ρ_{it} = korrelasie tussen item i en die toetstelling

p_i = moeilikheidswaarde van die i-de item

q_i = $1 - p_i$.

Die Kuder-Richardsonformule 14 kan weer uit die K-R-formule 8 afgelei word deur die addisionele aanname te maak dat die interkorrelasies tussen alle items in die toets gelyk is. Die formule is soos volg:

$$\rho_{xx} = \left(\frac{\sigma_x^2 - \Sigma p_i q_i}{(\Sigma\sqrt{p_i q_i})^2 - \Sigma p_i q_i}\right) \cdot \left(\frac{(\Sigma\sqrt{p_i q_i})^2}{\sigma_x^2}\right) \qquad \text{K-R 14}$$

waar die simbole dieselfde betekenis het as wat voorheen gedefinieer is.

Deur die verdere aanname te maak dat die variansie van al die items gelyk is, kan die Kuder-Richardsonformule 20 uit die K-R-formule 14 afgelei word. Die formule is soos volg:

$$\rho_{xx} = \left(\frac{n}{n-1}\right) \cdot \left(\frac{\sigma_x^2 - \Sigma p_i q_i}{\sigma_x^2}\right) \qquad \text{K-R 20}$$

waar n = aantal items in die toets.

Deur verder aan te neem dat al die items in die toets dieselfde moeilikheidswaarde het, kan die Kuder-Richardsonformule 21 afgelei word:

$$\rho_{xx} = \left(\frac{n}{n-1}\right) \cdot \left(\frac{\sigma_x^2 - M(1 - \frac{M}{n})}{\sigma_x^2}\right) \qquad \text{K-R 21}$$

waar M = gemiddelde vir die hele toets.

Daar kan wiskundig aangetoon word dat

$$\rho_{xx}(\text{K-R 20}) \geq \rho_{xx}(\text{K-R 21}).$$

Wanneer die voorwaardes vir die gebruik van die K-R-formule 20 geld, sal die numerieke waarde van die K-R-formule 21 kleiner of gelyk aan die numerieke waarde van die K-R-formule 20 wees. Die K-R-formule 21 gee, met ander woorde, dan 'n onderskatting van die betroubaarheids= koëffisiënt.

Kuder en Richardson (1937) beweer dat die numerieke waardes wat onder= skeidelik met die formules K-R 14 en K-R 20 deur hulle in praktiese situasies verkry is, met nie meer as 0,001 van mekaar verskil het nie.

In die algemeen kan dit gestel word dat die K-R-formule 8 die beste skatting vir die betroubaarheidskoëffisiënt van 'n toets gee wanneer aan die voorwaardes wat voorheen genoem is, voldoen word. Die for= mules K-R 14, K-R 20 en K-R 21 behoort slegs gebruik te word as prak= tiese oorwegings (byvoorbeeld die beskikbaarheid van rekenaargeriewe) die gebruik van die K-R-formule 8 baie moeilik of onmoontlik maak. In so 'n geval moet steeds daarna gestreef word om die formule met die laagste nommer te gebruik.

9.5 DIE KOëFFISIëNT ALPHA (α)

Die koëffisiënt alpha is 'n berekenbare hoeveelheid vir 'n saamgestelde telling wat sodanig is dat die betroubaarheidskoëffisiënt van die saam= gestelde telling 'n boonste perk is vir die koëffisiënt alpha. Gestel $y_1, y_2, \ldots y_n$ is metings van n veranderlikes en dat

$$Y = y_1 + y_2 + \ldots y_n.$$

Met die aannames van die klassieke toetsteorie kan aangetoon word dat

$$\rho_{yy} \geq \alpha \equiv \frac{n}{n-1} \quad \frac{\sigma^2(y) - \Sigma\sigma^2(y_i)}{\sigma^2(y)} \qquad \ldots (A)$$

waar $\sigma^2(y_i)$ = variansie van y_i

en $\sigma^2(y)$ = variansie van Y.

Nodige en voldoende voorwaardes vir $\rho_{yy} = \alpha$ is dat die ware telling vir elke komponent van Y 'n lineêre funksie van enigeen van hulle moet wees en dat die ware variansies van elke komponent gelyk aan mekaar moet wees. Met uitsondering van die vereiste van binêre items is hierdie voorwaardes ekwivalent aan die wat gelei het tot die K-R-formule 20. As die komponente van die saamgestelde telling binêre items is, is $\sigma^2(y_i) = p_i q_i$, waar p_i die moeilikheidswaarde van die i-de item is en $q_i = 1 - p_i$. Deur $\Sigma\sigma^2(y_i)$ in vergelyking A te vervang met $\Sigma p_i q_i$ word gevind dat die vergelyking vir α identies is aan die K-R-formule 20.

Die betroubaarheidskoëffisiënt is 'n boonste perk vir die waarde van die alphakoëffisiënt van 'n saamgestelde telling. Die enigste vereis= te wat vir die komponenttellings gestel word, is dat hulle aan die vereistes van die klassieke toetsteorie moet voldoen. Geen faktoriale homogeniteit van die komponente word vereis nie. Die komponenttel= lings kan ook itemtellings wees. Die itemtellings hoef nie slegs 'n strek van 0 tot 1 te besit nie. Al die items van die toets hoef ook nie dieselfde strek vir hulle itemtellings te besit nie. As α 'n be= vredigende waarde vir 'n toets het, kan aanvaar word dat die betrou= baarheidskoëffisiënt van die toets bevredigend sal wees.

9.6 FERGUSSON SE AANPASSING VAN DIE K-R-formule 20

Die K-R-formule 20 geld slegs wanneer itemtellings slegs 0 of 1 kan wees, dit wil sê vir binêre items. Fergusson (1951) het die formule aangepas vir die geval waar itemtellings van 0 tot m kan wees (m is 'n positiewe heelgetal). As die proporsie persone wat elke itemtel= ling vir item i verkry het, voorgestel word deur $p_0, p_1, \ldots, p_m$ kan die variansie van die betrokke item geskryf word as:

$$s_i^2 = \sum_{k=0}^{m} k^2 p_k - \left(\sum_{k=0}^{m} k p_k\right)^2.$$

Vir 'n binêre item met moeilikheidswaarde p is die variansie van item= tellings pq waar q = 1 - p. Fergusson se aanpassing kom daarop neer dat die term $\Sigma p_i q_i$ in die K-R-formule 20 vervang word deur Σs_i^2.

9.7 BETROUBAARHEID VAN 'N SAAMGESTELDE TELLING

Die behoefte bestaan soms om die betroubaarheidskoëffisiënt te vind vir 'n lineêre kombinasie van veranderlikes. Dit kan gedoen word mits die betroubaarheidskoëffisiënte en die interkorrelasies van die komponente (veranderlikes) bekend is. Die aanname is dat die klassieke toetsmodel vir elke komponent van die lineêre kombinasie geld. Gestel

$$X = w_1X_1 + w_2X_2 + \ldots + w_nX_n$$

waar n = aantal komponente van die lineêre kombinasie,

w_i = gewig van die i-de komponent,

X = lineêre kombinasie

en X_i = i-de komponent van die lineêre kombinasie.

Die betroubaarheidskoëffisiënt ρ_{XX} van X kan dan soos volg met behulp van Mosier (1943) se formule bepaal word:

$$\rho_{XX} = 1 - \frac{\Sigma w_j^2\sigma_j^2 - \Sigma w_j^2\sigma_j^2 r_{jj}}{\Sigma w_j^2\sigma_j^2 + 2\sum_{j=1}^{n-1}\sum_{k=2}^{n} w_j w_k \sigma_j \sigma_k r_{jk}}$$
$$(j < k)$$

waar r_{jj} = betroubaarheidskoëffisiënt van X_j,

σ_j^2 = variansie van X_j,

r_{jk} = korrelasie tussen X_j en X_k.

Sichel (1950) het 'n soortgelyke formule afgelei, maar hy het eenheidsgewigte aan die komponente in die lineêre kombinasie toegeken. Sy formule kan uit Mosier se formule gevind word deur $w_j = 1$ vir alle j te stel. Sichel beweer verder dat as die betroubaarheidskoëffisiënt van die i-de komponent van die kombinasie onbekend is, 'n konserwatiewe skatting daarvan gegee word deur die grootste een van die korrelasiekoëffisiënte van die i-de komponent met die ander komponente.

9.8 DIE SPEARMAN-BROWNFORMULE

Wanneer ons 'n totaaltelling van die tellings van n parallelle toetse kry, kan die betroubaarheidskoëffisiënt van die totaaltelling in terme van die betroubaarheidskoëffisiënt van die parallelle toetse (al die koëffisiënte is gelyk aan mekaar, kyk volgende paragraaf), geskryf word. Die Spearman-Brownformule, wat die genoemde verband gee, kan baie maklik uit die formule van Mosier gevind word deur daarop te let dat al die gewigte gelyk is aan 1, al die variansies gelyk is aan mekaar, al die betroubaarheidskoëffisiënte gelyk is aan mekaar en die korrelasiekoëffisiënte van enige twee toetse gelyk is aan hulle gesamentlike betroubaarheidskoëffisiënt. Die Spearman-Brownformule is soos volg:

$$\rho_{nn} = \frac{n\rho_{ii}}{1 + (n - 1)\rho_{ii}}$$

waar ρ_{nn} = betroubaarheidskoëffisiënt van die toets wat bestaan uit n parallelle komponente,

n = aantal komponente,

en ρ_{ii} = betroubaarheidskoëffisiënt van elkeen van die n komponente.

Die Spearman-Brownformule kan gebruik word om die betroubaarheidskoëffisiënt te skat van 'n nuwe toets wat verkry word wanneer 'n bestaande toets met bekende betroubaarheidskoëffisiënt met 'n aantal items verkort of verleng word. Die volgende vorm van die Spearman-Brownformule kan uit bostaande formule afgelei word:

$$\rho_{yy} = \frac{\frac{y}{x}\rho_{xx}}{1 + (\frac{y}{x} - 1)\rho_{xx}}$$

waar ρ_{xx} = bekende betroubaarheidskoëffisiënt van 'n toets met x items

ρ_{yy} = geskatte betroubaarheidskoëffisiënt van die nuwe toets met y items.

Met die afleiding van hierdie vorm van die Spearman-Brownformule is die aanname gemaak dat elke item 'n toets is en dat al die items parallelle toetse is.

9.9 PARALLELLE VORMS-METODE

In paragraaf 8.4.3 word parallelle toetse gedefinieer. Daar is aangetoon dat as twee toetse parallel is, is hulle korrelasiekoëffisiënte met enige ander toets gelyk en die twee toetse se gemiddeldes en variansies gelyk. Met ander woorde, as twee toetse X_1 en X_2 parallel is, volg dat:

$$\bar{X}_1 = \bar{X}_2$$

$$s(X_1) = s(X_2)$$

$$\rho(X_1, Y) = \rho(X_2, Y) \text{ vir alle } Y$$

waar $\rho(X_1, Y)$, $\rho(X_2, Y)$ = korrelasiekoëffisiënte van X_1 en X_2 met Y,

$\bar{X}_1$, $\bar{X}_2$ = gemiddeldes van X_1 en X_2,

en $s(X_1)$, $s(X_2)$ = variansies van X_1 en X_2.

Verder volg vir twee parallelle toetse (kyk paragraaf 8.4.3) dat

$$\rho(X_1, X_2) = \rho_{X_1X_1} = \rho_{X_2X_2}$$

waar $\rho(X_1, X_2)$ = korrelasiekoëffisiënt van toets 1 en toets 2,

$\rho_{X_1X_1}$ = betroubaarheidskoëffisiënt van toets 1

en $\rho_{X_2X_2}$ = betroubaarheidskoëffisiënt van toets 2.

Die betroubaarheidskoëffisiënte van parallelle toetse is dus gelyk aan mekaar en ook gelyk aan die korrelasiekoëffisiënt van enige twee van die parallelle toetse.

Die betroubaarheidskoëffisiënt van 'n toets kan dus bepaal word deur die toets en sy parallelle vorm op 'n geskikte groep persone toe te pas en die korrelasiekoëffisiënt tussen die twee tellings te bereken. Die grootste probleem van die metode is dat dit gewoonlik moeilik is om twee parallelle vorms van 'n toets op te stel. Die probleme verbonde aan die toets-hertoetsmetode geld verder ook vir die parallelle vorms-metode.

9.10 DIE HALFVERDELINGSMETODE

In hierdie metode word aanvaar dat die twee toetse wat verkry word deur respektiewelik die ewe genommerde items van 'n toets met n items en die onewe genommerde items te neem, twee parallelle toetse met elk n/2 items is.

Die korrelasiekoëffisiënt oor persone tussen die ewe en onewe items se totaaltellings is dan 'n skatting van die betroubaarheidskoëffisiënt van elk van die halwe toetse.

Die aanname dat die twee halwe toetse parallelle toetse is, kan oor die algemeen slegs geregverdig word as die items homogeen is, dit wil sê as al die items slegs een onderliggende faktor meet. Indien die items wel homogeen is, kan die betroubaarheid van die hele toets gevind word deur die Spearman-Brownformule. Die formule vir twee komponente is soos volg:

$$\rho_{xx} = \frac{2\rho_{\frac{x}{2}\frac{x}{2}}}{1 + \rho_{\frac{x}{2}\frac{x}{2}}}$$

waar $\rho_{\frac{x}{2}\frac{x}{2}}$ = betroubaarheidskoëffisiënt van elke halwe toets

ρ_{xx} = betroubaarheidskoëffisiënt van die hele toets

Kuder en Richardson (1937) voer die argument aan dat dit beter is om die K-R-formules te gebruik vir die skatting van 'n betroubaarheidskoëffisiënt as die halfverdelingsmetode aangesien laasgenoemde metode skattings kan lewer wat òf te hoog òf te laag kan wees terwyl die K-R-formules nooit oorskattings sal gee nie.

9.11 STEURINGSVERANDERLIKES

Die verdeling van die variansie van 'n toets vir 'n sekere groep persone in slegs twee komponente naamlik 'n komponent van ware variansie wat slegs toegeskryf kan word aan variansie as gevolg van die onderliggende konstruk en 'n variansiekomponent wat toegeskryf kan word aan toevallige foute (random errors) word dikwels bemoeilik as gevolg van die invloed

van steuringsveranderlikes. Steuringsveranderlikes kan gedefinieer word as veranderlikes (nie die onderliggende konstruk nie) wat in 'n variansie-ontleding van toetstellings geassosieer kan word met breuk= dele van die totale variansie. Steuringsveranderlikes is bronne van ongewenste variansie in toetstellings. Hierdie ongewenste variansie kan duidelik onderskei word van variansie wat te wyte is aan toeval= lige faktore wat moeilik aanwysbaar is.

Wanneer steuringsveranderlikes vir 'n bepaalde groep persone 'n rol speel by die variansie van 'n toets, kan die probleem soms opgelos word deur norms en betroubaarheidskoëffisiënte te bereken vir subgroepe van die groep wat homogeen sal wees ten opsigte van die belangrikste steurings= veranderlikes. Ongelukkig is dit nie altyd 'n oplossing nie, aangesien daar sterk oorwegings kan wees waarom sekere groepe nie vir normbepa= ling geskei behoort te word nie. In sulke gevalle kan die betroubaar= heidskoëffisiënte ten minste vir homogene groepe ten opsigte van die steuringsveranderlikes bepaal word.

Die volgende konkrete voorbeeld van steuringsveranderlikes kan genoem word. Veranderlikes soos ouderdom, skoolstanderd, ras, taal, geslag, toetsafnemer, ensovoorts, is dikwels steuringsveranderlikes wanneer die variansie van intelligensiemetings ondersoek word. Om hierdie rede word norms en betroubaarheidskoëffisiënte vir intelligensietoetse gewoonlik bereken vir groepe persone wat homogeen ten opsigte van die belangrikste steuringsveranderlikes is. Die belangrikste steuringsver= anderlike vir intelligensie by jongmense is waarskynlik ouderdom. Geen intelligensietoets van naam wat vir kinders bedoel is, sal dus norm= groepe aandui waar ouderdom 'n groot variasie binne 'n normgroep sal hê nie.

Die toetsopsteller kan die voorkoms van steuringsveranderlikes vermin= der deur items wat sydig teenoor groepe persone is sover moontlik te vermy. Die probleem van sydigheid is egter 'n moeilike een waarop hier nie dieper ingegaan kan word nie.

Die effek van steuringsveranderlikes op die waardes van betroubaarheids= koëffisiënte wat bepaal word met behulp van die formules wat voorheen bespreek is, is dat die verkreë waardes ongeregverdig verhoog word. Ongelukkig word soms in toetshandleidings waargeneem dat betroubaar=

heidskoëffisiënte bereken is vir groepe wat besonder heterogeen ten opsigte van steuringsveranderlikes is. Hierdie ongewenste praktyk is gewoonlik die gevolg van onkunde, maar kan ongelukkig ook die ge= volg wees van 'n toetsopsteller se begeerte om sy toets beter te laat vertoon as wat dit in werklikheid is.

9.12 STANDAARDMETINGSFOUT

Die betroubaarheidskoëffisiënt is 'n verhouding wat kan varieer tussen 0 en 1 en word nie in 'n eenheid uitgedruk nie. Dit is dikwels nodig om 'n aanduiding te hê van die grense waartussen 'n bepaalde meting waarskynlik lê. Dit is vir praktiese redes onmoontlik om die fout= komponent van 'n bepaalde meting te bepaal. Die variansie (en dus die standaardafwyking) van die foutkomponente kan maklik bepaal word. Hierdie variansie σ_E^2 is gelyk aan:

$$\sigma_E^2 = \sigma_X^2(1 - \rho_{XX})$$

waar σ_X^2 = variansie van die toets

en ρ_{XX} = betroubaarheidskoëffisient.

Die standaardafwyking van die foutkomponent word die *standaardmetings= fout* genoem.

Die standaardmetingsfout σ_E kan dus soos volg bepaal word:

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX}}$$

Die standaardmetingsfout kan byna gesien word as 'n gemiddelde van die absolute waardes van die foutkomponente. Enige meting plus of min die standaardmetingsfout gee dus grense waartussen die ware telling met redelike sekerheid sal lê.

Op hierdie punt moet gewaarsku word teen 'n verkeerde interpretasie van die standaardmetingsfout wat soms voorkom. Die standaardmetings= fout is 'n skatting van die standaardafwyking $\sigma(X|T)$ van waargenome tellings, gegee 'n spesifieke ware telling. Die standaardmetingsfout

is dus nie van toepassing wanneer 'n betroubaarheidsinterval van ware tellings, gegee 'n feilbare waargenome telling, opgestel moet word nie. In die laasgenoemde geval is die standaardafwyking $\sigma(T|X)$ van ware tellings vir 'n gegewe waargenome telling nodig. Daar kan aangetoon word dat $\sigma(X|T)$ groter is as $\sigma(T|X)$ en ernstige foute behoort nie te ontstaan deur te beweer dat die ware telling met redelike sekerheid in die interval $X \pm \sigma(X|T)$ lê nie. DIe bewering dat die ware telling met 68 % sekerheid in dié interval lê, mag egter nie gemaak word nie.

9.13 WANNEER IS 'N BETROUBAARHEIDSKOëFFISIëNT AANVAARBAAR?

Die eerste saak waaraan aandag gegee moet word om te bepaal of 'n betroubaarheidskoëffisiënt se grootte aanvaarbaar is, is om seker te maak dat 'n geskikte metode gebruik is om die koëffisiënt te bereken. Daarna kan na die grootte van die koëffisiënt gekyk word.

'n Eenduidige afsnypunt vir aanvaarbare betroubaarheid kan ongelukkig nie gegee word nie. In gevalle waar baie belangrike besluite in verband met die toekoms van 'n individu op grond van 'n enkele meting geneem moet word, word 'n hoë betroubaarheidskoëffisiënt (0,90 tot 1,00) vir die meetinstrument vereis. In gevalle waar besluite oor 'n individu geneem moet word en metings vir 'n relatief groot verskeidenheid eienskappe (wat nie geheel-en-al onafhanklik van mekaar is nie) beskikbaar is, kan 'n toets met 'n betroubaarheidskoëffisiënt so laag as 0,65 nog 'n betekenisvolle bydrae tot besluitneming lewer. In gevalle waar slegs 'n gemiddelde vir 'n eienskap oor 'n aantal persone bepaal moet word, kan 'n meetinstrument met 'n betroubaarheidskoëffisiënt so laag soos 0,30 nog nuttig gebruik word, mits genoeg persone betrek word om die gemiddelde te bepaal.

LITERATUURLYS VIR HOOFSTUK 9

1 CRONBACH, L.J. Test "reliability": Its meaning and determination. *Psychometrika, Vol. 12,* No. 1, March 1947: 1 - 16.

2 CRONBACH, L.J. & WARRINGTON, W.G. Time-limit tests: Estimating their reliability and degree of speeding. *Psychometrika, Vol. 16,* No. 2, June 1951: 167 - 188.

3 DUDEK, FRANK J. The Continuing misinterpretation of the standard error of measurement. *Psychological Bulletin,* 1979, Vol. 86, No. 2, 335 - 337.

4 FERGUSON, G.A. A note on the Kuder-Richardson formula. *Educational and psychological measurement, Vol. 11,* 1951: 612 - 615.

5 GLESER, G.C. et al. Generalizability of Scores influenced by multi= ple sources of variance. *Psychometrika, Vol. 30,* No. 4, December 1965: 395 - 418.

6 GUILFORD, J.P. *Fundamental statistics in psychology and education.* New York, McGraw Hill, 1965.

7 GULLIKSON, H. The Reliability of speeded tests. *Psychometrika, Vol. 15,* No. 3, September 1950.

8 KUDER, G.F. & RICHARDSON, M.W. The Theory of the estimation of test reliability. *Psychometrika, Vol. 2,* No. 3, September 1937: 151 - 160.

9 LORD, FREDERICK M. & NOVICK, MELVIN R. *Statistical theories of mental test scores.* Addison-Wesley Publishing Company. 1968.

10 MOSIER, C.I. On the reliability of a weighted composite. *Psychome= trika, Vol. 8,* No. 3, September 1943: 161 - 168.

11 NUNNALLY, J.C. *Psychometric theory.* New York, McGraw Hill. 1967.

12 SICHEL, H.S. Note on the reliability of combination of subtests, Tests or Criteria. *Bulletin, NIPR 2,* 1950: 57 - 60.

13 TUCKER, L.R. A note on the estimation of test reliability by the Kuder-Richardson formula 20. *Psychometrika, Vol. 14,* No. 3, June 1949: 117 - 119.

HOOFSTUK 10
VERALGEMEENBAARHEID

10.1 INLEIDING

In hoofstuk 9 wat oor betroubaarheid van 'n meetinstrument handel, is aangedui dat ongewenste faktore, byvoorbeeld geslag van die toetsafnemer, 'n meting kan beïnvloed en dat dit wenslik is om inligting oor die betroubaarheid van 'n meting te hê met die oog op interpretasie daarvan. As gevolg van sekere leemtes in die klassieke betroubaarheidsbenadering het die veralgemeenbaarheidsteorie geleidelik ontwikkel om 'n breër benaderingswyse daar te stel. Dit sal later uit die besprekings dan ook blyk dat die klassieke betroubaarheidsbenadering as 'n spesiale geval in die breër veralgemeenbaarheidsteorie beskou kan word.

'n Ernstige nadeel in die klassieke betroubaarheidsbenadering is die uitgangspunt van parallelle metings, wat met streng beperkings gepaard gaan en moeilik in die praktyk bereik kan word. Soos later sal blyk, stel die uitgangspunt van die veralgemeenbaarheidsteorie nie sulke streng vereistes nie. Die veralgemeenbaarheidsbenadering het verder onder andere ook die volgende voordele:

(a) Dit lei tot die *bewuste* oorweging van die verskillende aspekte (fasette) wat 'n meting kan beïnvloed en vermy gevolglik onduidelikhede en weglatings.

(b) Deur 'n multifasetbenadering kan ook die invloed van die interaksies tussen fasette op meting bestudeer word.

(c) Ondersoeke na die fasette wat 'n meting kan beïnvloed kan meer doeltreffend gedoen word aangesien die veralgemeenbaarheidsbenadering goedbeplande ontwerpe stimuleer sodat byvoorbeeld inligting uit 'n enkele datastel verkry kan word waarvoor andersins miskien verskeie datastelle nodig sou wees.

10.2 BEGRIPSOMSKRYWING

Veralgemening beteken volgens die Psigologiese Woordeboek van Gouws en andere (1979) "die proses waardeur 'n uitspraak gelewer word oor 'n

hele versameling van items (dikwels 'n oneindige aantal) op grond van die bestudering van 'n beperkte aantal daarvan".

In die gedragswetenskappe word die spesifieke stel omstandighede waar= onder waarnemings of metings kan plaasvind gewoonlik beskou as verteen= woordigend van 'n omvangryker stel omstandighede. Byvoorbeeld as waar= nemings uitgevoer word, verteenwoordig die waarnemers en geleenthede waarop waargeneem word, gewoonlik baie ander ewe aanvaarbare waarne= mers en waarnemingsgeleenthede. By meting byvoorbeeld word die items van 'n toets beskou as 'n steekproef uit 'n universum van soortgelyke stelle items, met ander woorde daar is ook baie ander ewe aanvaarbare stelle items.

Die gebruiker van 'n telling (meting of waarneming) stel selde net be= lang in byvoorbeeld die spesifieke respons ten opsigte van die spesi= fieke stimulus (vrae of voorwerp) volgens die spesifieke toetser/ waarnemer volgens die spesifieke tyd van toetsing/waarneming, aange= sien so 'n telling baie beperkend is en bitter min daaruit afgelei kan word. Verder is dit moontlik dat sommige van die spesifiekhede wel kan verander terwyl nogtans 'n ewe aanvaarbare telling verkry kan word. Die telling wat 'n persoon behaal volgens al die spesifieke omstandig= hede daarop van toepassing, kan dus beskou word as slegs een van 'n universum van ewe aanvaarbare tellings wat behaal sal kan word.

Die kernvraag is dan, met watter mate van sekerheid kan die spesifieke bevindings veralgemeen word na die hipotetiese gemiddelde telling wat die persoon byvoorbeeld sou behaal het as meting plaasgevind het deur die hele universum van byvoorbeeld toetsafnemers, deur die hele univer= sum van byvoorbeeld toetstye, deur die hele universum van byvoorbeeld die betrokke soort toetse, ensovoorts. Hierdie hipotetiese gemiddelde telling staan bekend as die *universumtelling* en beklee in die veral= gemeenbaarheidsteorie die plek wat die ware telling by die klassieke benadering beklee.

In die verdere bespreking sal sekere tegniese begrippe gebruik word en dit is wenslik om vooraf aandag aan enkeles te skenk. Twee basiese begrippe is faset en toestand. In hierdie bespreking word onder *faset* verstaan die verskillende *sistematiese* maar nie toepaslike (dit wil sê ongewenste) aspekte wat 'n persoon se telling kan beïnvloed, byvoor=

beeld toetsafnemer/waarnemer en tyd van meting. Onder *toestand* (condition) word verstaan die werklikhede soos dit binne elke faset kan voorkom, byvoorbeeld in 'n faset "toetsafnemer" kan dit wees afnemer A, B of C en in 'n faset "tyd van toetsing" 08h00, 13h00 of 16h00. Al die toelaatbare toestande wat binne 'n faset kan voorkom word die *universum van toelaatbare toestande van die faset* genoem en die stel universums van toelaatbare toestande van al die fasette heet die *universum van toelaatbare waarnemings*. Die term *universum* word soos ook in die voorgaande, gewoonlik met *toestande* geassosieer en die term *populasie* met die voorwerpe van waarneming, byvoorbeeld mense.

Die *universumtelling* van 'n persoon word dus gedefinieer as die gemiddelde telling van die persoon oor die universum van toelaatbare waarnemings. Indien 'n *steekproef* uit die universum van toelaatbare waarnemings geneem word, sal so 'n steekproef bestaan uit steekproewe van toelaatbare toestande in elke faset. Die *waargenome telling* van 'n persoon word gedefinieer as die gemiddelde telling van die persoon oor die steekproef van toelaatbare waarnemings. Die *verwagte waargenome telling* word gedefinieer as die gemiddelde waargenome telling oor al die moontlike steekproewe, met dieselfde grootte en ontwerp, uit die toelaatbare waarnemings.

Die *veralgemeenbaarheidskoëffisiënt* (ρ^2) word gedefinieer as die verhouding van die variansie (oor persone) van die universumtelling tot die verwagte variansie van die waargenome telling. Hierdie veralgemeenbaarheidskoëffisiënt kan gesien word as 'n indeks van vertroue waarmee die waargenome telling die universumtelling kan verteenwoordig. 'n Waarde van nul dui geen vertroue aan en 'n waarde van een, volkome vertroue.

Dit is dan ook nou duidelik dat daar baie veralgemeenbaarheidskoëffisiënte kan wees en dat 'n veralgemeenbaarheidskoëffisiënt bepaal word deur die kombinasie van toestande waarna veralgemeen word, naamlik die *universum van veralgemening*, en die aard en grootte van die steekproef van toelaatbare waarnemings uit die universum van veralgemening. Die universum van veralgemening kan ooreenstem met die universum van toelaatbare waarnemings maar gewoonlik word na beperkter universums veralgemeen.

10.3 DIE ROL VAN VARIANSIE-ONTLEDING IN DIE TEORIE VAN VERALGEMEENBAARHEID

Die teorie van veralgemeenbaarheid is hoofsaaklik deur Cronbach en sy medewerkers uitgewerk en geïntegreer uit die werke van andere oor sekere aspekte daarvan. Cronbach et al. (1972) meld egter dat die laaste woord nog nie gespreek is nie en dat verdere ontwikkelingswerk nog nodig is; 'n stelling wat vandag nog geld.

Die teorie maak hoofsaaklik van die ewekansige effek variansiemodel gebruik en gaan van die volgende aannames uit:

(a) Die toestande van 'n faset is eksperimenteel onafhanklik. 'n Persoon se telling volgens een toestand word dus nie beïnvloed deur die tellings wat volgens ander toestande behaal is nie.

(b) Die toestande wat gebruik word om die variansiekomponente te skat, is 'n ewekansige steekproef uit die universum van toelaatbare toestande.

(c) Die persone wat in die ondersoek gebruik word is 'n ewekansige steekproef uit die betrokke populasie.

Afgesien van dié aannames is dit 'n vereiste dat 'n universum van veralgemening baie duidelik gespesifiseer moet word, sodat met sekerheid bepaal kan word of 'n spesifieke toestand daartoe behoort of nie. Aannames oor die vorm van die verdelings van die fasette, persoonseienskappe of van gelyke variansies is nie nodig nie, aangesien toetse van statistiese betekenisvolheid nie eintlik in die teorie gebruik word nie.

10.4 EKSPERIMENTELE ONTWERPE VIR BEPALING VAN VERALGEMEENBAARHEID

In die teorie van veralgemeenbaarheid word kenmerkende terminologie gebruik wat soms verskil van die in gewone variansie-ontleding. So word die begrippe *fasette* en *toestande* gebruik as sinoniem vir faktore en peile (levels).

Dit is belangrik dat veralgemeenbaarheidsondersoeke deeglik beplan moet word om 'n skatting van variansiekomponente waarin 'n gebruiker

mag belangstel, moontlik te maak. So kan 'n faset volgens een van die volgende drie wyses gehanteer word:

(a) Die faset word verteenwoordig deur 'n steekproef van twee of meer toestande daarvan.

(b) Die faset word gekontroleer deur 'n enkele konstante toestand vir alle waarnemings te gebruik.

(c) Die faset word nie gekontroleer nie en toestande daarin wissel sonder enige eksperimentele beheer.

Dit is baie moeilik om alle toestande eksperimenteel te beheer en daarom word in die praktyk dikwels ongelukkig die weg van die minste moeite, naamlik (c) hiervoor gevolg. Hierdie optrede het tot gevolg dat die variansie van die ongekontroleerde aspekte (fasette) dan met die residue variansie vervleg is en nie onderskei en ondersoek kan word nie.

Uit die voorgaande volg dat daar velerlei ondersoekontwerpe kan wees en om 'n aanduiding hiervan te kry word voorbeelde van enkele ontwerpe kortliks verskaf.

Ontwerpe wat voorsiening maak dat elke persoon p onder elke toestand i waargeneem word, heet 'n *gekruisde* (crossed) ontwerp en kan aangedui word as i x p (i is gekruis met p). 'n Situasie waar elke persoon op verskillende geleenthede n_j gelyktydig deur elkeen van 'n groep waarnemers n_i waargeneem word, word aangedui as i x j x p. Hierdie ontwerp lewer dus 'n telling vir elke persoon p vir elke gepaarde toestand ij.

Indien die situasie gewysig word sodat die waarnemers nie meer die waarnemings gelyktydig doen nie, maar elkeen volgens sy eie tye, bestaan daar vir elke waarnemer i 'n afsonderlike stel waarnemingsgeleenthede n_j. In dié geval word gesê dat waarnemingstyd genestel is in die waarnemer en dit word aangedui as $j{:}i$. Die ontwerp van so 'n ondersoek word dan aangedui as $(j{:}i)$ x p.

Die aard van 'n ontwerp bepaal die veralgemenings wat uit 'n ondersoek gemaak kan word. 'n Enkelfasetontwerp verskaf min inligting oor aspekte wat die meting kan beïnvloed, aangesien daar nie onderskei kan word of toestande in ander fasette byvoorbeeld k (ouderdom) en l (kwalifikasie), konstant gehou word, of toegelaat word om saam met die faset van meting i te wissel nie. So 'n ontwerp het dus min veralgemeenbaarheidsmoontlikhede.

Uit 'n tweefasetontwerp kan meer inligting verkry word met gevolglik veralgemeningsmoontlikhede na

(i) die universum van alle toelaatbare waarnemings wat dus alle i en j insluit (i en j is die twee fasette);

(ii) 'n beperkte universum waar i vas is en j toegelaat word om te wissel; en

(iii) 'n beperkte universum waar j vas is en i toegelaat word om te wissel.

Slegs enkele moontlike ontwerpe is geskets maar dit is duidelik dat vele variasies kan voorkom en dat dit uiters kompleks kan word. 'n Volledige bespreking val buite die oogmerk van hierdie publikasie en vir verdere inligting kan die werke van Huysamen (1980) en veral Cronbach et al. (1972) geraadpleeg word.

10.5 VERALGEMEENBAARHEID- EN BESLUITNEMINGSONDERSOEKE

Dit is wenslik om in 'n bespreking oor veralgemeenbaarheid ook kennis te neem van bogemelde begrippe. In wese is daar nie verskil tussen die twee soorte ondersoeke nie maar om klemverskille in uitgangspunte duideliker te maak is dit nuttig om onderskeid tussen veralgemeenbaarheids- of inligtinginsamelingsondersoeke en besluitnemingsondersoeke te tref. In die literatuur word dikwels ook van die afkortings G-ondersoeke (generalisability studies) en D-ondersoeke (decision studies) gebruik gemaak.

Die doel van G-ondersoeke is insameling van genoeg inligting sodat skattings gemaak kan word van die variansiekomponente volgens sekere omstandighede (met ander woorde stelle toestande). G-ondersoeke be-

hoort so breed te wees dat genoeg inligting beskikbaar is vir veral= gemening na verskillende universums volgens die behoeftes van die waarskynlike gebruikers. Dit sou byvoorbeeld wenslik wees dat G-ondersoeke 'n deel moet uitmaak van die ontwikkelingsproses van enige meetinstrument, sodat daar voldoende inligting later vir die gebrui= ker sal wees.

In teenstelling met die breë ontwerp van G-ondersoeke is D-ondersoeke gemik om meer spesifieke inligting met die oog op besluitneming te verskaf. D-ondersoeke word gebaseer op inligting uit G-ondersoeke en dit is dus belangrik dat die G-ondersoeke so beplan word dat die uni= versums van veralgemening van D-ondersoeke waarin die gebruiker waar= skynlik sal belangstel, ooreenstem of bevat is in die universum van toelaatbare waarnemings van die G-ondersoek.

10.6 DIE ENKELFASET GEKRUISDE ONTWERP AS VOORBEELD

10.6.1 Die skatting van variansiekomponente

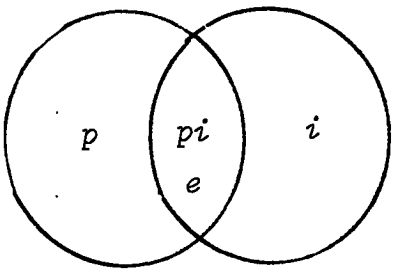
Variansiekomponente is die kerninligting wat gebruik word om die in= vloed van die verskillende aspekte op die meting te bepaal en dit is wenslik dat verdere aandag aan die saak bestee word. 'n Enkelfaset gekruisde ontwerpondersoek sal vervolgens in fyner besonderhede be= spreek word om aan te toon hoe die variansiekomponente geskat en 'n veralgemeenbaarheidskoëffisiënt bereken word.

Soos reeds vermeld, behels hierdie ontwerp dat elkeen van die persone p (wat 'n ewekansige steekproef uit die betrokke populasie is) 'n behan= deling volgens elk van die toestande i (wat 'n ewekansige steekproef is van die universum van toelaatbare toestande in die faset) ontvang. As voorbeeld kan dien 'n situasie waar elkeen van 'n groep persone 'n reeks enersoortige toetse afgelê het. Die ontwerp kan aangedui word as $i \times p$. Die tellings X_{pi} word skematies in figuur 10.1 voorgestel deur 'n tweerigtingtabel en die variansies deur 'n Venndiagram.

Toestande van faset i (Toetse)

Persone p	A	B	C
1	X_{1A}	X_{1B}	X_{1C}
2	X_{2A}	X_{2B}	X_{2C}
3	X_{3A}	X_{3B}	X_{3C}
4	X_{4A}	X_{4B}	X_{4C}
5	X_{5A}	X_{5B}	X_{5C}

Tweerigtingtabel (tellings)



Venndiagram (Variansies)

FIGUUR 10.1: SKEMATIESE VOORSTELLING VAN 'N ENKELFASET GEKRUISDE-ONTWERP

Uit 'n tweerigtingtabel soos in figuur 10.1 voorgestel, kan drie ver=skillende soorte gemiddeldes bereken word, naamlik

$\bar{X}_{p_i}$: die rygemiddelde van persoon p oor die verskillende toestande i

$\bar{X}_{i_p}$: die kolomgemiddelde van toestand i oor die verskillende perso=ne p

$\bar{X}_{ip}$: die totaalgemiddelde oor persone en toestande

Elkeen van die gemelde gemiddeldes kan beskou word as 'n steekproef van die betrokke universum van gemiddeldes. Die steekproefgemiddel=des kan elkeen gebruik word as 'n onsydige skatter van 'n ooreenstem=mende populasie of universumgemiddelde waar

$\mu_p = E(\bar{X}_{p_i})$: Die universumgemiddelde van elke persoon p wat bekend staan as die *universumtelling* van die persoon. (In die spesiale geval waar al die toestande parallel is, is die universumtelling dieselfde as die ware telling in die klassieke teorie.)

$\mu_i = E(\bar{X}_{i_p})$: Die populasiegemiddelde vir elke toestand i

$\mu_{ip} = E(\bar{X}_{pi})$: Die totaalgemiddelde.

Die telling van 'n spesifieke persoon vir 'n spesifieke stel toestande kan in 'n toutologiese uitdrukking weergegee word as die som van ver= skille tussen sekere gemiddeldes (tellingkomponente) sodat die invloed van die verskillende aspekte daaruit kan blyk (kyk tabel 10.1). Aan= gesien elkeen van die tellingkomponente 'n verdeling het, kan varian= sies ook bereken word.

Venndiagramme kan handig gebruik word om 'n duidelike beeld van die verskillende komponente te kry en dit veral na mate die ontwerpe in= gewikkelder raak. Volgens die Venndiagram in figuur 10.1 is daar by die gekruisde enkelfasetontwerp drie oppervlaktes en dit stem ooreen met die variansiekomponente p, i en (pi, e) soos verkry uit 'n twee= rigtingvariansie-ontleding. Besonderhede oor die komponente waaruit 'n telling van persoon p vir toestand i en ook sy variansie saamgestel kan word, verskyn in tabel 10.1.

TABEL 10.1

TELLING- EN VARIANSIEKOMPONENTE IN 'N ENKELFASET GEKRUISDE-ONTWERP

Bron	Telling komponent	Variansie komponent
Algemeen*	$X_{pi\cdot} = \mu(+)$	$\sigma^2(X_{pi\cdot}) = o(+)$
Persoon (p)	$+\mu_{p\cdot} - \mu$	$+\sigma^2(p)$
Toestand (i)	$+\mu_{\cdot i} - \mu$	$+\sigma^2(i)$
Residu (pi, e)	$+X_{pi\cdot} - \mu_{p\cdot} - \mu_{\cdot i} + \mu$	$+\sigma^2(pi, e)$

* Die telling $X_{pi\cdot}$ asook sy variansie bestaan uit die som van die terme van al die bronne, vandaar die "+"-teken daarnaas en voor die van elke ander bron.

Wat die telling $X_{pi\cdot}$ betref, verteenwoordig

μ	:	die totaalgemiddelde
$\mu_{p\cdot} - \mu$	:	die effek van persoon p
$\mu_{\cdot i} - \mu$	:	die effek van toestand i en
$X_{pi\cdot} - \mu_{p\cdot} - \mu_{\cdot i} + \mu$	:	die residu toe te skryf aan enige ander effek as die genoemdes

Die variansiekomponent van μ is gelyk aan 0 aangesien u 'n konstante is vir die populasie en universum. Die residuterm se variansie sluit, afgesien van die variansie van die interaksie tussen persoon en toe= stand, *ook* die variansie e in wat mag spruit uit wisselinge in enige ander ongespesifiseerde eienskap (faset). Dié twee soorte variansies is gekombineer en verstrengel en kan slegs deur 'n twee- of meerfaset= ondersoek geskei word waar meer as een waarneming vir elke p x i paar, beskikbaar is. Die variansiekomponent $\sigma^2(i)$ is die variansie van kon= stante foute wat geassosieer word met die verskillende toestande, by= voorbeeld die verskille in noukeurigheid van waarneming van die ver= skillende waarnemers. Die variansiekomponent $\sigma^2(p)$ stem ooreen met die ware tellingvariansie van die klassieke teorie. Dit blyk dus dat die begrip van ware telling in die klassieke teorie 'n baie eng beskou= ing is, want dit gee nie inligting oor hoe ander aspekte die meting beïnvloed nie, aangesien hierdie variansie met die interaksievarian= sie gekombineer en verstrengel is.

Die waardes van die telling- en variansiekomponente vir die universum en populasie is onbekend maar kan geskat word deur gebruikmaking van die verwagte gemiddelde som van kwadrate E(GK) aangesien dié uitgedruk kan word as geweegde somme van die variansiekomponente.

Besonderhede oor die berekening van die gemiddelde som van kwadrate (GK) asook oor die geweegde samestelling van die variansiekomponente soos dit van toepassing is op 'n enkelfaset gekruisde ontwerp word in tabel 10.2 weergegee.

TABEL 10.2

BESONDERHEDE VIR BEREKENING VAN GEMIDDELDE SOM VAN KWADRATE EN SAME= STELLINGS VAN VARIANSIEKOMPONENTE

Bron	SK	GV	GK	E(GK)
Persone p	SK_p	(n_p-1)	$SK_p/(n_p-1)$	$\sigma^2(pi, e)+n_i\sigma^2(p)$
Faset i	SK_i	(n_i-1)	$SK_i/(n_i-1)$	$\sigma^2(pi, e)+n_p\sigma^2(i)$
Residu pi	SK_{pi}	$(n_p-1)(n_i-1)$	$SK_{pi}/(n_p-1)(n_i-1)$	$\sigma^2(pi, e)$

SK: Som van kwadrate; GV: Grade van vryheid

Die berekening van die GK is soos vir gewone variansie-ontleding (kyk byvoorbeeld Glass & Stanley (1970)) en daaroor word nie verder uitgebrei nie. Belangrike inligting oor die samestelling van die uitdrukkings van die variansiekomponente kan uit die Venndiagram in figuur 10.1 afgelei word. Die samestelling van die uitdrukking kom ooreen met die segmente in die Venndiagram, daar is naamlik 'n term in die uitdrukking vir elke betrokke segment. Die gewigte kan ook uit die Venndiagram verkry word deur met n_p te vermenigvuldig wanneer 'n indeks *nie* p bevat nie en met n_i wanneer 'n indeks nie i bevat nie.

Soos gemeld is die variansiekomponente in die regterhandse kolom in tabel 10.2 onbekend. Dit kan egter geskat word deur die steekproefwaarde (GK) vir E(GK) te substitueer en dan die vergelykings op te los soos in tabel 10.3 aangedui.

TABEL 10.3
SKATTING VAN VARIANSIEKOMPONENTE

Geskatte variansiekomponent	Steekproefwaardes
$\hat{\sigma}^2(p)$	$(GK_p - GK_{res})/n_i$
$\hat{\sigma}^2(i)$	$(GK_i - GK_{res})/n_p$
$\hat{\sigma}^2(pi, e)$	(GK_{res})

Die resultate van die ontledings oor die skatting van die variansiekomponente is die kerninligting van die ondersoek en word gebruik vir gevolgtrekkings oor die invloed van die verskillende aspekte op die meting en vir die berekening van veralgemeenbaarheidskoëffisiënte.

11.6.2 Bepaling van die veralgemeenbaarheidskoëffisiënt

Soos reeds gemeld is die veralgemeenbaarheidskoëffisiënt (ρ^2) 'n indeks van die verhouding van die variansie van die universumtelling tot die variansie van die verwagte waargenome telling. Die variansie van die universumtelling is per definisie gelyk aan $\sigma^2(p)$. Uit tabel 10.3 kan gesien word dat 'n skatting van die variansie van die universumtelling soos volg gemaak kan word:

$$\hat{\sigma}^2(p) = (GK_p - GK_{res})/n_i$$

waar n_i gelyk is aan die aantal toelaatbare toestande van die enkele faset wat in die G-ondersoek betrek is. Daar kan verder aangetoon word dat in die geval van 'n enkelfaset gekruisde ontwerp die verwagte waarde van die variansie van die waargenome telling gegee word deur

$$\sigma^2(p) + \sigma^2(pi, e)/n_i'$$

waar n_i' gelyk is aan die aantal toelaatbare toestande van die enkele faset wat in die D-ondersoek betrek is. Uit tabel 10.3 kan gesien word dat 'n skatting van $\sigma^2(pi, e)$ gegee word deur

$$\hat{\sigma}^2(pi, e) = GK_{res}$$

Volgens die definisie van die veralgemeenbaarheidskoëffisiënt is die formule vir die berekening daarvan in 'n enkelfaset gekruisde ontwerp die volgende

$$\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(pi, e)/n_i'} \qquad (10.1)$$

Soos reeds gemeld is die waardes van die variansiekomponente in vergelyking 10.1 onbekend maar dit kan geskat word deur die steekproefwaardes soos in tabel 10.3 verstrek daarin te vervang

$$E(\rho^2) = \frac{(GK_p - GK_{res})/n_i}{(GK_p - GK_{res})/n_i + GK_{res}/n_i'} \qquad (10.2)$$

10.6.3 Die veralgemeenbaarheidskoëffisiënt en die intrakorrelasiekoëffisiënt

Volgens die definisie van die veralgemeenbaarheidskoëffisiënt, naamlik die verhouding van die universumtellingvariansie tot die verwagte waargenome tellingvariansie is die koëffisiënt ongeveer gelyk aan die verwagte waarde van die gekwadreerde korrelasie tussen die waargenome en universumtellings (Cronbach, 1972, p. 82).

Hierdie verwagte waarde van die gekwadreerde korrelasiekoëffisiënt, aangedui deur $E(\rho^2)$ is 'n intrakorrelasiekoëffisiënt. Die waarde van $E(\rho^2)$ is onbekend maar kan bereken word uit herhaalde toepassings van die ontwerp op dieselfde groep persone of geskat word uit gegewens van 'n enkele steekproef. Die skatting is nie onsydig nie, maar wel stabiel. Die besonderhede vir 'n enkelfaset gekruisde ontwerp word in formule 10.3 gegee wat in wese dieselfde is as 10.2.

$$\hat{\rho}^2 \approx r_{intraklas} = \frac{GK_p - GK_{res}}{GK_p + \left(\frac{n_i - n_i'}{n_i'}\right) GK_{res}} \qquad (10.3)$$

Dié formule wat volgens Cronbach et al. (1972) deur Buros in hierdie vorm ontwikkel is, maak voorsiening vir berekening van die intraklas= korrelasiekoëffisiënt in gevalle waar die aantal toelaatbare toestande kan verskil van die in die oorspronklike ondersoek. Indien $n_i = n_i'$ vereenvoudig die formule na

$$r_{interklas} = 1 - \frac{GK_{res}}{GK_p} \qquad (10.4)$$

In hierdie geval is die veralgemeenbaarheidskoëffisiënt algebraïes ge= lyk aan Cronbach se koëffisiënt alpha (α) en ook aan die K-R-formule 20. Dit bevestig vorige opmerkings oor die beperkinge van die klas= sieke betroubaarheidskoëffisiënt daar dit as 'n spesiale geval in die veralgemeenbaarheidsteorie beskou kan word.

Besonderhede oor die berekening van die veralgemeenbaarheidskoëffi= siënt by komplekser ontwerpe kan in die werke van Cronbach et al. (1972) en Huysamen (1980) gevind word.

10.6.4 Numeriese voorbeeld

Om die leierskaphoedanighede van 6 persone te evalueer is hulle in 'n leierlosegroepbespreking deur vyf beoordelaars waargeneem en aan die hand van 'n sewepuntskaal beoordeel. Die besonderhede van die beoor= delings verskyn in tabel 10.4.

TABEL 10.4

EVALUERING VAN SES PERSONE DEUR VYF BEOORDELAARS

Persoon	Beoordelaar				
	A	B	C	D	E
1	6	5	7	5	6
2	4	4	5	4	4
3	7	6	7	6	5
4	5	5	6	5	4
5	3	2	4	4	3
6	2	1	3	2	4

Die voorbeeld verteenwoordig 'n enkelfaset gekruisde ontwerp van persone (p) x beoordelaars (i). Indien 'n ewekansige effek variansie-ontleding hierop uitgevoer word, lewer dit die resultate soos in tabel 10.5 weergegee.

TABEL 10.5

VARIANSIE-ONTLEDINGSRESULTATE VIR 6 PERSONE WAT DEUR 5 BEOORDELAARS GEëVALUEER IS

Bron	GV	SK	GK
Persone (p)	6 - 1	55,067	11,013
Faset (i)	5 - 1	7,133	1,783
Residu (pi)	(6 - 1)(5 - 1)	9,266	0,463
Totaal	29	71,466	-

Vir die enkelfaset gekruisde ontwerp kan ρ^2 bereken word volgens formule 10.4.

$$\hat{\rho}^2 \approx r_{interklas} = 1 - \frac{0,463}{11,013}$$

$$= 0,958$$

Indien belanggestel word om te weet of drie beoordelaars nog 'n aanvaar= bare veralgemeenbaarheidskoëffisiënt sal lewer, kan formule 10.3 gebruik word.

$$\hat{\rho}^2 \approx r_{intraklas} = \frac{11{,}013 - 0{,}463}{11{,}013 + \left(\frac{5 - 3}{3}\right)0{,}463}$$

$$= 0{,}932$$

Drie beoordelaars lewer dus nog 'n heeltemal aanvaarbare veralgemeen= baarheidskoëffisiënt. Dit beteken weer dat die gemiddelde beoordeling van drie waarnemers met aansienlike vertroue in die plek van die gemid= delde beoordeling van vyf waarnemers gebruik kan word.

LITERATUURLYS VIR HOOFSTUK 10

1 CRONBACH, L.J., GLESER, G.C., NANDA, H., RAJARATNAM, N. *The dependability of behavioral measurements*. New York: Wiley & Sons, 1972.

2 GLASS, G.V. & STANLEY, J.C. *Statistical methods in education and psychology*. Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

3 GOUWS, L.A., LOUW, D.A., MEYER, W.F. & PLUG, C. *Psigologiewoordeboek*. Johannesburg: McGraw-Hill, 1979.

4 HUYSAMEN, G.K. *Psychological test theory*. Durbanville: Boschendal, 1980.

HOOFSTUK 11
GELDIGHEID

11.1 INLEIDING

Die geldigheid van 'n meetinstrument word dikwels gedefinieer as die mate waarin dit aan sy doel beantwoord. 'n Evaluering van 'n toets se geldigheid geskied dus altyd met betrekking tot 'n spesifieke gebruik daarvan. Dit is moontlik dat 'n toets vir 'n bepaalde doel geldig is maar vir 'n ander doel beslis nie of vir 'n bepaalde beslissing wel en vir 'n ander een nie. Vir elke moontlike beslissing (of gebruik of afleiding) wat op grond van toetstellings gemaak gaan word, sal die geldigheid afsonderlik vasgestel moet word.

Dit is dan ook die rede waarom daar dikwels beweer word dat dit sinne= loos is vir 'n toetsouteur om te beweer dat sy toets "geldig" is sonder om terselfdertyd te sê waarvoor dit geldig is (watter afleidings op grond van toetsprestasie geregverdig is) of na watter soort geldig= heid hy verwys.

Aan die ander kant is dit ook so dat aan alle doelstellings ten aan= sien van toetsgebruik 'n gemeenskaplike kenmerk ten grondslag lê wat 'n mate van regverdiging bied om in die algemeen van die begrip "geldig= heid" te praat. By sielkundige toetsing gaan dit nooit om die toets= gedrag vanself nie - dit moet altyd "verleng" word na iets anders, 'n begrip of 'n kriterium wat buite die toetsgedrag geleë is. In hierdie sin het die begrip "geldigheid" betrekking op die vraag of die sprong van die toetsgedrag na iets anders, die "verlenging" van die toetsge= drag, verantwoord is. Die insameling van bewyse vir hierdie regver= diging word die valideringsproses genoem en die mate waarin die sprong vanaf toetsgedrag na iets anders geregverdig is, word geëvalueer in die lig van die geldigheidsgegewens wat beskikbaar is oor die toets.

Daar word dikwels onderskei tussen verskillende soorte geldigheid op grond van die plek en die rol wat die kriterium in die gedagtegang inneem. By een soort geldigheid gaan dit uitsluitlik om die krite= rium vanself en het 'n hoë geldigheidsin omdat die kriterium beter voorspel kan word.

By die tweede soort geldigheid gaan dit om die kriterium as verteenwoordigend van 'n hipotetiese konstruk (begrip). In dié geval moet die toets-kriterium-korrelasie hoog wees omdat deur dié verband iets sinvols gesê kan word aangaande die begrip self. Hier het die valideringsproses alleen sin in 'n raamwerk van die betekenisanalise van die toets. Die vraag is nie wat die toets kan *voorspel* nie, maar wat die toets *meet*, met watter begrip die toetsgedrag verklaar kan word.

Kortliks dus, die geldigheid van 'n toets dui aan in hoeverre dit aan sy doel beantwoord en dit moet altyd gebaseer word op die relasie tussen die toets en 'n kriterium. 'n Nadere differensiëring van die geldigheid van 'n toets is gegrond op 'n differensiëring van die rol van die kriterium: Is dit 'n doel op sigself of het die kriterium slegs sin as operasionalisering van 'n hipotetiese begrip? Die plek en rol van die kriterium in die valideringsproses sal hopelik duideliker word in die hieropvolgende bespreking van verskillende soorte geldigheid.

11.2 VERSKILLENDE SOORTE GELDIGHEID

Daar word verskillende klassifikasies van geldigheid deur verskillende skrywers gemaak. Die belangrikste klassifikasie is dié van 'n spesiale komitee van die American Psychological Association (1954, 1966). Hulle onderskei tussen die volgende drie tipes geldigheid:

(a) Kriteriumverwante of empiriese geldigheid,
(b) inhoudsgeldigheid en
(c) konstrukgeldigheid.

Vir die meeste skrywers (byvoorbeeld die APA se spesiale komitee) het elk van die verskillende soorte geldigheid besondere relevansie tot 'n spesifieke doel met die toetsing. Die verskillende soorte geldigheid en die situasies waarop elkeen betrekking het, word soos volg in die APA se "Recommendations" beskryf:

Inhoudsgeldigheid, waarna soms verwys word as "geldigheid per definisie" of "logiese geldigheid" het hoofsaaklik betrekking op 'n beoordeling van die verteenwoordigendheid van die toetsitems ten opsigte van die betrokke universum van gedragswyses. Dit is veral ter sprake wanneer die toetsgebruiker wil bepaal hoe 'n persoon op die huidige tydstip

sal presteer in 'n gegewe universum van situasies waarvan die toetssitu=asie 'n steekproef uitmaak.

Kriteriumverwante of empiriese geldigheid wat voorspellingsgeldigheid en gelyktydige geldigheid insluit, verwys na die verband, gewoonlik uitgedruk as 'n korrelasiekoëffisiënt, tussen toetstellings (voorspel=lertellings) en kriteriumtellings. *Voorspellingsgeldigheid* is veral belangrik wanneer die toetsgebruiker wil voorspel hoe 'n persoon in die toekoms sal presteer in die toets of 'n ander eksterne veranderlike. *Gelyktydige of saamvallende geldigheid* is ter sprake wanneer die toets=gebruiker 'n skatting wil maak van 'n persoon se huidige status op een of ander veranderlike wat ekstern is aan die toets.

Konstrukgeldigheid het betrekking op die mate waarin 'n toets 'n teore=tiese konstruk of trek meet. Dit is belangrik wanneer die toetsgebrui=ker 'n skatting wil maak van die mate waarin een of ander trek of kwali=teit (konstruk), wat aanvaar word dat dit in die toetsgedrag weerspieël word, aanwesig is.

In hul "Recommendations" wys die APA se spesiale komitee daarop dat in=houdsgeldigheid besonder belangrik is in die geval van prestasietoetse, dat voorspellingsgeldigheid geëvalueer moet word wanneer 'n latere krite=riummeting soos skoolsukses, beroepsukses of reaksie op terapie voorspel moet word, dat gelyktydige geldigheid van belang is by psigiatriese sif=tingsinstrumente, by toetse om beroepsgroepe te differensieer of om pa=siënte te klassifiseer en laastens dat konstrukgeldigheid veral van be=lang is wanneer die toetsafnemer geen definitiewe kriteriummeting van die eienskap waarin hy belangstel tot sy beskikking het nie, of wanneer daar geen operasionele definisie van die konstruk bestaan nie.

Die voorafgaande beteken nie dat die verskillende soorte geldigheid onderling uitsluitend is nie. In die hieropvolgende vollediger bespre=king van elke tipe geldigheid sal die onderlinge verband tussen al drie soorte hopelik duideliker word.

11.3 INHOUDSGELDIGHEID

Hier moet onderskei word tussen *gesigsgeldigheid* (wat Huysamen, 1978, *voorkomsgeldigheid* noem) en *logiese* of *steekproefgeldigheid*.

11.3.1 Gesigsgeldigheid

Anastasi (1964) wys daarop dat inhoudsgeldigheid nie met gesigsgeldigheid verwar moet word nie. Sy skryf: *Content validity should not be confused with face validity. The latter is not validity in the technical sense; it refers not to what the test actually measures but to what it appears superficially to measure* (p. 138). Sy meen dat dit in die reël belangrik is dat 'n toets gesigsgeldigheid behoort te besit aangesien die afwesigheid daarvan moontlik tot swak samewerking (rapport) kan lei. Cattell (1964) stem in 'n mate saam met Anastasi maar wys ook daarop dat hoe meer gesigsgeldigheid daar in toetsitems is, hoe makliker kan toetstellings verdraai word. Hy skryf: *In some trivial sense face or faith validity perhaps still has a role, but in diplomacy rather than psychology, as when an industrial psychologist is pressured to make tests which a chief executive will, from the depths of his ignorance, commend or sanction as measuring what he conceives to be this or that trait* (p. 8).

11.3.2 Logiese geldigheid

Omdat die evaluering van inhoudsgeldigheid essensieel 'n rasionele en logiese beoordeling van toetsinhoud behels, word dit soms *logiese* of *steekproefgeldigheid* genoem. Die belangrikste oorweging by die evaluering daarvan is of die toetsitems 'n bevredigende steekproef is van 'n spesifiek gedefinieerde universum van gedrag (take). Behalwe miskien in die geval van skolastiese prestasietoetse is dit nie 'n maklike taak nie aangesien die inhoudsuniversum in die reël net teoreties bestaan.

Hoewel dit die verantwoordelikheid van die toetsopsteller is om toe te sien dat sy meetinstrument inhoudsgeldigheid besit, is dit wenslik dat hy die items aan 'n paneel van bekwame beoordelaars voorlê. Uiteindelik verwag hy dat sy instrument deur ander verstaan en aanvaar word en dus is dit noodsaaklik dat hy reeds in 'n vroeë stadium van toetsontwikkeling die opinies van ander spesialiste op die gebied kry.

Wanneer 'n toets aan so 'n paneel voorgelê word vir die evaluering van die inhoudsgeldigheid daarvan, is dit die taak van die paneel om elke item te beoordeel ten opsigte van die veronderstelde relevansie daar-

van tot die eienskap wat die toets voorgee om te meet. Om dit te kan doen is eerstens 'n duidelike omskrywing van die inhoudsgebied noodsaaklik. Met ander woorde, die universum van take moet so goed moontlik gedefinieer word. Tweedens is 'n analise nodig van die totale gebied in kategorieë wat al die vernaamste aspekte van die gebied verteenwoordig. Die beoordelaars moet dus voorsien word van spesifieke aanwysings om hul beoordeling te maak. By skolastiese prestasietoetse word dikwels van tweerigtingspesifikasietabelle gebruik gemaak met verskillende inhoudsaspekte aan die een kant en doelstellings aan die ander kant. Hieruit is dit ook duidelik dat inhoudsgeldigheid reeds van meet af aan in 'n toets ingebou kan word.

Uiteindelik is 'n toets net so goed as die toetsouteur en die beoordelaars van die toetsitems. As daar 'n onuitgesproke aanname van die kant van albei is dat sekere inhoudsaspekte onbelangrik is, sal items wat die betrokke aspekte dek, nie in die toets verskyn nie. Huysamen (1978) wys ook daarop dat deskundiges kan verskil oor die definisie van die universum, dit wil sê oor die vraag of sekere take belangrik is of relevant is. Sy aanbeveling hieroor stel hy soos volg: *Die opdrag van die deskundige is egter om, gegee die toetsopsteller se universumdefinisie, te beoordeel of (i) die gekose items verteenwoordigend is van die take of situasies wat in die universumdefinisie gespesifiseer is, en (ii) of hierdie take of situasies bevredigend in die items ingebou is. Indien die meerderheid deskundiges op die gebied saamstem dat aan hierdie vereistes voldoen is, dan moet die inhoudsgeldigheid van die toets as bevredigend beskou word* (p. 99).

Die meeste skrywers is dit eens dat die inhoudsgeldigheid van 'n meetinstrument nie in terme van 'n kwantitatiewe indeks (byvoorbeeld 'n korrelasiekoëffisiënt) uitgedruk kan word nie. Vir Helmstadter (1970, p. 299) egter, maak die mees gesofistikeerde vorm van inhoudsvalidering gebruik van die tegniek van faktorontleding om te bepaal in watter mate 'n toets verskillende inhoudsgebiede meet. Omdat 'n faktorontleding die inwin van empiriese gegewens in verband met 'n verskeidenheid van metings behels, klassifiseer die meeste skrywers eerder faktoriale geldigheid onder een of ander vorm van empiriese geldigheid. Helmstadter redeneer dat dit logies beter inpas by wat hy as inhoudsgeldigheid beskou wanneer die hoofdoel met die faktorontleding die analise van toetsinhoud is.

Vir die eerste toetsopstellers was inhoudsgeldigheid (veral gesigsgel-

digheid) die vernaamste kriterium by die opstel van vraelyste en daar is min aandag gegee aan die voorspellingswaarde van items. Later het die pendulum na die teenoorgestelde kant geswaai en het sommige skrywers aangevoer dat iteminhoud irrelevant is en dat groter klem op empiriese geldigheid geplaas moet word. Op grond van heelwat navorsing (byvoorbeeld Duff, 1965; Norman, 1963; Neill & Jackson, 1970; Goldberg & Slovic, 1967) kan die stelling gemaak word dat inhoudsgeldigheid 'n noodsaaklike (maar nie voldoende nie) voorwaarde is vir empiriese geldigheid. Tenopyr (1977) meen ook dat daar 'n noue verband bestaan tussen inhoudsgeldigheid en konstrukgeldigheid. Sy skryf: *At the minimum, content validation may be considered propaedeutic to construct validation. At the maximum, content validity may be assumed to be one type of evidence of construct validity* (p. 51).

Tenopyr stel verder voor dat met inhoudsgeldigheid volstaan kan word net in die geval van aangeleerde vaardighede, eenvoudige (in teenstelling met komplekse) konstrukte soos goed gedefinieerde fisiese eienskappe, spesifieke kennis, goed gedefinieerde voorkeure (byvoorbeeld vir skofwerk) en so meer. Haar algemene raad is: *If you want to use inferences about test construction to justify inferences about test scores, stay with simple, well defined constructs with easily observable manifestations* (p. 54).

Cronbach (1971) wys daarop dat: *It is improper to report high item-test correlations as evidence of content validity* (p. 457). Vir Cronbach is daar niks in die logika van inhoudsvalidering wat vereis dat die universum van die toets homogeen moet wees wat die inhoud betref nie: *Low item intercorrelations do not necessarily imply failure of the test content to fit the definition. Indeed, if the universe is heterogeneous, consistently high item intercorrelations imply inadequate sampling ... When the test constructor routinely discards the items whose intercorrelations with the total score for the pool are low, he risks making the test less representative of the defined universe* (p. 457). Hiermee sal Cattell waarskynlik saamstem. Hy voer oortuigende argumente aan (1964) dat 'n *baie* homogene vraelys (as die betrokke gedragsuniversum heterogeen is) net in sekere situasies of by sekere steekproewe bevredigende empiriese geldigheid sal toon. Met ander woorde, 'n vraelys met hoë homogeniteit wat inhoud betref, sal minder oordraagbaarheid besit as 'n vraelys waarin 'n wye verskeidenheid toepaslike situasies betrek word.

11.4 KRITERIUMVERWANTE GELDIGHEID

Kriteriumverwante geldigheid kan gedefinieer word as die akkuraatheid waarmee tellings in die toets die tellings in 'n kriterium voorspel. Twee soorte kriteriumverwante geldigheid kan onderskei word, naamlik *voorspellingsgeldigheid* en *gelyktydige of saamvallende* geldigheid wat hoofsaaklik verskil met betrekking tot die tyddimensie. Albei soorte geldigheid staan in verband met die gebruik van meetinstrumente vir die neem van besluite soos byvoorbeeld by die keuring van aansoekers om werk, die klassifikasie van personeel in verskillende werksrigtings, of die kliniese diagnose van 'n pasiënt vir nuwe opname in 'n inrigting vir geestesversteurdes. In al die genoemde gevalle word die toetstelling gebruik om persone se prestasies of posisies ten opsigte van 'n ander veranderlike - die kriterium - te voorspel.

11.4.1 Voorspellingsgeldigheid

Die evaluering van 'n toets as voorspeller is primêr 'n empiriese en statistiese evaluering. Vir hierdie rede word daar soms gepraat van *empiriese* of *statistiese* geldigheid. Dit gaan hoofsaaklik om die korrelasie tussen toetstellings en een of ander geskikte kriteriummeting van sukses. Hierdie korrelasie word dikwels die *geldigheidskoëffisiënt* genoem. 'n Geldigheidskoëffisiënt het betrekking op die voorspelling van 'n spesifieke kriterium met 'n spesifieke voorspeller vir 'n spesifieke populasie proefpersone.

Die basiese prosedure is gewoonlik om die toets toe te pas op 'n groep aan die begin van 'n opleidings- of werkprogram, hulle later op te volg en vir elkeen in die steekproef (groep) 'n gespesifiseerde kriteriummeting te verkry en korrelasies te bereken tussen toetstellings en kriteriummetings. Hoe hoër die korrelasie, hoe effektiewer verrig die toets die doel waarvoor dit opgestel is.

'n Mens vra jouself dikwels af hoe hoog die korrelasie tussen 'n toetstelling en 'n kriteriumtelling moet wees voordat die toets as 'n nuttige voorspeller beskou kan word. Een moontlike manier om die praktiese betekenisvolheid van 'n korrelasiekoëffisiënt te beoordeel, word in tabel 11.1 gegee.

TABEL 11.1

AKKURAATHEID VAN VOORSPELLING VIR 'N AANTAL WAARDES VAN DIE GELDIG= HEIDSKOëFFISIëNT (1 000 GEVALLE IN ELKE RY OF KOLOM)

Geldigheids= koëffisiënt	25 % van groep volgens voorspeller	25 % van groep volgens kriterium	1ste	2de	3de	4de
0,00		1ste	250	250	250	250
		2de	250	250	250	250
		3de	250	250	250	250
		4de	250	250	250	250
0,40		1ste	428	277	191	104
		2de	277	277	255	191
		3de	191	255	277	277
		4de	104	191	277	428
0,50		1ste	480	279	168	73
		2de	279	295	258	168
		3de	168	258	295	279
		4de	73	168	279	480
0,60		1ste	537	277	141	45
		2de	277	318	264	141
		3de	141	264	318	277
		4de	45	141	277	537
0,70		1ste	601	270	107	22
		2de	270	253	270	107
		3de	107	270	253	270
		4de	22	107	270	601

Die gegewens in tabel 11.1 behoort soos volg geïnterpreteer te word:

As die voorspeller en kriterium gesamentlik normaal verdeel, kan vir 'n gegewe geldigheidskoëffisiënt afgelees word hoe 1 000 persone wat volgens die voorspeller val in die beste 25 %, tweede beste 25 %, derde beste 25 % of vierde beste 25 % van die populasie, waarvoor die geldigheidskoëffisiënt bepaal is, volgens die kriterium sal ver= deel in die beste 25 %, tweede beste 25 %, derde beste 25 % en vierde beste 25 % van genoemde populasie. As byvoorbeeld die geldigheids= koëffisiënt 0,40 is, sal uit 1 000 persone wat volgens die voorspeller in die derde beste 25 % van die populasie val, 191 val in die beste 25 %, 255 in die tweede beste 25 %, 277 in die derde beste 25 % en 277 in die vierde beste 25 % van die populasie soos bepaal deur die kriterium.

Die gebruik van tabel 11.1 kan aan die hand van die volgende voorbeeld verduidelik word:

Veronderstel 'n departement van Wiskunde aan 'n universiteit wil volgende jaar minstens 25 studente tot die eerstejaarkursus toelaat, maar wil verder vereis dat 'n aansoeker om toegelaat te word minstens 'n 75 % kans moet hê om op universiteit beter te presteer as die swakste 50 % van vanjaar se eerstejaarstudente. Veronderstel verder dat die korrelasie= koëffisiënt van Wiskunde op skool (voorspeller) en eerstejaareksamen= punte in Wiskunde (kriterium) vir vanjaar se studente 0,40 is. Daar word ook aanvaar dat evalueringstandaarde en sillabusse op skool en universiteit en die verspreiding van vermoëns in die aansoekgroepe nie noemenswaardig van jaar tot jaar verander nie.

Indien daar 100 aansoekers is en volgens Wiskundepunt op skool die beste 50 % aansoekers toegelaat word (wat meer is as die vereiste minimum van 25), sal volgens tabel 11.1 (277 + 277 + 277 + 428) uit elke 2 000 van die toegelatenes oftewel 63 % van die toegelate aan= soekers (wat minder is as die vereiste minimum van 75 %) waarskynlik op universiteit op die vereiste vlak presteer. Deur minder aansoekers toe te laat, kan hulle kans om op die vereiste vlak te presteer, ver= hoog word. Veronderstel nou dat volgens Wiskundepunt op skool die beste 25 % aansoekers toegelaat word (wat presies gelyk is aan die vereiste minimum van 25). Volgens tabel 11.1 sal (277 + 428) uit elke 1 000 oftewel 71 % van die toegelate aansoekers (wat nog steeds

minder is as die vereiste minimum van 75 %) waarskynlik op universiteit op die vereiste vlak presteer. 'n Keuringstrategie wat die vereiste resultate kan oplewer, kan dus nie gevolg word deur van wiskundeprestasie op skool gebruik te maak nie. 'n Voorspeller met 'n hoër korrelasiekoëffisiënt as 0,40 met die kriterium is nodig indien die keuringsvereistes nie verlaag mag word nie.

a. Kriteriumprobleme

Dat die kriterium waarteen 'n toets se geldigheid bepaal word, probleme kan skep, is reeds vyftig jaar gelede deur Hull (aangehaal deur Guion, 1974) ingesien: ... *the most formidable problem encountered by the aptitude psychologist is the location of a trial group of subjects from whom a valid and reliable quantitative criterion of aptitude may be obtained* (p. 287).

Voorheen is daar verwys na die korrelasie tussen toetstellings en *geskikte kriteriummetings*. Die frase *geskikte kriteriummeting* kan probleme veroorsaak. Een van die moeilikste probleme wanneer dit gaan om keuring- en plasingtoetse is om 'n bevredigende meting van uiteindelike gedrag te vind (of te skep) wat kan dien as 'n kriterium vir die bepaling van die voorspellingsgeldigheid van 'n toets. Daar word dikwels pogings aangewend om toetsgebruike te valideer teen kriteria wat onmiddellik beskikbaar is, maar nie relevant is nie. Vir hierdie rede beklemtoon skrywers (byvoorbeeld Magnussen, 1967; De Groot, 1961) die feit dat daar onderskei moet word tussen *ware* kriteria en tussentydse of substituutkriteria.

Verder is daar baie soorte werkers (byvoorbeeld geneeshere, verpleegsters, onderwysers, skakelbeamptes) vir wie objektiewe rekords van prestasie feitlik nooit beskikbaar is nie. Wanneer daar wel rekords bestaan, kan 'n persoon se sukses nog deur verskillende faktore wat buite sy eie beheer is, beïnvloed word. 'n Assuransieverkoper se sukses is byvoorbeeld 'n funksie nie alleen van sy eie doeltreffendheid as verkoopsman nie, maar ook van die gebied waarin hy moet werk; 'n fabriekswerker se doeltreffendheid kan ook afhang van die tipe apparaat waarmee hy moet werk.

Die vier basiese vereistes vir wenslike kriteriummaatstawwe kan kortliks in volgorde van belangrikheid soos volg opgesom word:

(i) Relevansie

'n Kriterium is relevant in die mate waarin 'n persoon se stand ten opsigte van 'n kriteriummeting ooreenkom met sy werklike sukses in 'n opleidingsprogram of werk. Geen empiriese bewyse om dit te beoordeel kan gevind word nie. In die geval van die inhoudsgeldigheid van prestasietoetse is genoem dat 'n mens jou moet verlaat op die beste beskikbare professionele oordeel om te bepaal of die toetsinhoud akkuraat die betrokke doelwitte verteenwoordig. Op dieselfde manier is dit nodig dat wat kri= teriummetings betref, daar staatgemaak word op professionele oordeel aangaande die mate waarin 'n beskikbare (gedeeltelike) kriteriummeting relevant is tot die uiteindelike kriterium van sukses.

(ii) Sydigheid

Die kriteriummeting moet sodanig wees dat irrelevante faktore nie 'n persoon se telling beïnvloed nie. Daar behoort dus daar= teen gewaak te word dat faktore soos geslag en ras 'n rol speel in kriteriummetings wat nie geregverdig kan word nie.

(iii) Betroubaarheid

'n Kriterium wat vanself nie betroubaar is nie, kan nie deur iets anders voorspel word nie.

(iv) Beskikbaarheid of gerieflikheid

Daar is gewoonlik praktiese probleme in verband met die beskik= baarheid van 'n kriterium (byvoorbeeld die onkoste om kriterium= metings te verkry of die ontwrigting in 'n opleiding- of werksi= tuasie om toepaslike gegewens in te win). Enige keuse van 'n geskikte kriterium moet dus praktiese beperkings in ag neem.

Laastens kan in gedagte gehou word dat

(a) daar gewoonlik baie soorte kriteriummetings is wat verkry kan word vir die validering van seleksietoetse en

(b) byna alle kriteriummetings net 'n gedeelte van werksukses of net die voorvereistes vir latere werklike sukses meet.

b. Kruisvalidering

Ten slotte is dit van groot belang, veral in gevalle waarin die oorspronklike geldigheidsberekening eksploratief van aard was, om 'n kontrolevalidasie (kruisvalidasie) uit te voer op 'n nuwe, onafhanklike steekproef. Die kontrolevalidasie kan gesien word as hipotesetoetsing ook wat die sterkte van die verband betref, mits vooraf eksplisiete verwagtinge oor die (minimum) sterkte van die verband uitgespreek is.

Wat in die oog gehou moet word, is dat die operasionele opvatting van voorspellingsgeldigheid alleen 'n bevredigende antwoord gee op die vraag na die geldigheid van 'n veranderlike as die veranderlike as voorspeller bedoel is in 'n bepaalde ondersoekkonteks, vir 'n spesifieke voorspellingsdoel waar die kriterium self voldoende (geldig) meetbaar is. (Vergelyk De Groot, 1961.)

11.4.2 Gelyktydige (samevallende, kongruente) geldigheid

In die voorafgaande bespreking oor voorspellingsgeldigheid is aanvaar dat tellings in die toets waarvan die geldigheid bepaal word, eers verkry is en dat die kriteriumtellings later verkry is. Vir verskillende redes word kriteriumgegewens soms tesame met die toetsgegewens ingesamel. Wanneer hierdie metode vir 'n geldigheidsondersoek gebruik word, word gepraat van *saamvallende* geldigheid wat gedefinieer kan word as die akkuraatheid waarmee die toets 'n identifikasie of diagnose van huidige gedrag of status van individue verskaf.

Soos in die geval van voorspellingsgeldigheid is die mees algemene metode om hierdie tipe geldigheid te ondersoek die berekening van 'n korrelasiekoëffisiënt tussen toetsdata en kriteriumdata (Pearson-produk-momentkorrelasie, rangordekorrelasie, biseriële korrelasie, tetrachoriese korrelasie of 'n korrelasie-ratio). Hoe hoër die korrelasiekoëffisiënt (positief) wat verkry word, hoe beter is die kriteriumgeldigheid van die toets.

Die gewone prosedure om saamvallende geldigheid te bereken, is om die

toets op 'n verteenwoordigende steekproef van die populasie vir wie die toets bedoel word, toe te pas. Terselfdertyd word kriteriumgegewens ook ingesamel (byvoorbeeld skoolpunte, beoordelings deur werkgewers, produktiwiteitsindekse, ensovoorts). Hierna kan die verband tussen toetstellings en indekse van kriteriumstatus bepaal word. Een belangrike probleem is dat sulke kriteriummetings dikwels uiters onbetroubaar is.

In die algemeen is voorspellingsgeldigheid veral relevant by aanlegtoetse en belangstellingtoetse wat vir die seleksie en klassifikasie van applikante vir werk en gespesialiseerde opleidingskursusse gebruik word. Saamvallende geldigheid, daarenteen, is belangriker in die geval van persoonlikheids- en skolastiese toetse wat vir die diagnosering van onderskeidelik persoonlikheidsafwykings en skolastiese gebreke gebruik word (Huysamen, 1978, p. 267). Die geldigheid van 'n toets wat bedoel word om persoonlikheidafwykings te diagnoseer, kan dan ondersoek word deur te bepaal of tellings in die toets onderskei tussen pasiënte in die onderskeie psigiatriese kategorieë. Hier kan die geldigheid van psigiatriese diagnoses egter probleme veroorsaak.

11.5 KONSTRUKGELDIGHEID

Die begrip "konstrukgeldigheid" het in die vyftigerjare sy verskyning gemaak in die woordeskat van toetsspesialiste. Die rasionaal daarvoor het ontwikkel uit persoonlikheidstoetsing. Daar was 'n behoefte aan 'n nuwe benadering tot geldigheid omdat nòg inhoudsgeldigheid nòg kriteriumverwante geldigheid as fundamentele doelwit die *begrip* van die trek (konstruk) wat 'n toets meet, gehad het. Vir 'n instrument wat byvoorbeeld 'n persoonlikheidskonstruk soos egosterkte meet, is daar geen unieke, pertinente kriterium om te voorspel nie, of 'n goedafgebakende inhoudsgebied waaruit 'n steekproef gedragswyses getrek kan word nie. Daar is wel 'n teorie wat die veronderstelde aard van die konstruk (trek, eienskap) omskryf. In hierdie lig kan konstrukgeldigheid gedefinieer word as die mate waarin die toets in werklikheid die teoretiese konstruk meet wat dit veronderstel is om te meet.

Sommige skrywers (byvoorbeeld Stilson, 1966) beklemtoon die feit dat 'n enkele kwantitatiewe indeks van konstrukgeldigheid nie voldoende is nie en dat konstrukgeldigheid slegs geëvalueer kan word in die lig van alles wat bekend is oor, of bewyse wat geakkumuleer word aangaande die

konstruk. Uit hierdie oogpunt is Stilson (1966) se definisie toepaslik: *(It is) essentially the amount of integrated knowledge we have of a measure. The more extensive and detailed the empirically verified theoretical context of a scale becomes, the greater is its construct validity* (p. 128).

a. Die nomologiese netwerk

'n Teorie wat 'n verskynsel probeer verklaar, bestaan uit 'n hele reeks onderling verwante begrippe, beginsels en wette. Hierdie stelsel van aaneengeskakelde wette noem Cronbach en Meehl (1955) 'n *nomologiese netwerk*. Die wette bring waarneembare eienskappe in verband met ander waarneembare eienskappe en met teoretiese konstrukte of een teoretiese konstruk met 'n ander een. Die relasies tussen die wette kan òf deterministies òf statisties wees, alles-of-niks van aard wees of dit kan moontlikhede wees (probabilistic). Die essensiële is dat die definisie van die konstruk teruggevoer moet kan word na wette of begrippe wat geanker is in waarneembare gegewens.

Om helderheid te kry oor enige konsep in die netwerk moet die netwerk uitgebrei word om baie regmatige relasies in te sluit of om die bestaande relasies meer spesifiek en definitief te omskryf. Sielkundige toetse is maar een (let wel *net een*) metode om die betekenis van 'n besondere konstruk te ontwikkel.

Die voorafgaande het belangrike implikasies. Om die betekenis van 'n konstruk af te lei moet daar eerstens in een of ander stadium waarneembare gegewens wees wat 'n toets byvoorbeeld kan voorsien. Tweedens moet die gevolgtrekkings en afleidings wat op grond van die waarneembare gegewens aangaande die betekenis van die konstruk gemaak word openlik en eksplisiet gestel word anders kan die geldigheid van die afleidings nie gekontroleer word nie. Derdens, tensy almal wat die konstruk gebruik, nie essensieel dieselfde netwerk in gedagte het nie, is kommunikasie en ooreenkoms tussen verskillende navorsers nie moontlik nie. Almal heg byvoorbeeld nie dieselfde betekenis aan konstrukte soos *intelligensie*, *introversie* of *aggressiwiteit nie*.

b. Metodes vir die bepaling van konstrukgeldigheid

Soos in die geval van inhoudsgeldigheid en kriteriumverwante geldigheid kan konstrukgeldigheid ook met behulp van verskillende tegnieke

ondersoek word. Hierdie tegnieke word deur verskillende skrywers op verskillende maniere geklassifiseer. Huysamen (1978) bespreek byvoorbeeld (a) korrelasionele benaderings (waaronder hy die multitrek-multimetode van Campbell en Fiske, 1959, insluit asook faktorontleding) en (b) die eksperimentele benadering. Thorndike en Hagan (1977) noem vyf kategorieë naamlik (a) intratoetsmetodes, (b) intertoetsmetodes, (c) kriteriumverwante studies, (d) eksperimentele studies (dit wil sê studies wat die manipulering van veranderlikes behels) en (e) "generalizability" studies. Voordat 'n aantal valideringsmetodes bespreek word, is dit belangrik om in gedagte te hou dat *konstrukvalidering hoofsaaklik die ondersoek van voorspellings (die toets van hipoteses) aangaande toetstellings en ander veranderlikes wat uit 'n teorie afgelei is, behels.*

c. Intratoetsmetodes

Hierdie kategorie bestaan uit tegnieke wat die interne struktuur van die toets bestudeer - die inhoud, die onderlinge verband tussen items en subtoetse en die prosesse wat gemoeid is by die beantwoording van die items. Omdat hierdie metodes net te doen het met die interne struktuur van die toets en nie eksterne veranderlikes in ag neem nie, is hulle op sigself nie voldoende om die konstrukgeldigheid van 'n toets en die trek (konstruk) wat dit meet, te steun nie. Met ander woorde, hulle kan iets bydra tot die *begrip* van die aard van die konstruk maar niks omtrent die verband tussen die konstruk en ander veranderlikes (byvoorbeeld wette in die nomologiese netwerk) nie.

Voorspellings moet natuurlik geformuleer word in ooreenstemming met die onderliggende teorie. Indien die teorie hoë item-interkorrelasies voorspel, kan sulke korrelasies konstrukgeldigheid impliseer, anders kan hoë interkorrelasies lei tot 'n vermindering van die konstrukgeldigheid van die instrument (Edmunds & Kendrick, 1980, p. 30).

'n Faktorontleding van die items van die toets kan ook beskou word as konstrukvalidering wanneer die doel met die ontleding is om hipoteses aangaande die struktuur van die instrument te toets.

d. Intertoetsmetode

In hierdie kategorie val al die metodes waar verskeie toetse tegelyk beskou word, maar ander eksterne veranderlikes nie in ag geneem word nie. In die algemeen word die eienskappe wat gemeenskaplik is aan verskillende toetse deur sulke metodes aangedui, maar geen direkte afleidings kan gemaak word aangaande die verband tussen toetstellings en eksterne veranderlikes nie.

Die eenvoudigste metode in hierdie kategorie is om 'n nuwe toets met 'n bestaande toets, waarvan die konstrukgeldigheid aanvaar word, te korreleer. Nuwe intelligensietoetse word byvoorbeeld dikwels met erkende, bestaande toetse vergelyk. In hierdie geval is De Groot (1961) se term *soortgenootgeldigheid* verkieslik omdat dit verwarring met saamvallende geldigheid (bespreek in paragraaf 4.3) uitskakel.

Omdat die betekenis van tellings in die ouer toets reeds in 'n mate vasgestel is, word aanvaar dat tellings in die nuwe toets op dieselfde manier verband hou met ander veranderlikes. Die gevaar by hierdie benadering, tensy die twee toetse baie hoog korreleer (die r ongeveer dieselfde is as die betroubaarheid van die ouer toets), kan die faktore wat meewerk om die korrelasies te verlaag juis die relevante faktore wees wat verantwoordelik is vir die verband tussen die toets en die eksterne veranderlike(s). Gevolglik sal afleidings wat op die basis van die interkorrelasies tussen die toetse gemaak word, nie geldig wees nie.

'n Tweede benadering is om 'n faktorontleding van die interkorrelasies tussen 'n aantal toetstellings (wat praktiese kriteria kan insluit) te doen. Die resultate van die faktorontleding kan verskillende soorte inligting verskaf. Dit sal aantoon watter toetse gemeenskaplike variansie besit en dus dieselfde konstruk meet. 'n Ondersoek van die inhoud van die toetse wat op dieselfde faktor laai, sal lig werp op die aard van die konstruk en 'n naam daarvoor suggereer (Thorndike & Hagen, 1977, p. 147). 'n Faktorontleding sal ook toon hoeveel gemeenskaplike faktorvariansie elke toets besit en in watter mate tellings in die toets afhanklik is van spesifieke variansie. Volgens die definisie van Kerlinger (1964) is die proporsie van die totale variansie in die toetstellings wat gemeenskaplike variansie is, 'n kwantitatiewe indeks van konstrukgeldigheid.

Campbell en Fiske (1959) het die *multitrek-multimetode* voorgestel om korrelasies te bestudeer ten einde 'n uitspraak te maak oor die konstruk= geldigheid van 'n toets sonder om 'n faktorontleding te doen.

Verskillende skrywers (byvoorbeeld Althauser en Herberlein, 1971) meen dat die metode van die genoemde skrywers nie beskou moet word as 'n stel= sel waarvolgens outomaties bepaal kan word of 'n toets konstrukgeldigheid besit of nie. Inderdaad is die hele begrip van konstrukgeldigheid teen= strydig met so 'n moontlikheid. Daar word egter toegegee dat die metode die sinvolle interpretasie van belangrike inligting in verband met kon= strukgeldigheid vergemaklik en dat dit 'n manier voorsien om vas te stel watter stappe moontlik geneem kan word om die meting van 'n besondere trek te verbeter. (Vir 'n bespreking van die verband tussen die metode van Campbell en Fiske en faktorontledings asook variansie-ontleding kyk Boruch, Larkin, Wolins en MacKinney, 1970 en Boruch en Wolins, 1970.)

e. Die multitrek-multimetodetegniek

Campbell en Fiske staan 'n valideringsproses voor wat vereis dat die interkorrelasies bereken word tussen toetstellings van toetse wat min= stens twee trekke verteenwoordig en wat elk deur minstens twee verskil= lende metodes gemeet word. 'n Multitrek-multimetodematriks word gevorm deur die korrelasies tussen (a) tellings in die twee (of meer) trekke wat met dieselfde meetmetode verkry is, (b) tellings in dieselfde trek wat met die verskillende metodes verkry is, en (c) tellings in die ver= skillende trekke wat met verskillende meetmetodes verkry is. Hierdie skrywers meen dat 'n trek nie onafhanklik van een of ander metode gemeet kan word nie. Gevolglik kan 'n gedeelte van die toetsvariansie toege= skryf word aan die besondere metingsmetode eerder as die trek wat ge= meet word. Terwyl hierdie metodevariansie nie 'n besondere probleem is in 'n suiwer empiriese situasie nie (waar die doel bloot voorspelling van 'n kriterium is), kan dit misleidend wees wanneer 'n poging aange= wend word om te bepaal of 'n toets 'n spesifieke trek meet. Dit is denkbaar dat twee afsonderlike toetse 'n redelik hoë verband kan toon bloot omdat dieselfde metode vir die meting van die trek gebruik is en nie omdat hulle inderdaad dieselfde trek meet nie.

Tweedens beklemtoon Campbell en Fiske die belangrikheid van die idee (genoem deur Cronbach en Meehl) dat enige meetmiddel alleen duidelik

beskryf word deur 'n gesamentlike metode van ooreenkomste en verskille; met ander woorde, dit is noodsaaklik om te sê wat 'n toets meet sowel as wat dit nie meet nie. Bewyse vir die konstrukgeldigheid van 'n toets moet derhalwe gebruik maak van twee beginsels naamlik die beginsels van konvergensie en diskriminasie. Volgens eersgenoemde beginsel sal twee metings van dieselfde trek hoog met mekaar korreleer selfs al verskil die meetmetodes. Volgens die beginsel van diskriminasie sal twee metings vir twee verskillende trekke nie hoog met mekaar korreleer nie, selfs al is die meetmetodes dieselfde.

Campbell en Fiske se metode kan egter baie praktiese en teoretiese probleme by die toepassing daarvan oplewer (Lumsden, 1976, p. 269-270). In die eerste plek is dit moeilik om toetse vir dieselfde trek volgens verskillende metodes op te stel. Die probleme kom ooreen met die waar betroubaarheid volgens die metode van parallelle vorms bereken word en 'n toets as betroubaar beskou word slegs wanneer die toets 'n "identiese tweeling" het. Die multitrek-multimetodebenadering om konstrukgeldigheid te ondersoek, vereis klaarblyklik dat 'n toets 'n "identiese tweeling" van teenoorgestelde geslag moet hê! Krause (1972) en Lumsden en Ross (1978) wys ook op verskeie gebreke in dié metode.

f. Kriteriumverwante ondersoeke

Die aard en tipe kriteria wat toetstellings kan voorspel kan ook 'n aanduiding gee van die aard van die konstruk wat die toets meet. Dus kan gegewens uit sulke kriteriumverwante geldigheidstudies relevante inligting verskaf vir die evaluering van konstrukgeldigheid (Thorndike & Hagen, 1977, p. 148).

g. Groepverskille

Die vermoë van toetstellings om tussen groepe wat natuurlik bestaan of eksperimenteel saamgestel is (dit wil sê maklik herkenbare groepe) te onderskei, kan ook ondersteuning bied vir die konstrukgeldigheid van 'n instrument. Getuienis vir die konstrukgeldigheid van 'n toets vir kunsaanleg, byvoorbeeld, kan uitgedruk word in terme van die verskil in gemiddelde tellings van persone in die algemeen, kunsstudente en professionele kunstenaars. As die gemiddelde tellings van die drie groepe toeneem vanaf die eersgenoemde groep tot die laasgenoemde, kry 'n mens meer vertroue in die doeltreffendheid waarmee die toets die taak

verrig waarvoor dit opgestel is.

'n Meetinstrument wat opgestel is op grond van verskille tussen maklik herkenbare groepe (dit wil sê uit die poel items wat toegepas is, is slegs dié gekies wat onderskei het tussen die groepe) beskik natuurlik uit die aard van die saak oor kongruente (gelyktydige, saamvallende) geldigheid. Strong, byvoorbeeld, het hierdie metode gebruik om sy belangstellingsvraelys op te stel (SVIB). Die skaal vir geneeshere bestaan uit die items wat geneeshere onderskei het van mans-in-die-algemeen, dié vir eiendomsagente uit die items wat onderskei het tussen eiendomsagente en mans-in-die-algemeen, en so aan.

'n Alternatiewe benadering is ook moontlik. Groepe kan saamgestel word op grond van hul toetstellings (byvoorbeeld die boonste kwart en die onderste kwart) en die onderskeidende eienskappe van die groepe nagegaan word. Hierdie eienskappe dien dan om die konstruk te definieer. In 'n studie waarin hierdie benadering gebruik is, het Barron (1963, p. 125-126) universiteitstudente wat hoë tellings behaal het op sy toets vir egosterkte, vergelyk met die wat lae tellings behaal het volgens gedragsbeoordelings. Eersgenoemdes is beskryf as wakker, avontuurlustig, gedetermineerd, onafhanklik, uitgesproke, volhardend, betroubaar, vindingryk en verantwoordelik, terwyl die groep met lae tellings beskryf is as afhanklik, verwyfd, oordrewe goedgemanierd.

Sulke beskrywings weerspieël die aard van die konstruk wat deur die toets gemeet word.

Geldigheidskoëffisiënte kan ook relevante gegewens voorsien. 'n Toets wat bedoel is as 'n meetinstrument vir skolastiese aanleg behoort vanselfsprekend prestasie in akademiese vakke te kan voorspel; 'n toets vir vingervaardigheid behoort 'n goeie voorspeller te wees van sukses in beroepe soos byvoorbeeld die van 'n horlosiemaker waar vingervaardigheid 'n essensiële komponent van die taak is; 'n toets vir sosiopatiese neigings behoort jeugmisdadigheid te kan voorspel. Indien daar vir sulke voorspellings ondersteuning gevind word, lei dit weer tot meer vertroue in die toets as 'n meetmiddel van die veronderstelde trek.

h. Eksperimentele manipulering

Weer eens, indien die teoretiese konstruk lei tot 'n voorspelling dat daar 'n verskil in toetstellings sal wees na een of-ander eksperimentele manipulering, sal so 'n uitslag bydra om die konstrukgeldigheid van die instrument en die daarmee saamhangende teorie te ondersteun.

Hierdie kategorie sluit ook getuienis in aangaande verskille wat die gevolg is van natuurlike gebeure. Byvoorbeeld, as die definisie van die betrokke konstruk (trek) impliseer dat die trek besonder stabiel is (min verander met verloop van tyd) en onbeïnvloedbaar is deur omgewingstoestande sal 'n mens in alle waarskynlikheid 'n hoë stabiliteitskoëffisiënt verkry as die toets na 'n redelik lang tydsverloop weer toegepas word. In hierdie geval dien betroubaarheidsgegewens ook as een soort bewys van konstrukgeldigheid.

i. Ander metodes

Oorsaaklike verbande kan nie met behulp van gewone korrelasietegnieke ondersoek word nie. Wanneer 'n teorie 'n oorsaaklike verband tussen konstrukte spesifiseer, kan ander korrelasietegnieke soos *path analysis* gebruik word om oorsaaklikheid (causality) te ondersoek (kyk byvoorbeeld Blalock, 1964; Crano, Kenny & Campbell, 1972; Yee & Gage, 1968).

j. Bydrae van konstrukgeldigheid tot sielkundige toetsing

'n Belangrike uitvloeisel van die konsep van konstrukgeldigheid was om die aandag te vestig op sielkundige toetse as instrumente van sielkundige teorie. Met ander woorde, dit het pertinent die aandag gevestig op die noodsaaklikheid daarvan dat toetskonstruksie op 'n eksplisiet erkende teoretiese grondslag moet berus. Sowel by die samestelling van 'n nuwe toets as by die beplanning van valideringsprosedures is dit nodig dat sielkundige hipoteses geformuleer word. Die voorstanders van konstrukvalidering het probeer om sielkundige toetsing beter te integreer met sielkundige teorie en eksperimentele metodes.

Toetse word vandag in breër perspektief gesien as blote hulpmiddels by praktiese besluitneming. Selfs wanneer geen formele teorie ter sprake is nie, dwing vrae in verband met konstrukgeldigheid die toets-

opsteller om die trek wat hy meet presies te definieer en om die ver= band tussen toetstellings en ander veranderlikes te spesifiseer en te ondersoek. Dit het ook die feit beklemtoon dat 'n verskeidenheid ge= gewens nodig is om stellings aangaande 'n toets se konstrukgeldigheid te staaf, eerder as om tevrede te wees met net een tegniek.

Die volgende drie stappe in die proses van konstrukvalidering wys op die belangrikheid van teorie in die proses:

(i) Die formulering van hipoteses op grond van die definisie van die konstruk en die verbandhoudende wette

(ii) Die empiriese toetsing van hierdie hipoteses

(iii) Die modifiëring van die teorie (insluitende die konstruk en die wette) op die basis van die empiriese gegewens. Positiewe resultate werp lig op die betekenis van die konstruk; nega= tiewe resultate lei tot 'n hersiening van die teorie en/of in= strument asook die hipoteses.

11.6 MOONTLIKE WANBEGRIPPE AANGAANDE DIE RELATIEWE BELANGRIKHEID VAN TOETS= BETROUBAARHEID EN TOETSGELDIGHEID

Vir sommige skrywers is die konvensionele betroubaarheidsteorie on= toepaslik op baie gebiede van sielkundige meting. Reeds gedurende 1957 het Loevinger so ver gegaan om te beweer dat, indien 'n toets konstrukgeldigheid besit, geen betroubaarheidsindekse nodig is nie. Karon (1966) sê: ... *for personality tests validity coefficients are important and ... reliability is largely an irrelevant conside= ration* (p. 226). Hy is ook oortuig dat hoewel: *It is often assumed that internal consistency must be higher than temporal consistency, (this is) a conclusion which is as false as the assumption that in= ternal consistency limits validity* (p. 224).

Lumsden (1976) betreur die feit dat daar nog geen vooruitgang op die gebied van toetsteorie gedurende die laaste tiental jare gemaak is nie. Hy skryf die toedrag van sake enersyds toe aan die moeilike probleme op die gebied en andersyds aan die beheptheid van toetsteo= ritici met die betroubaarheid van toetse. Hy skryf: *The problems of reliability should be assimilated into validity and scaling theory*

where they have already been partly solved (p. 265). Sy beswaar teen tradisionele toetsteorie is dat dit gegrond is op 'n aanname wat selde geregverdig is naamlik dat 'n toets eendimensioneel is.

Hoewel Loevinger en Lumsden albei te velde trek teen die belangrikheid wat aan toetsbetroubaarheid toegedig word, verskil hulle wat betref die relatiewe belangrikheid van verskillende soorte geldigheid. Vir Loevinger is konstrukgeldigheid die belangrikste eienskap van 'n toets. Lumsden meen weer dat ... *construct validation of tests is ... impossible* (p. 271) vir die redes wat Lumsden en Ross (1973) noem, naamlik dat 'n konstrukvalideringsprogram die volgende vereis: (i) *test unidimensionality and theoretical singularity*, (ii) *operational criteria for all the theoretical terms used to describe tests*, and (iii) *multiple theoretical linkages for the terms* (p. 271). Hoofsaaklik omdat teorieë aangaande individuele verskille nog nie ver genoeg ontwikkel is nie, kan daar selde (indien ooit) aan die derde vereiste voldoen word. Lumsden meen dus dat daar meer aandag aan inhoudsgeldigheid en die opstel van eendimensionele toetse gegee moet word.

Karon (1966) het 'n oplossing vir diegene wat aandring op betroubaarheidskoëffisiënte. Hy beveel aan dat ... *students who work with personality tests may take the highest validity coefficient (the highest correlation of their test with anything), square it, and report this estimate of a lower bound to the reliability, which indeed it is. That this lower bound to the reliability so estimated, is frequently higher than the internal consistency directly measured, is simply evidence of the fact that mental test theory, with its assumption of random error uncorrelated with anything, does not apply to their domain* (p. 227).

Karon meen nie dat so 'n prosedure neerkom op *capitalizing on chance* nie. Dit mag waar wees indien 'n navorser ongeveer 100 geldigheidskoëffisiënte het waaruit hy die hoogste kies, maar dit is nie gewoonlik die geval nie. In die gewone loop van 'n ondersoek word selde meer as 1-3 geldigheidskoëffisiënte bereken en dus is *capitalizing on chance* nie 'n probleem nie.

11.7 SAMEVATTING EN GEVOLGTREKKINGS

In die breedste sin beskou het die geldigheid van 'n toets betrekking op veral twee vrae naamlik wat die toets meet en wat dit kan voorspel. Uit 'n wetenskaplike navorsingsoogpunt is eersgenoemde vraag waarskynlik die belangrikste. Hoewel hierdie aspek nie altyd by voorspelling so noodsaaklik is nie, is dit nogtans so dat die gebruik van teoreties relevante veranderlikes by voorspelling bydra tot die *begrip* en interpretasie van verbande wat gevind word.

Wat die laaste vraag betref, naamlik wat toetstellings kan voorspel, is dit goed om in gedagte te hou dat toetse ook kan verskil wat geldigheidsbreedte betref. Daar word verwys na die vermoë van 'n toets om meer as een betekenisvolle diskriminasie te kan maak. 'n Toets wat net een enkele kriterium en net een kan voorspel, is waarskynlik nie van veel teoretiese belang nie.

Die toetsopsteller kan byna nooit al die ondersoeke wat relevant is aan 'n besondere toets doen nie. Die omvang van die valideringsprogram sal bepaal word deur die omstandighede van die toetsontwikkeling. Aan die een uiterste sal 'n klasonderwyser nie 'n stelselmatige studie maak van 'n toetsie wat hy vir sy studente gee nie. Aan die ander kant kan die uitgewer of opsteller van 'n aanlegtoets wat omvattend gebruik sal word vir voorligting, twee jaar en langer bestee aan navorsing voordat die toets vrygestel word en periodiek daarna nog addisionele ondersoeke doen solank as wat die toets gebruik word. Selfs in die eerste geval is dit nuttig vir die onderwyser om te weet watter vrae aangaande geldigheid kan ontstaan al is dit net om die beperkings van sy toets te erken.

Wat konstrukgeldigheid betref, is validering van 'n toets feitlik nooit afgehandel nie. Die positiewe resultate van elke studie dra by tot meer vertroue in 'n toets as 'n geldige meetinstrument vir 'n besondere konstruk.

Wanneer 'n toets opgestel word, moet van meet af aan, aan ten minste inhoudsgeldigheid en konstrukgeldigheid aandag gegee word. Die verhoging van die betroubaarheid van 'n toets is in 'n hoë mate 'n bloot tegniese saak maar geldigheid behels baie meer as tegniek daar dit

diepgaande filosofiese grondslae het. Wanneer 'n toets eers by die itemontledingstadium is, kan onbevredigende geldigheid nie so maklik soos onbevredigende betroubaarheid reggestel word nie.

Die gewoonte om toetsbetroubaarheid voor toetsgeldigheid te bespreek, is logies verdedigbaar slegs op grond daarvan dat 'n mens wil weet hoe betroubaar iets gemeet word voordat 'n mens vra wat dit is wat gemeet word. Uit die oogpunt van persoonlikheidsielkunde kom dit egter voor asof die gewone volgorde eerder omgeruil behoort te word. Indien 'n toets nie iets wat psigologies belangrik is, meet nie, maak dit nou juis saak hoe betroubaar die toets meet wat dit nou ook al meet?

LITERATUURLYS VIR HOOFSTUK 11

ALTHAUSER, R.P. & HERBERLEIN, T.A. *Validity and the multitrait-multi=method matrix.* In: BORGETTA, E.F. (ed.), Sociological Methodology. San Francisco, Jossey-Bass, 1971.

AMERICAN PSYCHOLOGICAL ASSOCIATION. *Standards for educational and psychological tests and manuals.* Washington, APA, 1966.

AMERICAN PSYCHOLOGICAL ASSOCIATION. *Technical recommendations for psychological techniques.* Washington, APA, 1954.

ANASTASI, A. *Psychological testing.* New York, MacMillan, 1964.

BLALOCK, H.M. *Causal inferences in non-experimental research.* Chapel-Hill, University of Carolina Press, 1964.

BLOOM, B.S. (ed.) *Taxonomy of educational objectives.* Handbook 1. Cognitive domain. New York, McKay, 1956.

BORUCH, R.F., LARKIN, J.D., WOLINS, L. & MACKINNEY, A.C. Alternative methods of analysis: Multitrait-multimethod data. *Educational and Psychological Measurement* 30, 1970: 833 - 853.

BORUCH, R.F. & WOLINS, L.A. A procedure for estimation of trait, method and error variance attributable to a measure. *Educational and Psychological Measurement* 30, 1970: 547 - 574.

BROWN, F.G. *Principles of educational and psychological testing.* Illinois, Dryden Press, 1970.

CAMPBELL, D. & FISKE, D. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56, 1959: 81 - 105.

CATTELL, R.B. Validity and reliability: A proposed more basic set of concepts. *Journal of Educational Psychology* 55, 1964: 1 - 22.

CRANO, W.D., KENNY, D.A. & CAMPBELL, D.T. Does intelligence cause achievement? A cross-lagged panel analysis. *Journal of Educational Psychology* 63, 1972: 258 - 275.

CRONBACH, L.J. *Test validation.* In: THORNDIKE, R.L. Educational measurement. Washington, D.C., American Council on Education, 1971.

CRONBACH, L.J. & MEEHL, P. Construct validity in psychological tests. *Psychological Bulletin* 52, 1955: 281 - 302.

DE GROOT, A.D. *Methodologie.* 's-Gravenhage, Mouton, 1961.

DUFF, F.L. Item subtlety in personality inventory scales. *Journal of Consulting Psychology* 29, 1965: 565 - 570.

EDMUNDS, G. & KENDRICK, D.C. *The measurement of human aggressiveness.* New York, Wiley, 1980.

GOLDBERG, L.R. & SLOVIC, P. Importance of test-item content: an analysis of a corollary of the deviation hypotheses. *Journal of Counselling Psychology* 30, 1967: 199 - 206.

GUION, R.M. *Personnel testing.* New York, McGraw-Hill, 1965.

GUION, R.M. *Recruiting, selection, and job placement.* In: DUNNETT, M.D. (ed.), Handbook of Industrial-Organizational Psychology. New York, Rand McNally, 1976.

GULLIKSEN, N. *Theory of mental tests.* New York, Wiley, 1950.

FREEMAN, F.S. *Theory and practice of psychological testing.* New York, Holt, Rinehart & Winston, 1962.

HELMSTADTER, G. *Principles of psychological measurement.* New York, Appleton-Century-Crofts, 1964.

HELMSTADTER, G. *Research concepts in human behavior.* New York, Appleton-Century-Crofts, 1970.

HUYSAMEN, G.K. *Beginsels van sielkundige meting.* Pretoria, Academica, 1978.

KARON, B.P. RELIABILITY: Paradigm of paradox, with especial reference to personality tests. *Journal of Projective Techniques and Personality Assessment* 30, 1966: 223 - 227.

KERLINGER, F.N. *Foundations of behavioral research.* New York, Holt, Rinehart & Winston, 1964.

KRAUSE, M.S. The implication of convergent and discriminant validity data for instrument validation. *Psychometrika* 37, 1972: 179 - 186.

LOEVINGER, J. Objective tests as instruments of psychological theory. *Psychological Reports* 3, 1957: 635 - 694.

LUMSDEN, J. Test theory. *Annual Review of Psychology* 27, 1976: 254 - 280.

LUMSDEN, J. & ROSS, J. Validity as theoretical equivalence. *Australian Journal of Psychology* 25, 1973: 191 - 197.

NEILL, J.A. & JACKSON, D.N. An evaluation of item selection strategies in personality scale construction. *Educational and Psychological Measurement* 30, 1970: 647 - 661.

SHAVELSON, R.J. & STANTON, G.C. Construct validation: Methodology and application to three measures of cognitive structure. *Journal of Educational Measurement* 12, 1975: 67 - 85.

STILSON, D.W. *Probability and statistics in psychological theory.* San Francisco, Holden-Day, 1966.

TENOPYR, M.L. Content-construct confusion. *Personnel Psychology* 30, 1977: 47 - 54.

THORNDIKE, R.L. *Educational measurement.* Washington, American Council for Education, 1971.

THORNDIKE, R.L. & HAGEN, E. *Measurement and evaluation in Psychology and education.* New York, Wiley, 1977.

WIGGINS, J.S. Personality and prediction. *Principles of personality assessment.* Reading, Addison-Wesley, 1973.

WISEMAN, S. The effect of restriction of range upon correlation coef= ficients. *British Journal of Education Psychology* 37, 1967: 248 - 252.

YEE, A.H. & GAGE, N.L. Techniques for estimating the source and direction of causal inference in panel data. *Psychological Bulletin* 70, 1968: 115 - 126.

HOOFSTUK 12
ITEMRESPONSTEORIE (LATENTE TREKTEORIE)

Vir 'n goeie begrip van hierdie hoofstuk is dit nodig dat die leser 'n gangbare kennis sal besit van hoërskoolwiskunde en statistiese begrippe soos gemiddelde, standaardafwyking, verdeling en normaalverdeling. Dele wat in raampies geplaas is, is bedoel om addisionele inligting te verskaf aan lesers met 'n kennis van universiteitswiskunde. Al sou die leser hierdie dele weglaat, behoort hy steeds 'n redelike begrip van itemresponsteorie uit die res van die hoofstuk te kry.

12.1 DIE KLASSIEKE TOETSTEORIE TEENOOR ITEMRESPONSTEORIE

12.1.1 Die gebondenheid aan groepnorme

In die klassieke toetsteorie is dit slegs moontlik om 'n redelike begrip van 'n persoon se vermoë* te verkry deur die persoon se vermoë in verhouding tot die gemiddelde en standaardafwyking of ander statistieke van die vermoë van een of ander groep persone te beskou. Sodra 'n persoon nie aan die normgroep behoort nie en nog boonop met 'n ander toets vir dieselfde vermoë getoets word, is dit met klassieke toetsteorie onmoontlik om 'n goeie begrip van sy vermoë te vorm. Die volgende situasie klink miskien verspot maar dit is analoog aan die toestand wat bestaan in 'n teorie waar altyd na een of ander groepnorm verwys moet word. Gestel ons kon die lengte van 'n kleuter slegs uitdruk in terme van die gemiddelde en standaardafwyking van die lengte van alle kleuters en dat die lengte van 'n volwassene slegs gegee kan word in terme van die gemiddelde en standaardafwyking van die lengte van alle volwassenes. In so 'n geval sou ons geen verband tussen die lengte van 'n kleuter en die lengte van 'n volwassene kon vind nie ten spyte van die feit dat daar 'n onderliggende veranderlike bestaan waarmee die lengte van alle persone gemeet kan word.

Deur die woord *intelligensie* te vervang vir *lengte* in die voorgaande, word gevind dat die situasie nie verspot is nie maar ooreenstem met die werklike toedrag van sake. Die fisikus sou nooit gelukkig slaap as hy nie 'n verband kon kry tussen die eenheid waarmee die massa van atome gemeet word en die eenheid waarmee die massa van sterre gemeet word nie.

---

* In hierdie hoofstuk word vermoë gesien as een of ander eienskap wat met 'n psigometriese toets gemeet moet word.

Daar word dus gesoek na 'n metode om weg te kom van die verpligting om die meting van 'n persoon se vermoë aan een of ander normgroep te koppel. Wat nodig is, is steekproefvrye skattings van 'n persoon se vermoë en dié vermoë moet kan strek oor die hele spektrum van vermoëns waarin die psigometrikus moontlik kan belangstel.

12.1.2 Die spesifiekheid van meetinstrumente

'n Ernstige probleem wat uit die aard van die klassieke toetsteorie voortvloei, is dat die meting van vermoë afhang van die spesifieke instrument waarmee dit gemeet word. Ons kry byvoorbeeld nie vergelykbare IK's wanneer die NSAG en die SSAIS vir die meting van die IK van 'n persoon gebruik word nie. Die klassieke toetsteorie bied ook geen manier waarmee 'n vergelykbare meting gevind kan word nie. Die situasie kan weer met 'n voorbeeld in die natuurkunde vergelyk word. As 'n fisikus die temperatuur van water gemeet het, is dit gewoonlik nie vir hom nodig om te spesifiseer dat hy 'n kwik- of 'n alkoholtermometer gebruik het nie. Watter een ook al gebruik is, die meting beteken dieselfde. Ons moet dus daarna streef om metings van vermoë te kry wat onafhanklik is van 'n spesifieke toets en die spesifieke items wat in 'n toets gebruik word.

12.1.3 Vlak van meting

'n Ander probleem van die klassieke toetsteorie is dat dit slegs skynbaar metings op 'n intervalskaal gee alhoewel hierdie probleem dikwels geïgnoreer word wanneer sekere wiskundige modelle gebruik word. Die praktiese betekenis van die verskil in verstandsouderdom tussen 4- en 5-jaar is beswaarlik dieselfde as dié tussen 12- en 13-jaar. Net so min is die praktiese betekenis van die verskil van 15 IK-punte tussen 50 en 65 dieselfde as 'n verskil van 15 punte tussen 120 en 135. Om die maklike en goeie interpretasie van metings te bevorder, is dit nodig om metings van vermoë te vind wat op 'n verhoudingskaal gemeet kan word of minstens op 'n intervalskaal sodat kwantifiseerbare gevolgtrekkings gemaak kan word uit die feit dat die vermoë van twee persone met 'n sekere hoeveelheid verskil.

12.1.4 Meting oor 'n wye strek van vermoë

Die rede waarom mense met 'n vermoëtoets getoets word, is om betroubaar

tussen persone met verskille in vermoë te onderskei. Dit is onvermydelik dat 'n vermoëtoets vir sommige persone te maklik sal wees en hulle sal die maksimumtelling of 'n telling daar naby behaal terwyl ander weer lae tellings sal behaal wat slegs op kansvlak of selfs laer lê. As 'n toets vir 'n groep persone te maklik is, kan daar nie effektief tussen hulle onderskei word nie. 'n Nuttige analogie is die situasie waar daar nie tussen 'n groep goeie hoogspringers onderskei kan word nie indien die dwarslat konsekwent te laag gestel word. Net so min kan tussen die swakker springers onderskei word as die dwarslat te hoog gestel word. Om dus effektief tussen persone oor die hele vermoëspektrum te onderskei, is dit nodig om toetse op verskillende vlakke te hê en daar moet 'n metode wees om prestasie op verskillende toetse op dieselfde skaal te plaas. Itemresponsteorie maak dit moontlik.

12.1.5 Die voordele van itemresponsteorie

In teenstelling met die klassieke toetsteorie gee die itemresponsteorie metodes waarmee item- en toetsstatistieke oor verskillende groepe persone en verskillende toetse van dieselfde konstruk vergelyk kan word. In die itemresponsteorie is

- item- en toetsstatistieke invariant (afgesien van 'n lineêre transformasie) oor toetsgroepe en toetse,
- dit moontlik om die eienskappe van 'n toets noukeurig te beskryf voordat die toets toegepas word indien slegs die itemparameters bekend is,
- dit moontlik om vooraf 'n toets te ontwerp wat baie doeltreffend vir die spesifieke doel van die toets sal wees.

12.2 PERSOONLIKHEIDSTREKKE EN DIE DIMENSIONALITEIT VAN ITEMS IN 'N TOETS

'n Teorie van persoonlikheidstrekke neem aan dat 'n persoon se toekomstige gedrag in 'n aansienlike mate voorspel kan word deur gebruik te maak van metings van 'n relatiewe klein aantal menslike eienskappe wat trekke genoem word. Die probleem is egter om die trekke te identifiseer en dan metings van 'n persoon se peil in elke trek deur byvoorbeeld psigometriese toetsing te maak.

Dit moet duidelik gestel word dat daar nie deur voorgaande geïmpliseer word dat trekke in 'n fisiologiese sin bestaan nie (alhoewel dit dalk so

kan wees). Dit is vir praktiese doeleindes voldoende dat 'n persoon op= tree asof hy in besit is van 'n sekere hoeveelheid van elk van die aantal onderliggende trekke en dat dit voorkom asof hierdie hoeveelhede sy ge= drag in 'n aansienlike mate bepaal.

Die doel met psigometriese toetsing is nie om net 'n persoon se telling op 'n spesifieke groep items te kry nie. Die persoon wat toets, wil afleidings maak oor die toetsling se tipiese of verwagte prestasie op 'n groot klas van items soos die wat toegepas is. Hierdie verwagte pres= tasie word die toetsling se vermoë genoem in die psigologiese konstruk wat deur die klas items gedefinieer word.

As items in 'n toets baie heterogeen is ten opsigte van die trekke wat hulle meet, het vermoë soos dit hierbo gedefinieer is, weinig psigolo= giese waarde. 'n Totaaltelling van korrekte antwoorde in 'n toets met nie-verwante items kan psigologies nie sinvol geïnterpreteer word nie. Die rede hiervoor is dat enige bepaalde toetstelling nie 'n eenduidi= ge betekenis het nie. Toetslinge met totaal uiteenlopende eienskappe kan dieselfde toetstelling kry. Begrip van sake word dus die beste bevorder deur 'n toets te gebruik met 'n klas items wat voldoende ho= mogeen is, dit wil sê items wat 'n enkele trek meet, sodat die pres= tasie van 'n persoon in die toets met 'n enkele getal voorgestel kan word.

12.3 ITEMRESPONSFUNKSIES

Van die belangrikste probleme in die psigometrika is om trekke in terme van waarneembare verskynsels te definieer en om te bepaal watter trekke in 'n gegewe gedragskonteks belangrik is.

Die psigometrikus wat 'n persoon se response op die items van 'n sielkun= dige toets wil gebruik om afleidings te maak oor die persoon se vermoë (stand op die latente trek), moet vir elke item weet watter verband daar tussen die onderliggende trek Θ en die itemtelling U is. Hierdie verband word gegee deur die regressievergelyking van die itemtelling U van 'n item op die latente trek Θ. Die regressievergelyking

$$\mu(U|\Theta) = f(\Theta)$$

waar $\mu(U|\Theta)$ die gemiddelde itemtelling U aandui vir persone met vermoë Θ en $f(\Theta)$ aandui dat dié gemiddelde 'n funksie van Θ is, word die *item-responsfunksie (irf)* van die betrokke item genoem. Die uiteensetting wat volg, word beperk tot die geval waar Θ eendimensioneel is. Om hierdie rede sal die irf invariant wees (deur dieselfde wiskundige vergelyking voorgestel word) vir alle groepe toetslinge. As dit nie so is nie moet daar nog onderliggende trekke wees wat tussen die itemtellings van die groepe onderskei en dan kan Θ nie eendimensioneel wees nie. Omdat slegs met 'n eendimensionele trek gewerk word, sal enige parameter wat 'n eendimensionele irf $f(\Theta)$ beskryf, dus wat groepe persone betref, 'n invariante parameter wees.

Die irf spesifiseer dus presies hoe die voorwaardelike gemiddelde waargenome itemtelling van 'n populasie toetslinge vir 'n spesifieke item verband hou met hul vermoë op die latente trek. Die irf is dus 'n punt waar begin kan word om inferensies te maak oor die peil van die latente trek, wat nie direk waarneembaar is nie, vanaf die itemresponse wat wel direk waarneembaar is. Om sulke inferensies te maak, is 'n basiese taak van sielkundige toetsing.

Hoe kan 'n itemresponsfunksie gevind word? Die latente trek is dan nie direk waarneembaar nie. 'n Baie belangrike sprong in die ontwikkeling van die itemresponsteorie is dat sekere aannames gemaak moet word oor die vorm van die itemresponsfunksie. Hierdie aannames, 'n model genoem, saam met die toetslinge se werklike response op die items is voldoende om al die oorblywende relevante informasie oor die itemresponsfunksies en die onbekende Θ's (vermoëns) vir elke persoon te vind.

'n Wiskundige model om vir elke toetsling 'n waarde vir Θ te vind vanaf sy itemresponse, sal vervolgens beskryf word. Die eerste aanname wat gemaak word, is dat die items in die toets die eienskap van lokale onafhanklikheid besit. Hiermee word bedoel dat vir enige groep toetslinge met dieselfde waarde vir die latente trek Θ, sê Θ_1, die voorwaardelike verdeling van itemtellings vir elk van die items onafhanklik van mekaar is. Dit beteken eenvoudig dat as al die persone met 'n waarde Θ_1 uit 'n groep geneem word en 'n proporsie P_g van hulle het item g korrek beantwoord dan sal P_g nie verander as die groep toetslinge of die samestelling van ander items in die toets verander nie. Die beteken weer

dat die waarskynlikheid dat 'n toetsling 'n item korrek beantwoord, nie beïnvloed word deur antwoorde wat hy op ander items in die toets gegee het nie.

As in die algemeen die waarskynlikheid dat 'n gebeurtenis U sal plaasvind, gegee dat gebeurtenis Θ reeds plaasgevind het, aangedui word deur $prob(U|\Theta)$, kan vir n binêre items die beginsel van lokale onafhanklikheid wiskundig soos volg gestel word:

$$Prob(U_1, U_2, U_3, \ldots, U_g, \ldots U_n|\Theta) = \prod_{g=1}^{n} Prob(U_g|\Theta) \qquad (1)$$

waar U_g = itemtelling van item g
en Θ = waarde van vermoë.

Die waarskynlikheid om 'n spesifieke gegewe responspatroon vir al die items saam waar te neem, is dus gelyk aan die produk van die waarskynlikhede om die gegewe respons vir elke item waar te neem.

Ons kan vergelyking (1) ook in vektornotasie skryf:

$$Prob(\underline{U}|\Theta) = \prod_{g=1}^{n} Prob(U_g|\Theta) \qquad (2)$$

Die voorwaardelike waarskynlikheid P_g, gegee Θ, dat 'n binêre item g korrek geantwoord word, kan geskryf word as

$$P_g \equiv P_g(\Theta) \equiv Prob(U_g = 1|\Theta) = \mu(U_g|\Theta)$$

$P_g(\Theta)$ is vir 'n binêre item niks anders nie as die gemiddelde itemtelling vir persone met vermoë Θ, dit wil sê die regressie van die itemtelling op Θ, met ander woorde die itemresponsfunksie van item g.

Definieer nou ook

$$Q_g = Prob(U_g = 0|\Theta) = 1 - P_g$$

Vir 'n binêre item g kan die waarskynlikheid vir enige itemtelling U_g dan soos volg geskryf word:

$$Prob(U_g|\Theta) = (P_g)^{U_g}.(Q_g)^{1-U_g} \qquad (3)$$

want as $U_g = 1$ is $Prob(U_g|\theta) = (P_g).(Q_g)^{1-1} = P_g$

en as $U_g = 0$ is $Prob(U_g|\theta) = (P_g)^0.(Q_g)^{1-0} = Q_g$

Uit (2) en (3) volg dan dat vir enige responspatroon $\underline{U}$

$$Prob(\underline{U}|\theta) = \prod_{g=1}^{n} (P_g)^{U_g}.(Q_g)^{1-U_g} \qquad (4)$$

Vergelyking (4) gee die waarskynlikheid om vir 'n vasgestelde θ 'n sekere responspatroon $\underline{U}$ waar te neem, met ander woorde die kans dat 'n persoon met 'n vermoë θ sekere spesifieke items korrek en die res van die items verkeerd sal beantwoord. Let asseblief op dat die som van die waar=skynlikhede oor alle moontlike responspatrone moet sommeer tot 1. Dus

$$\sum_{\forall \underline{U}} Prob(\underline{U}|\theta) = 1.$$

Wanneer $\underline{U}$ vasgehou word by een of ander responspatroon, word die funk=sie

$$Prob(\underline{U}|\theta) = L(\underline{U}|\theta)$$

'n aanneemlikheidsfunksie (likelihood function) genoem. Die aannames wat gemaak word oor die vorm van die itemresponsfunksies bring mee dat 'n eksplisiete vorm vir $L(\underline{U}|\theta)$ neergeskryf kan word. As dit ge=beur, soos in die model wat bespreek gaan word, dat die funksie $L(\underline{U}|\theta)$ vir elke $\underline{U}$ 'n maksimum het wat by 'n unieke waarde van θ bereik word, word die maksimumaanneemlikheidswaarde (maximum likelihood value) $\hat{\theta} = \hat{\theta}(\underline{U})$ gedefinieer om hierdie waarde te besit. Daar sal later van die maksimumaanneemlikheidswaarde gebruik gemaak word om die parame=ters van die itemresponsfunksies en 'n waarde vir θ vir elke persoon te vind.

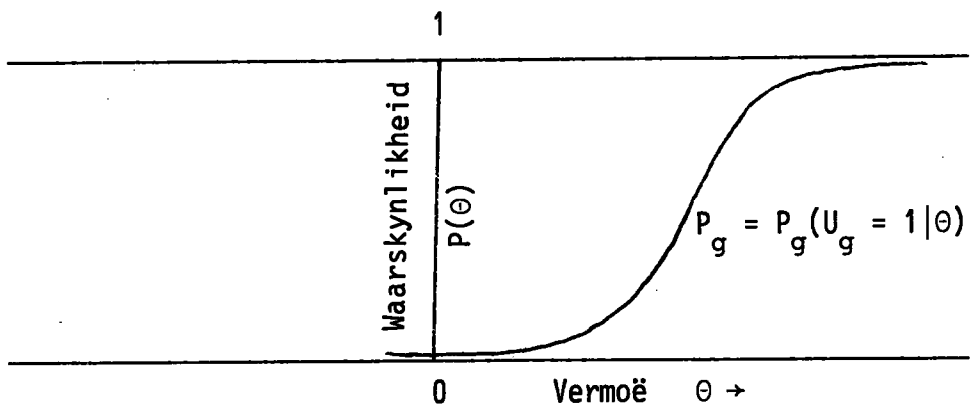
12.4 ENKELE MODELLE VIR ITEMRESPONSFUNKSIES

12.4.1 Die algemene vorm van die irf

Vir elke toetsling kan 'n waarde vir die latente trek θ gevind word en

ook die ontbrekende informasie (itemparameters) vir elke item se item= responsfunksie, indien sekere aannames oor die vorm van die *irf* gemaak word.

Voordat die genoemde aannames gemaak word, kan met bestaande kennis reeds iets oor die vorm van 'n *irf* gesê word. Kyk na figuur 12.1.



FIGUUR 12.1 DIE VORM VAN DIE ITEMRESPONSFUNKSIE

Daar kan aanvaar word dat die waarskynlikheid baie naby 1 is dat per= sone met baie hoë Θ 'n item g korrek sal beantwoord. Net so kan aan= vaar word dat vir persone met lae vermoë die waarskynlikheid 0 is dat hulle die item korrek sal beantwoord. Daar word voorlopig aangeneem dat die kans 0 is dat die korrekte antwoord van 'n item geraai kan word. Daar word aangeneem dat die *irf* 'n kontinue funksie van Θ is en die vorm daarvan moet dus min of meer wees soos aangedui in figuur 12.1.

12.4.2 Die wiskundige formule vir die *irf*

Dit is nie genoeg dat ons net weet wat die vorm van 'n *irf* is nie. 'n Wiskundige vergelyking moet vir die *irf* gepostuleer word voordat die probleem om 'n Θ vir elke persoon te vind, opgelos kan word.

Twee families funksies het die verlangde vorm en kan as model gebruik word:

(a) Die normaalogiefmodel

$$P_g(\Theta) = \Phi(a_g\,(\Theta - b_g)) \equiv \int_{-\infty}^{a_g(\Theta - \beta_g)} \phi(x)\,dx \qquad (5)$$

waar $\phi(x) = (2\pi)^{-\frac{1}{2}}\ exp(-x^2/2)$
en a_g en b_g itemparameters is wat bepaal word deur die diskrimina= sie- en moeilikheidseienskappe van die item.

(b) Die tweeparameter logistiese model

$$P_g(\Theta) = \Psi(Da_g(\Theta - b_g)) \equiv \int_{-\infty}^{Da_g(\Theta - b_g)} \psi(x)\,dx \qquad (6)$$

waar $\psi(x) \equiv e^{-x}/(1 + e^{-x})^2$

D = 'n konstante = 1,7
en a_g en b_g itemparameters is wat bepaal word deur die diskriminasie- en moeilikheidseienskappe van die item.

In hierdie geval kan 'n algebraïese vergelyking vir Ψ gevind word naamlik†

$$\Psi(Da_g(\Theta - b_g)) = \{1 + exp(-Da_g(\Theta - b_g))\}^{-1}$$

Dit is moontlik om die konstante D so te kies dat die normaalogiefmodel en die logistiese model feitlik dieselfde resultate lewer. Dit kan be= wys word dat

$$|\ \Phi(x) - \Psi(1{,}7x)\ | < 0{,}01 \text{ vir alle } x.$$

Om hierdie rede word D gewoonlik gelyk aan 1,7 gekies.

† Per definisie is $exp(x) = e^x$ waar e die grondtal van die natuurlike logaritmes is ($e \cong 2{,}7183$).

Die bekende Raschmodel vir itemresponsfunksies kan beskou word as 'n spesiale geval van die logistiese model. As vermoë in die Raschmodel voorgestel word deur Θ^* en die volgende substitusies in die logistiese model gemaak word, naamlik

$$a_g = 1 \text{ vir alle } g,$$

$$\Theta^* = 1{,}7\Theta$$

en $$b_g^* = 1{,}7\, b_g$$

word die itemresponsfunksie vir die Raschmodel gevind, naamlik

$$P_g(\Theta^*) = \Psi(\Theta^* - b_g^*) = 1 + exp(-\Theta^* + b_g^*))^{-1}$$

$$= exp(\Theta^* - b_g^*)/(1 + exp(\Theta^* - b_g^*))$$

12.4.3 Die betekenis van die itemparameters

Wat is die betekenis van a_g en b_g in die logistiese model? In figuur 12.2 word die itemresponsfunksies vir drie items van die tweeparameter logistiese model

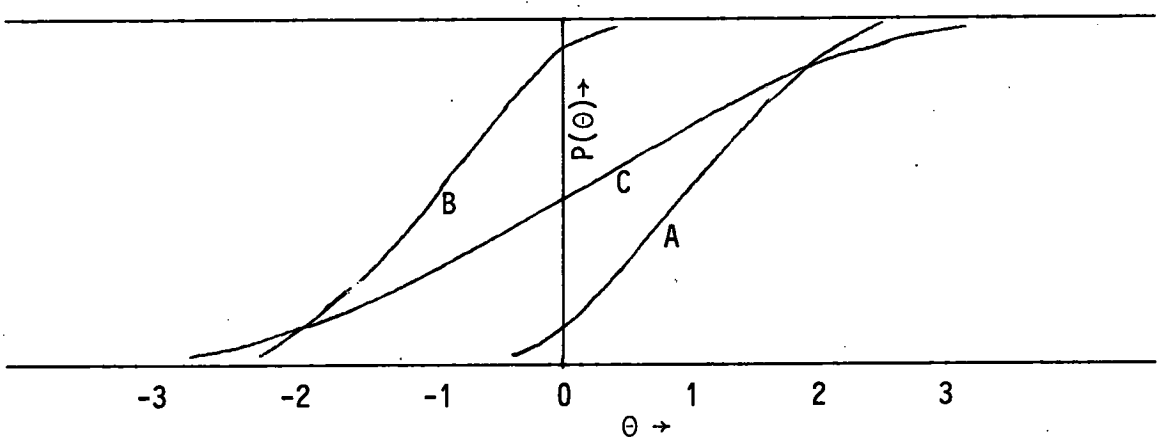
$$P_g(\Theta) = (1 + exp(-1{,}7\, b_g(\Theta - b_g)))^{-1}$$

getoon waar

$a_1 = 1,\ b_1 = +1$ ---------- A

$a_2 = 1,\ b_2 = -1$ ---------- B

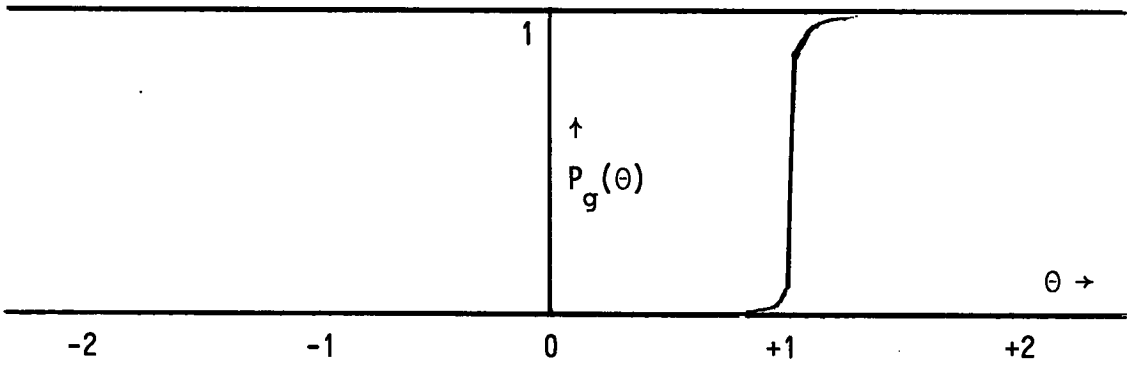
en $a_3 = 0{,}5,\ b_3 = 0$ ---------- C



FIGUUR 12.2: DRIE ITEMRESPONSFUNKSIES

Ons merk dat die parameter b_g die *irf* van links na regs skuif as dit (b_g) toeneem. As b_g toeneem terwyl a_g konstant bly, neem die waarskynlikheid dus af dat die persone met 'n bepaalde Θ die item korrek sal beantwoord. Om hierdie rede word b_g die *moeilikheidsindeks* van die item genoem. Al= hoewel daar wel 'n verband is, kan daar nie 'n formule vir die verband tussen b_g en die moeilikheidswaarde van 'n item in die klassieke toets= teorie gevind word nie. Die rede hiervoor is dat die moeilikheidswaarde van die klassieke toetsteorie afhang van die eienskappe van die groep toetslinge terwyl b_g en a_g invariante parameters is. In beginsel kan b_g varieer tussen $-\infty$ en $+\infty$ maar as gevolg van skaling van Θ (kyk para= graaf 12.5.1) is die variasie gewoonlik (maar nie noodwendig nie) tus= sen -3 en +3.

Die parameter a_g kan potensieel varieer tussen 0 en ∞. In figuur 12.3 word 'n *irf* getoon met $b_g = 1$ en $a_g = \infty$ (baie groot).



FIGUUR 12.3: 'N ITEMRESPONSFUNKSIE MET $a_g = \infty$

Die parameter a bepaal die helling van die *irf*. Die helling van die *irf* by 'n punt Θ bepaal die diskriminasievermoë by Θ met ander woorde hoe goed die item kan onderskei tussen persone met vermoë Θ en persone met vermoë $\Theta + \Delta$. Hoe groter die helling is hoe groter sal $P_g(\Theta + \Delta) - P_g(\Theta)$ vir 'n gegewe Δ wees. Hoe groter $P_g(\Theta + \Delta) - P_g(\Theta)$ vir 'n gegewe Δ is, hoe makliker is dit om tussen persone met vermoë $\Theta + \Delta$ en persone met vermoë Θ te onderskei want persone met vermoë $\Theta + \Delta$ het dan 'n groter kans om die item korrek te kry as persone met vermoë Θ. Die parameter a_g word die *diskriminasie-indeks* van die item g genoem. Let daarop dat die klassieke diskriminasiewaarde van 'n item bepaal word deur die samestelling van die groep toetslinge (met ander woorde of die groep gemiddelde, lae of hoë vermoë het, terwyl a_g inva= riant is.

Let op die volgende:

As $\Theta = b_g$ is

$P_g(\Theta) = 0$

$P'_g(\Theta) = Da_g/4$

en $P''_g(\Theta) = 0$

waar $P'_g(\Theta)$ = eerste afgeleide van $P_g(\Theta)$ met betrekking tot Θ,

en $P''_g(\Theta)$ = tweede afgeleide van $P_g(\Theta)$ met betrekking tot Θ.

By $\Theta = b_g$ het die twee parameter itemresponsfunksie dus 'n punt van inf leksie, met ander woorde 'n punt waar die helling 'n maksimum bereik. Die diskriminasievermoë van 'n item is dus maksimaal by $\Theta = b_g$ waar dit die waarde van $Da_g/4$ aanneem.

As die diskriminasie-indeks a_g baie groot word, sal die diskriminasievermoë van die item in die onmiddellike omgewing van $\Theta = b_g$ groot word maar vinnig val na nul sodra weg beweeg word van $\Theta = b_g$. Hier het ons met 'n skynbare paradoks te doen naamlik hoe hoër die diskriminasie-indeks a_g van 'n item is hoe beter diskrimineer dit in die onmiddellike omgewing van b_g en hoe swakker diskrimineer dit as van b_g weg beweeg word. Aan die ander kant is dit weer waar dat items met relatief kleiner waardes van a_g weer nie so goed by $\Theta = b_g$ diskrimineer nie maar tog oor 'n wyer gebied bevredigend diskrimineer. Hierdie probleem sal weer later aangeraak word. Vir die huidige kan gemeld word dat 'n toets wat bestaan uit items wat almal dieselfde moeilikheidsindeks het (let wel *nie* die moeilikheidswaarde van die klassieke toetsteorie nie) en dieselfde nie te groot diskriminasie-indeks ($a_g \cong 1$) het, oor 'n redelike wye gebied van Θ goed sal diskrimineer.

1.4 Verband van die itemparameters met die klassieke toetsmodel

In die *spesiale* geval waar die normaalogiefmodel geld en Θ normaal verdeel met gemiddelde 0 en standaardafwyking 1, kan vir elke item g 'n matematiese verband tussen a_g en b_g aan die een kant en die klassieke diskriminasiewaarde ρ_g en die klassieke moeilikheidswaarde π_g aan die ander kant gevind word. Hierdie verband is nie eenvoudig nie.

Die verband is soos volg

$$\pi_g = \frac{1}{\sqrt{2\pi}} \int_{\gamma_g}^{\infty} e^{-x^2/2}\, dx$$

waar $\gamma_g = a_g\, b_g\, (1 + a_g^2)^{-\frac{1}{2}}$

$a_g = \rho_g\, (1 - \rho_g^2)^{-\frac{1}{2}}$

π_g = proporsie persone wat item g korrek beantwoord het

en ρ_g = biseriaalkorrelasie tussen itemtelling en toetstotaaltelling

12.4.5 Die drieparameter logistiese model

Die twee modelle wat in vergelykings (5) en (6) gegee is, sal slegs bevredigend werk as daar geen moontlikheid is dat 'n persoon die korrekte antwoord van 'n item kan raai nie. Wanneer daar 'n waarskynlikheid c_g is dat die korrekte antwoord van item g verskaf kan word deur te raai, sal die onderste asimptoot van die *irf* nie by $P_g = 0$ wees nie maar by $P_g = c_g$. Daar kan in 'n drieparametermodel voorsiening gemaak word vir raai van korrekte antwoorde deur vergelykings (5) en (6) respektiewelik soos volg aan te pas:

$$P_g(\Theta) = c_g + (1 - c_g)\, \Phi\, (a_g(\Theta - b_g)) \qquad (5a)$$

$$P_g(\Theta) = c_g + (1 - c_g)\, \Psi\, (Da_g(\Theta - b_g)) \qquad (6a)$$

12.4.6 Die waarde van c

Die c-waarde is 'n raaiparameter en dui die waarskynlikheid aan dat persone met baie lae vermoë die item korrek sal antwoord. Oor die algemeen word gevind dat c nie presies gelyk is aan a^{-1} waar a die aantal afleiers per item is nie. 'n Verklaring wat gewoonlik hiervoor aangebied word is dat toetslinge nie regtig ewekansig raai wanneer hulle nie die antwoord op 'n item ken nie. Oor die algemeen is c vir items met vyf afleiers

egter ongeveer gelyk aan 0,20.

12.5 DIE SKATTING VAN ITEMPARAMETERS EN WAARDES VIR Θ VIR ELKE PERSOON

12.5.1 Die maksimum aanneemlikheidsbeginsel

Omdat die logistiese model die model is wat die beste in die praktyk gebruik kan word, sal oorsigtelik aangetoon word hoe itemparameters en waardes vir Θ daarvoor uit die toetsgegewens gevind word.

Die drieparameter logistiese model stel die waarskynlikheid dat 'n persoon d met vermoë Θ_d 'n item i met itemparameters a_i, b_i en c_i korrek sal antwoord as

$$P_{id} = Prob(U_{id} = 1 | a_i, b_i, c_i, \Theta_d)$$

$$= c_i + (1 - c_i)\Psi(\ell_{id}) \qquad (8)$$

waar $\ell_{id} = Da(\Theta_d - b_i)$

$\Psi(\ell_{id}) = (1 + exp(-\ell_{id}))^{-1}$

en $D = 1,7$

As gevolg van die lokale onafhanklikheidsbeginsel en die onafhanklikheid van die beantwoording van verskillende persone van enige item, kan met behulp van vergelyking (4) die waarskynlikheid vir 'n bepaalde uitkoms van die toetsing van N persone met n items soos volg geskryf word:

$$L = \prod_{d=1}^{N} \prod_{i=1}^{n} P_{id}^{U_{id}} Q_{id}^{1-U_{id}} \qquad (10)$$

Deur nou vergelyking (8) in (10) te substitueer, word 'n vergelyking van die volgende vorm verkry:

$$L = L(\underline{U}, \underline{\Theta}, \underline{a}, \underline{b}, \underline{c}) \qquad (11)$$

Die maksimum aanneemlikheidsbeginsel aanvaar dat die werklike uitkoms van 'n eksperiment dié uitkoms is wat volgens die model die aanneemlikste is. Die onbekende parameters $\underline{\Theta}$, $\underline{a}$, $\underline{b}$ en $\underline{c}$ moet dus so gekies word dat L 'n maksimum sal wees onderhewig aan die waargenome responspatrone van al die toetslinge.

Deur die parsiële afgeleides van L met betrekking tot $\underline{\theta}$, $\underline{a}$, $\underline{b}$ en $\underline{c}$ gelyk aan 0 te stel, word N + 3n gelyktydige vergelykings met N + 3n onbekendes naamlik $\underline{\theta}$, $\underline{a}$, $\underline{b}$ en $\underline{c}$ verkry.

'n Probleem is egter dat slegs N + 3n - 2 van die gelyktydige vergelykings onafhanklik van mekaar is. Hierdie feit kan wiskundig aangetoon word, wat nie hier gedoen sal word nie, maar kan ook aanneemlik gemaak word deur daarop te let dat as $a_i\ (\theta_d - b_i)$ gelyk aan 'n konstante is, is die eenheid en oorsprong van die θ-skaal nie uniek bepaal nie.

Daar is dus N + 3n onbekendes en slegs N + 3n - 2 vergelykings. Deur nou verder twee arbitrêre vergelykings te stel wat die oorsprong en eenheid van θ uniek sal bepaal, word N + 3n vergelykings verkry wat opgelos kan word. Die twee arbitrêre vergelykings is gewoonlik die volgende:

$$\sum_{d=1}^{N} \theta_d/N = 0 \qquad (12)$$

$$\sum_{d=1}^{N} \theta_d^2/N = 1 \qquad (13)$$

Dit is nie moontlik om die N + 3n komplekse vergelykings met die hand op te los nie. Dit is selfs vir 'n redelike groot en kragtige rekenaar nie 'n onaansienlike taak nie. Die oplossing geskied deur 'n iterasieproses wat soms lank neem om te konvergeer. Vanweë die aard van die maksimum-aanneemlikheidsbeginsel is dit wenslik vir konvergensie dat die aantal persone redelik groot (N > 1 000) moet wees en die aantal items (n > 40) ook.

12.5.2 Die keuse van die θ-skaal

Vergelykings (12) en (13) kan die indruk skep dat die meting van θ, net soos in die klassieke toetsteorie, nog steeds noodwendig in terme van die gemiddelde en standaardafwyking van θ is. Dit is egter nie die geval nie. Enige ander twee lineêre onafhanklike vergelykings soos byvoorbeeld

$$\theta_1 - \theta_2 = 1 \qquad (12a)$$

$$\theta_1 = 0 \qquad (13a)$$

sou ook doen. Daar is weer 'n pragtige analogie met die meting van tem= peratuur waar eers twee punte op die skaal naamlik 0° en 100° arbitrêr vasgestel moes word voordat die temperatuurskaal nuttig gebruik kan word.

Indien daar twee verskillende toetse is wat dieselfde vermoë meet en wat op verskillende groepe toegepas is, sal die metode van oplossing wat in die vorige paragraaf geskets is nie Θ's, a's en b's vir items van albei toetse gee wat op dieselfde skaal lê nie. Dit is egter moontlik om 'n lineêre transformasie van die een skaal na die ander te maak, net soos 'n lineêre transformasie van die Celsius- na die Fahrenheit-temperatuur= skaal gemaak kan word.

Wanneer dieselfde groepie items (ankeritems) in twee verskillende toetse op twee verskillende groepe persone toegepas is, kan die a- en b-waardes van die items en die Θ's soos verkry in die een situasie lineêr getrans= formeer word na die waardes soos dit in die ander situasie gevind sou word. Die C-waardes van die itemresponsfunksies is reeds op dieselfde skaal. Die transformasieformules is die volgende:

$$b^* = S^*/S\ (\Theta - \bar{b}) + \bar{b}^*$$
$$\Theta^* = S^*/S\ (\Theta - \bar{b}) + \bar{b}^*$$
en $$a^* = a.S/S^*$$

waar b = waarde van b-parameter in groep 1-skaal,
a = waarde van a-parameter in groep 1-skaal,
Θ = waarde van vermoë in groep 1-skaal,
$\bar{b}$ = gemiddelde van die b-parameters vir die gemeenskaplike items vir die groep 1-skaal,
S = standaardafwyking van die b-parameters vir die groep 1-skaal,
en $*$ dui ooreenstemmende waardes vir groep 2 aan.

Om te verseker dat die standaardafwyking van die b-waardes akkuraat be= paal word, word aanbeveel dat daar meer as 17 ankeritems moet wees.

12.5.3 Meer oor die vlak van meting

Op hierdie stadium kan gerus weer iets gesê word oor ratio- en inter= valskale. Deur na vergelykings (5) en (6) te kyk, word gemerk dat item= moeilikheidsindekse en Θ op dieselfde skaal is. Verder kan die prak=

tiese afleiding gemaak word dat 'n persoon met vermoë Θ_1 'n kans van 50 % het om alle items met $b_g = \Theta_1$ korrek te beantwoord. In die algemeen kan noukeurige algemene stellings nie gemaak word oor hoe 'n persoon met vermoë Θ_1 op items met $b_g \neq \Theta_1$ sal vaar nie. Die tekortkominge van die klassieke teorie wat in paragraaf 12.1.3 genoem is, geld dus nog tot 'n sekere mate vir die logistiese model.

In die geval van die Raschmodel is die posisie egter beter en kan 'n transformasie van die Θ-skaal na 'n Θ'-skaal gevind word waar laasgenoemde die eienskappe van 'n ratioskaal toon. Op die Θ'-skaal kan dan noukeurige algemeen geldende afleidings gemaak word oor die implikasies van 'n verskil in vermoë tussen twee persone met vermoëns Θ_1' en Θ_2' respektiewelik.

Die Raschmodel gee vir item g met itemparameter b_g

$$P_g(\Theta) = exp(\Theta - b_g)/(1 + exp(\Theta - b_g)$$
$$Q_g(\Theta) = 1 - P_g(\Theta) = (1 + exp(\Theta - b_g))^{-1}$$

Dus is vir item g die relatiewe kans O_g van sukses teenoor die kans op mislukking (odds for getting the item right)

$$O_g = P_g/Q_g = exp(\Theta)/exp(b_g)$$

Transformeer nou $\Theta = \log_e \Theta'$

en $b = \log_e b_g'$

Dan is $$O_g = P_g/Q_g = \Theta'/b_g'$$

Aangesien Θ kan varieer tussen $-\infty$ en $+\infty$ sal Θ' varieer tussen 0 en ∞. Ons het dus 'n skaal met 'n natuurlike of absolute nulpunt wat zero vermoë sal verteenwoordig.

Verder sal die waarde van O_g vir alle g verdubbel as Θ' verdubbel. Met ander woorde 'n persoon met vermoë $2\Theta_1'$ se relatiewe kans op sukses in enige item is twee keer so groot as die relatiewe kans op sukses van 'n persoon met vermoë Θ_1'.

Die Θ'-skaal toon dus die eienskappe wat ons van 'n ratio-skaal verwag.

Dit het 'n natuurlike nulpunt en gelyke intervalverskille op Θ' het dieselfde betekenis op enige vlak van Θ'.

12.6 TOETSKARAKTERISTIEKE FUNKSIES

12.6.1 Die verband tussen toetstelling en Θ

Die toetstelling X van 'n persoon is die getal items wat hy korrek beantwoord het in 'n toets met n items. Dus

$$X = \sum_{g=1}^{n} U_g$$

In vergelyking (4) het ons gesien dat die kans vir 'n spesifieke responsiepatroon gegee word deur:

$$Prob(\underline{U}|\Theta) = \prod_{g=1}^{n} P_g^{U_g} Q_g^{1-U_g} \qquad (14)$$

Die spesifieke responspatroon $\underline{U}$ is egter maar een van die moontlike maniere om dieselfde totaaltelling X te verkry omdat hierdie telling ook verkry kon word deur enige X van die n items korrek te kry. Die kans van 'n persoon met vermoë Θ om 'n telling X te kry, kan dan soos volg geskryf word:

$$Prob(X|\Theta) = \sum_{\Sigma U_g = X} \left(\prod_{q=1}^{n} P_g^{U_g} Q_g^{1-U_g} \right). \qquad (15)$$

In (15) gaan die sommasie oor al die moontlike responspatrone wat 'n totaaltelling X gee. Die uitdrukking aan die regterkant is die digtheidsfunksie vir die saamgestelde binomiaalverdeling (compound binomial distribution) met die bekende gemiddelde

$$\sum_{g=1}^{n} P_g.$$

Uit die voorgaande kan die volgende belangrike formule vir die gemiddelde $\mu(X|\Theta)$ van X vir 'n gegewe Θ gevind word:

$$\mu(X|\theta) = \sum_{g=1}^{n} P_g(\theta) = n\,\bar{P}(\theta) \qquad (16)$$

waar $\bar{P}(\theta)$ die gemiddelde van die itemresponsfunksies by θ is.

Die voorwaardelike variansie $\sigma^2(X|\theta)$ van X, gegee θ, kan ook maklik gevind word. As gevolg van die beginsel van lokale onafhanklikheid is die variansie eenvoudig die som van die variansies by θ. Dus

$$\sigma^2(X|\theta) = \sum_{g=1}^{n} P_g\,Q_g \qquad (17)$$

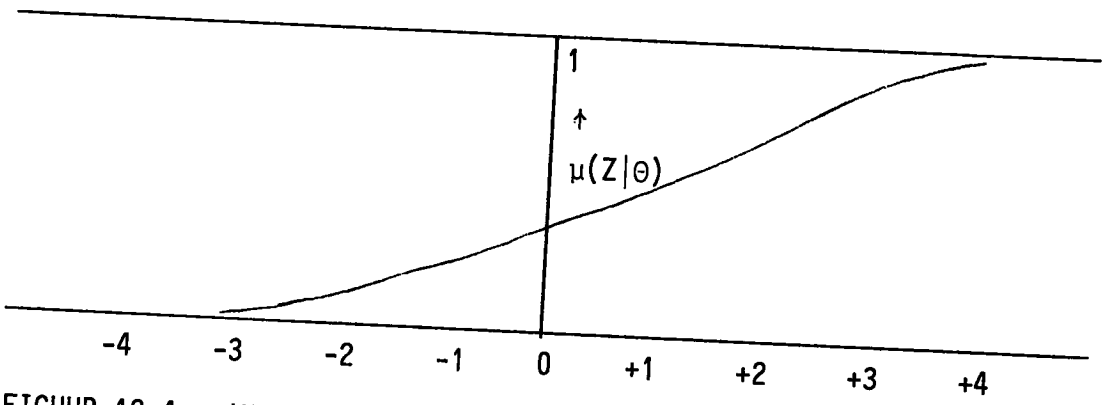
waar $Q_g = 1-p_g$.

Vir latere gebruik is dit nuttig om vergelykings (16) en (17) oor te skryf in terme van die proporsie-korrek-telling $Z = X/n$:

$$\mu(Z|\theta) = \frac{1}{n}\sum_{g=1}^{n} P_g = \bar{P} \qquad (18)$$

en $$\sigma^2(Z|\theta) = \frac{1}{n^2}\sum_{g=1}^{n} P_g\,Q_g \qquad (19)$$

Vergelyking (18) is van fundamentele belang in itemresponsteorie en dit staan as die *toetskarakteristieke funksie (tkf)* bekend. Die *tkf* gee die regressievergelyking van die proporsie-korrek-telling Z op θ. Figuur 12.4 toon 'n tipiese *tkf*.



FIGUUR 12.4: 'N TIPIESE TOETSKARAKTERISTIEKE FUNKSIE (*tkf*)

12.6.2 Ware telling

Die ware telling (true score) T van 'n persoon met vermoë θ in 'n toets met n items word gedefinieer as die verwagte proporsie items wat hy korrek sal beantwoord op 'n toets wat bestaan uit 'n oneindige aantal

parallelle vorms wat elk n items het.

Die vorms is parallel in dié sin dat die g-de item in elke vorm dieselfde *irf* het ($g = 1, 2, \ldots, n$).

Laat U_{gj} die itemtelling wees van die g-de item van die j-de parallelle vorm. Dan is die ware telling met behulp van vergelyking (18)

$$T = \lim_{m \to \infty} \frac{1}{nm} \sum_{g=1}^{n} \sum_{j=1}^{m} U_{gj}$$

$$= \frac{1}{n} \sum_{g=1}^{n} P_g$$

$$= \mu(Z|\Theta).$$

Die variansie van T kan met behulp van vergelyking (19) soos volg bereken word:

$$\sigma^2(T|\Theta) = \lim_{m \to \infty} \left(\frac{1}{nm}\right)^2 \sum_{g=1}^{n} \sum_{i=1}^{m} Q_{ij} P_{ij}$$

$$= 0$$

Die gesamentlike verdeling van Θ en T degenereer dus en daar is geen strooiing om die regressie van T op Θ nie.

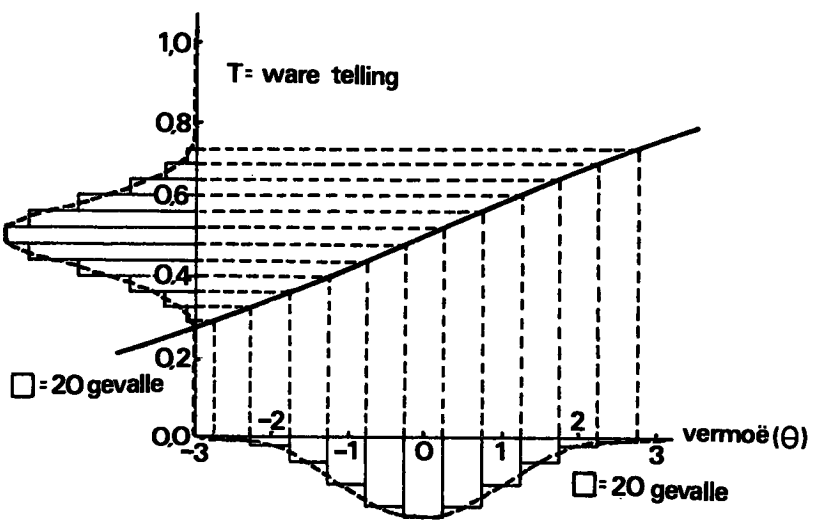
Uit die voorgaande blyk dus dat die ware telling 'n funksionele eerder as 'n statistiese verband met Θ het. Dit beteken dat in 'n eendimensionele latente trekruimte Θ dieselfde is as die ware telling T behalwe vir die skaal van meting waarmee elkeen beskryf word. T en Θ is dus funksioneel verbind met 'n monotoon stygende funksie wat gelyk is aan die toetskarakteristieke funksie, naamlik

$$T = \frac{1}{n} \sum_{g=1}^{n} P_g(\Theta) \qquad (21)$$

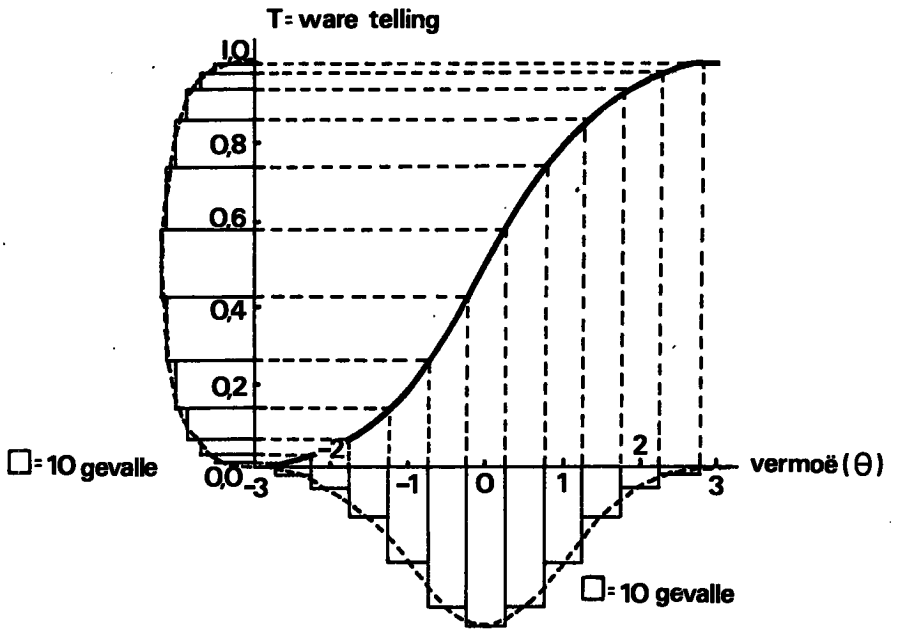
12.6.3 Die verdeling van Θ en ware tellings

Tot dusver is geen aannames gemaak oor die verdeling van Θ in enige populasie toetslinge nie. Deur aannames te maak oor die verdeling van Θ in 'n populasie kan die vorm van die *tkf* gebruik word om die verdeling van die ware telling T te infereer.

Figuur 12.5 toon 'n gedeelte van 'n *tkf* wat oor 'n redelike wye gebied feitlik 'n reguit lyn is met relatiewe lae helling. Hierdie tipe *tkf* word verkry deur items wat relatief swak diskrimineer of deur items met 'n wye reeks moeilikheidswaardes in die betrokke toets op te neem. As die *tkf* min of meer 'n reguit lyn is in die vermoëgebied van die toetsgroep, sal T min of meer dieselfde vorm van verdeling as Θ besit.



FIGUUR 12.5: AFLEIDING VAN DIE VERDELING VAN T (WARE TELLING) VIR 'N TOETS WAT BESTAAN UIT ITEMS WAT SWAK DISKRIMINEER EN GEMIDDELDE MOEILIKHEIDSINDEKSE BESIT



FIGUUR 12.6: AFLEIDING VAN DIE VERDELING VAN T (WARE TELLING) VIR 'N TOETS MET HOë DISKRIMINASIE EN GEMIDDELDE MOEILIKHEIDSWAARDE

In figuur 12.6 word die *irf* van 'n toets wat bestaan uit items met hoë diskriminasie-indekse en gemiddelde moeilikheidswaardes getoon. Dit blyk dat persone naby die kante van die verdeling van Θ saamgedruk word in die verdeling van T. As Θ dus soos in figuur 12.6 normaal verdeel, sal T platikurties verdeel. Dit is te wagte in die lig van die onbeperkte strek van Θ en die beperkte strek van T.

12.7 INFORMASIEFUNKSIES

12.7.1 Definisie van 'n informasiefunksie

In die vorige paragraaf is die regressievergelyking van die totaaltel= ling van die toets X op vermoë Θ ondersoek (vergelykings (15) en (16)). Die totaaltelling is slegs 'n spesiale geval van 'n meer algemene funksie $X = X(\underline{U})$ waar X nou enige funksie van die responspatroon $\underline{U}$ kan wees.

Daar is 'n statistiese verband tussen Θ en X en om hierdie rede kan 'n spesifieke waarde vir Θ nie gegee word as 'n waarde van X bekend is nie. Met ander woorde as 'n waarde vir X bekend is, kan slegs met onvolmaakte sekerheid (sê 95 % sekerheid) beweer word dat Θ in 'n sekere interval (sê tussen 3,16 en 3,20) sal lê.

Die vraag wat nou beantwoord moet word, is die volgende: Hoe akkuraat kan Θ bepaal word as 'n waarde van X, wat 'n statistiese verband met Θ het, gegee word. Statistici het 'n algemene indeks ontwikkel om hierdie akkuraatheid aan te dui. Die waarde van die indeks kan varieer tussen 0, wanneer geen afleiding oor die waarde van Θ vir 'n gegewe X gemaak kan word nie en ∞, wanneer 'n presiese waarde van Θ gegee kan word vir 'n gegewe waarde van X. Dié indeks staan bekend as die informasiefunksie en dit word aangedui deur die simbook $I(\Theta, X)$.

Die informasiefunksie word soos volg gedefinieer:

$$I(\Theta, X) = \left(\frac{\partial \mu(X)}{\partial \Theta}\right)^2 \left(\sigma^2(X|\Theta)\right)^{-1}$$

As aangedui word dat X 'n funksie van U is (wat 'n statistiese verband met Θ het), word die vergelyking vir I:

$$I(\Theta, X(\underline{U})) = \left\{\frac{\partial}{\partial \Theta} \mu(X(\underline{U})|\Theta)\right\}^2 \left\{\sigma^2(X(\underline{U})|\Theta\right\}^{-1} \qquad (22)$$

Die $\mu(X(\underline{U})|\Theta)$ dui die verwagte waarde en $\sigma^2(X|\Theta)$ die variansie van $X(\underline{U})$ aan vir 'n gegewe Θ.

Let daarop dat X nie 'n veranderlike argument van I is nie maar slegs aandui van watter funksie van Θ die informasie-eienskappe ondersoek word.

12.7.2 Iteminformasiefunksies en toetsinformasiefunksies

As $X(\underline{U})$ die vorm van 'n geweegde lineêre som van die itemtellings het, naamlik

$$X(\underline{U}) = \sum_{g=1}^{n} W_g U_g \qquad (23)$$

kan aangetoon word dat die gemiddelde en standaardafwyking van X gegee word deur (kyk vergelyking (16)):

$$\mu(X|\Theta) = \sum_{g=1}^{n} W_g P_g(\Theta) \qquad (24)$$

en $$\sigma^2(X|\Theta) = \sum_{g=1}^{n} W_g^2 P_g(\Theta) Q_g(\Theta) \qquad (25)$$

waar $Q_g(\Theta) = 1 - P_g(\Theta)$.

Die informasiefunksie $I(\Theta, X)$ van X met betrekking tot Θ word gevind deur (24) en (25) in (22) te substitueer:

$$I(\Theta, X) = \left\{\sum_{g=1}^{n} W_g P_g'\right\}^2 \left\{\sum_{g=1}^{n} W_g^2 P_g Q_g\right\}^{-1} \qquad (26)$$

waar $P_g' = \frac{\partial}{\partial\Theta} P_g(\Theta)$ = die parsiële afgeleide van P_g met betrekking tot Θ

Vergelyking (24) gee die gemiddelde, vergelyking (25) die standaardafwyking en vergelyking (26) die informasiefunksie van 'n lineêre kombinasie van itemtellings vir 'n gegewe Θ in terme van die itemresponsfunksies van die items.

As 'n spesiale geval van 'n lineêre tellingsformule kan die geval geneem word waar $W_h = 0$ vir alle h ongelyk aan g. Dan is

$$X = W_g U_g$$

Met ander woorde X is 'n funksie van 'n enkele itemtelling. Die *iteminformasiefunksie (irf)* is dus

$$I(\Theta, U_g) \equiv I(\Theta, W_g U_g) = (P_g')^2/(P_g Q_g) \qquad (27)$$

In vergelyking (27) word die informasie in 'n enkele item in terme van die itemresponsfunksie van die item gegee.

Daar kan wiskundig aangetoon word dat dit algemeen geld dat as

$X = \Sigma W_g U_g$ volg dat

$$I(\Theta, X) \leq \sum_{g=1}^{n} I(\Theta, U_g) \qquad (28)$$

en dat die gelykheid slegs geld as

$$W_g = P_g'/(P_g\,Q_g) \qquad g = 1, 2, 3, \ldots, n \tag{29}$$

Uit vergelyking (28) kan dus gesien word dat die informasie oor Θ in 'n *lineêre kombinasie* van itemtellings minder of gelyk is aan die som van die informasies wat in die individuele itemtellings teenwoordig is.

Omdat $P_g'(\Theta)/(P_g(\Theta).Q_g(\Theta))$ oor die algemeen 'n funksie van Θ is, geld die gelykheid in (28) dus slegs as die W's lokale beste gewigte is. Beste gewigte wat onafhanklik is van Θ, word slegs in logistiese modelle aangetref (paragraaf 12.8.2).

Die regterkant van vergelyking (28) wat eenvoudig die som van die itemsinformasies is, staan bekend as die *toetsinformasiefunksie* en dit word deur 'n spesiale simbool aangedui:

$$I(\Theta) \equiv \sum_{g=1}^{n} I(\Theta, U_g) = \sum_{g=1}^{n} (P_g')^2/(P_g\,Q_g) \tag{30}$$

Let daarop dat $I(\Theta)$ slegs deur die toetsmodel bepaal word aangesien dit die som van die informasiefunksies van die items is. As gevolg van vergelyking (28) is die toetsinformasie $I(\Theta)$ 'n boonste perk vir alle moontlike informasiefunksies $I(\Theta, X)$ wat verkry kan word vir verskillende keuses vir lineêre tellingsformules. Dit kan aangetoon word dat $I(\Theta)$ die maksimum waarde van die informasiefunksie is vir enige moontlike funksie $X(\underline{U})$ met ander woorde ook vir nie-lineêre funksies X van $\underline{U}$.

Uit vergelyking (26) kan die informasiefunksie van die gewone aantal-items-korrektelling X ($W_g = 1$ vir alle g) neergeskryf word, naamlik

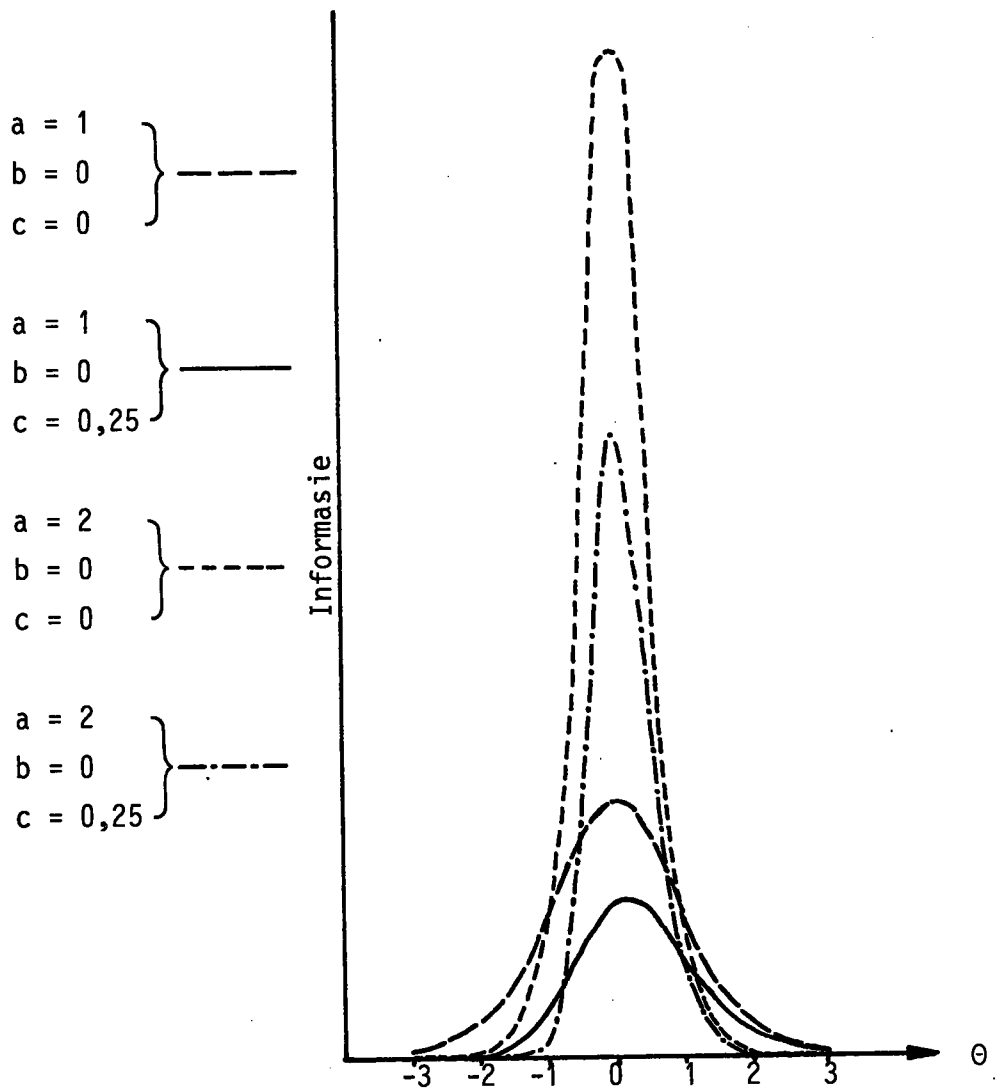
$$I(\Theta, X) = (\Sigma P_i')^2/(\Sigma P_i Q_i)$$

As X die aantal-items-korrektelling is, word $I(\Theta, X)$ die *tellinginformasiefunksie* genoem. Volgens vergelyking (28) is die tellinginformasiefunksie kleiner of gelyk aan die toetsinformasiefunksie. As die logistiese model geld, kan aangetoon word dat vir hoë Θ die tellinginformasiefunksie byna gelyk aan die toetsinformasiefunksie is. As Θ kleiner word, word die verhouding tussen die tellinginformasiefunksie en die toetsinformasiefunksie kleiner. Laasgenoemde feit dui op die ondoel-

treffendheid van die gebruik van die aantal-items-korrektelling om lae vermoëns te skat.

12.7.3 Die invloed van itemparameters op iteminformasie

Omdat die iteminformasiefunksies beskou kan word as die boustene van die toetsinformasiefunksie, is dit nuttig om die eienskappe van die itemin= formasiefunksies vir die logistiese model van nader te beskou. Figuur 12.7 toon grafiese voorstellings van vier iteminformasiefunksies.



FIGUUR 12.7: VOORBEELDE VAN ITEMINFORMASIEFUNKSIES

Vir die drieparameter logistiese model wat in vergelyking (6a) gedefi= nieer is, word die iteminformasiefunksie gegee deur:

$$I(\Theta,U) = \frac{(1.7a)^2(1 - c)}{\{c + exp(1.7a(\Theta - b))\}\{1 + exp(-1.7a(\Theta - b))\}^2}$$

Hierdie iteminformasiefunksie bereik 'n maksimum by Θ_m waar

$$\Theta_m = b + \frac{1}{1,7a}\left\{log_e(0,5 + 0,5\ (1 + 8c)^{-\frac{1}{2}}\right\} \qquad (33)$$

Dit is duidelik dat $\Theta_m \geq b$ en dat Θ_m slegs gelyk sal wees aan b wanneer $c = 0$.

Die iteminformasiefunksie is simmetries as $c = 0$ en na regs skeef as c ongelyk is aan nul (kyk figuur 12.7). Die skeefheid na regs beteken dat die c-waarde meer informasie by lae waardes van Θ vernietig as by hoë waardes. Hierdie verskynsel klop met wat 'n mens verwag aangesien verwag kan word dat persone met laer vermoë meer raai as persone met hoër vermoë.

Die totale informasie A in 'n item word gedefinieer as die som van die informasie wat dit by alle moontlike waardes van Θ het.

Dus

$$A = \int_{-\infty}^{+\infty} I(\Theta, U)\ d\Theta$$

Vir die drieparameter logistiese model is

$$A = 1,7a\ (1 + (c\ log_e\ c)/(1 - c)) \qquad (34)$$

Vir 'n vaste waarde van c sal die totale informasie in 'n item dus styg as a styg. Vir 'n vaste waarde van a sal die informasie A 'n maksimum wees as $c = 0$ en 'n minimum as $c = 1$. As c dus styg vernietig dit informasie. Dit is dus voordelig om meer afleiers vir 'n item te kies omdat c daal as die aantal afleiers toeneem.

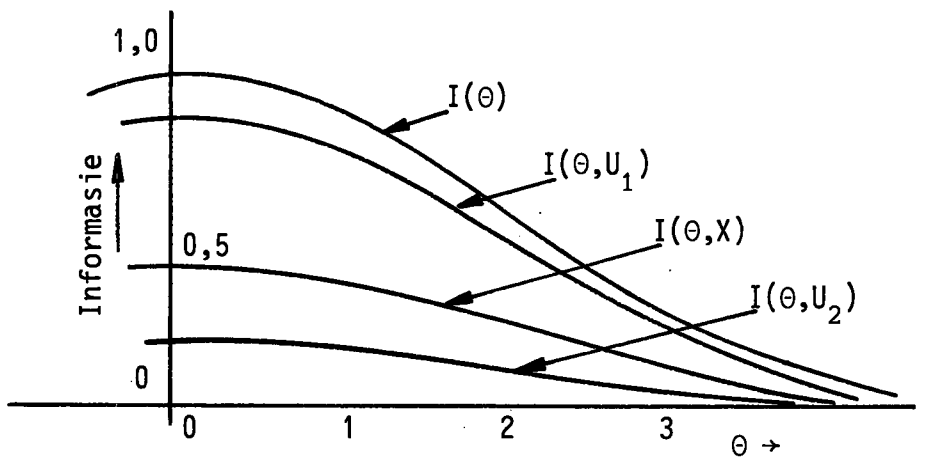
Alhoewel die totale informasie volgens vergelyking (34) styg as a groter word, word die informasie oor 'n kleiner gebied van Θ versprei as a groter word (kyk figuur 12.7). Hierdie verskynsel word die *bandwyteparadoks* genoem. 'n Kompromis moet dus soms gemaak word tussen die totale informasie wat 'n item verskaf en die gebied van Θ waaroor die informasie versprei. In die praktyk word gewoonlik items met die hoogste a-waardes gekies aangesien a-waardes groter as 1 selde gevind word.

12.7.4 Die opbou van toetsinformasie uit iteminformasie

Omdat die toetsinformasiefunksie eenvoudig die som van die iteminforma= siefunksies is van die items waaruit die toets bestaan, kan die vorm van die toetsinformasiefunksie tydens toetsopstelling bepaal word deur items in die toets te plaas wat die gewenste informasie-eienskappe besit.

Ter illustrasie van die effek van weging van itemtellings op die infor= masie in die totaaltelling word 'n toets geneem wat slegs uit twee soorte items bestaan. Veronderstel die helfte van die items het $a_g = a_1 = 0{,}95$ en die ander helfte het $a_g = a_2 = 0{,}20$ terwyl $b_g = 0$ vir alle items.

Die funksie $I(\Theta)$ gee die toetsinformasiefunksie dit wil sê waar die item= tellings volgens vergelyking (29) geweeg is, $I(\Theta, X)$ gee die tellingin= formasiefunksie van die ongeweegde totaaltelling van itemtellings, $I(\Theta, U_1)$ gee die tellinginformasiefunksie waar al die tipe 1-items 'n ge= wig van 1 het en die tipe 2-items 'n gewig van 0 het en $I(\Theta, U_2)$ die tel= linginformasiefunksie waar al die tipe 1-items 'n gewig van 0 het en die tipe 2-items 'n gewig van 1. Genoemde informasiefunksies word in figuur 12.8 voorgestel.



FIGUUR 12.8: INFORMASIEFUNKSIES VIR VERSKILLENDE LINEêRE SAMESTELLINGS VAN TWEE TIPES ITEMS

Die waardes van a_1 en a_2 wat hier gekies is, verteenwoordig die uiterstes wat in die praktyk in 'n gestandaardiseerde toets gevind word. Daar word opgemerk dat in hierdie uiterste voorbeeld die items met $a_g = 0{,}20$ in werk=

likheid die informasie laat daal as die tellings daarvan ongeweeg by die tellings van die items met a_g = 0,95 getel word. Hieruit kan gesien word dat die informasie in die totaaltelling van 'n toets oor die hele spektrum van Θ verminder kan word as items wat swak diskrimineer gevoeg word by items wat goed diskrimineer.

12.7.5 Metingsfout en die verband met klassieke toetsteorie

Omdat 'n toets nie volkome akkuraat meet nie is dit nodig om die moontlike metingsfout in gedagte te hou wanneer 'n persoon se vermoë Θ met 'n toets geskat word. Die *standaardmetingsfout* S_e vir 'n gegewe Θ word gedefinieer as die standaardafwyking van die Θ's wat met die meetinstrument geskat word vir alle persone met werklike vermoë Θ.

Die standaardmetingsfout het 'n eenvoudige verband met die toetsinforma= siefunksie, naamlik

$$S_e = (I(\Theta))^{-\frac{1}{2}}$$

Die standaardmetingsfout is dus 'n funksie van Θ.

Wanneer 'n onsydige skatting van Θ uit itemtellings gemaak word, kan met behulp van die standaardmetingsfout van Θ, soos bereken by die geskatte waarde $\hat{\Theta}$, grense bepaal word waarbinne die ware telling waarskynlik sal lê. Die ware telling sal met redelike sekerheid lê in die interval $\hat{\Theta} - S_e(\hat{\Theta})$ tot $\hat{\Theta} + S_e(\hat{\Theta})$.

Die gemiddelde waarde $\bar{S}_e$ oor toetslinge van die standaardmetingsfout het 'n baie interessante verband met die betroubaarheidskoëffisiënt ρ_{xx} van die klassieke toetsteorie, naamlik

$$\rho_{xx} = 1 - (\bar{S}_e/S_t)^2$$

waar S_t^2 = variansie van waargenome Θ,s oor toetslinge.

Die laasgenoemde vergelyking beklemtoon weer die gebreke van die klassieke toetsteorie en in besonder van die betroubaarheidskoëffisiënt. 'n Toets met 'n hoë betroubaarheidskoëffisiënt kan vir 'n bepaalde doel ongeskik wees

omdat dit lae informasie bevat by sekere belangrike waardes van Θ. Soortgelyk kan 'n toets met 'n lae betroubaarheidskoëffisiënt vir sekere doeleindes baie geskik wees omdat dit hoë informasie het waar dit vir 'n spesifieke doel nodig is.

Die vergelyking wys ook op die afhanklikheid van die betroubaarheidskoëffisiënt van die verspreiding van vermoë in die toetsgroep. As baie toetslinge op die vlak van die vermoëskaal lê waar die toets hoë informasie bevat, sal die betroubaarheidskoëffisiënt hoër wees as wanneer die toetslinge versprei word in gebiede van Θ waar die toetsinformasie laag is.

12.8 RELATIEWE DOELTREFFENDHEID

12.8.1 Definisie

Laat $I_1(\Theta, X)$ en $I_2(\Theta, Y)$ die tellinginformasiefunksie wees van twee toetse van dieselfde vermoë met die volgende tellingformules:

$$X = X(\underline{U}_1) \text{ en } Y = Y(\underline{U}_2).$$

Die verhouding

$$RD(\Theta, X, Y) \equiv I_1(\Theta, X)/I_2(\Theta, Y) \tag{40}$$

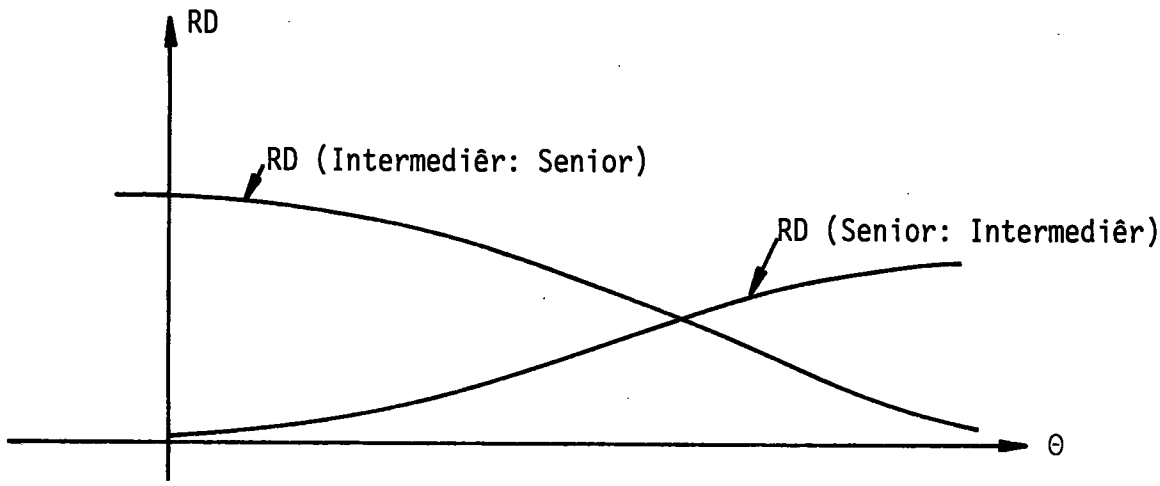
word die *relatiewe doeltreffendheid* (RD) by Θ van die twee tellingsformules X en Y genoem.

In die spesiale geval waar die twee toetse dieselfde is en $Y = Y(U)$ sodanig is dat $I(\Theta, Y) = I(\Theta)$ vir alle Θ word die verhouding

$$Eff(\Theta, X) \equiv I(\Theta, X)/I(\Theta) \tag{41}$$

die doeltreffendheid van X by Θ genoem.

Vergelyking (40) gee 'n metode om die doeltreffendheid van twee verskillende toetse, wat albei Θ meet, met mekaar te vergelyk. Figuur 12.9 gee 'n voorbeeld van wat verwag kan word as die relatiewe doeltreffendheid van die NSAG (Intermediêr) relatief tot die NSAG (Senior) teenoor Θ grafies aangetoon sou word.



FIGUUR 12.9: RD VAN DIE NSAG (INTERMEDIêR) EN NSAG (SENIOR)

12.8.2 Doeltreffendheid van die totaaltelling vir die logistiese model

Dit is insiggewend om die doeltreffendheid van 'n ongeweegde som van itemtellings X te ondersoek as die tweeparameter logistiese model van toepassing is. Uit vergelykings (41), (30), (26) en (6) kan gevind word dat

$$Eff(\Theta, X) = (\Sigma a_g \psi_g)^2 \ (\Sigma \psi_g)^{-1} \ (\Sigma a_g^2 \psi_g)^{-1} \qquad (44)$$

Uit vergelyking (44) kan gesien word dat die doeltreffendheid van 'n ongeweegde totaaltelling X gelyk is aan 1 as die Raschmodel geld aangesien vir hierdie model $a_g = 1$ vir alle g. As die Raschmodel geld, is daar dus nie 'n funksie van die itemtellings wat 'n akkurater skatting van Θ kan gee as die ongeweegde totaaltelling nie. Hierdie pragtige eienskap van die Raschmodel ontneem ongelukkig nie die feit dat die model nie goed werk vir die meeste praktiese situasies nie.

In vergelyking (29) is aangetoon dat in die algemeen $I(\Theta) = I(\Theta, X)$ slegs wanneer die itemgewigte W_g in 'n geweegde som van itemtellings gelyk is aan

$$W_g = P_g'/(P_g Q_g).$$

As die tweeparameter logistiese model van toepassing is, kan aangetoon word dat

$$W_g = Da_g.$$

Dit volg dus dat as die tweeparameter logistiese model van toepassing is dan is die doeltreffendheid van 'n geweegde totaaltelling X van item-tellings U_g maksimaal (gelyk aan 1) vir alle θ as

$$X = D \sum_{g=1}^{n} a_g U_g$$

dit wil sê as

$$X = \sum_{g=1}^{n} a_g U_g \tag{45}$$

aangesien $I(\theta, DX) \equiv I(\theta, X)$.

Die verband tussen die gemiddelde van X en θ volg dan uit vergelyking (24):

$$\mu(X|\theta) = \sum_{g=1}^{n} a_g P_g(\theta) \tag{46}$$

As die tweeparameter logistiese model geld, is dit relatief maklik om die akkuraatste skatting van 'n persoon se vermoë θ te maak as sy item-response en die itemparameters van al die items in die toets bekend is. Vergelyking (46) kan grafies voorgestel word. 'n Geweegde totaaltelling X kan vir 'n persoon volgens vergelyking (45) bereken word en die ooreenstemmende waarde van θ kan in die grafiek afgelees word.

Let daarop dat as die toetskarakteristieke funksie (gemiddelde ongeweegde totaaltelling Y teenoor θ) bekend is, kan 'n waarde vir θ ook afgelees word as Y bekend is. Hierdie metode sal egter in die algemeen 'n minder akkurate skatting van θ gee as vergelyking (46). Slegs as die Rasch-model geld, kan die akkuraatste skatting vir θ met 'n ongeweegde totaaltelling gemaak word.

LITERATUURLYS VIR HOOFSTUK 12

1 BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. In: LORD, F.M. & NOVICK, M.R. *Statistical theories of mental test scores*. Reading, Mass., Addison-Wesley, 1968, Hoofstukke 17-20.

2 BOCK, R.D. & LIEBERMAN, M. Fitting a response model for N dichotomously scored items. *Psychometrika* 35, 1970, 179-197.

3 HAMBLETON, R.K. & COOK, L. Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement* 14, 1977, 75-96.

4 LORD, F.M. & NOVICK, M.R. *Statistical theories of mental test scores*. Reading, Mass., Addison-Wesley, 1968.

5 RASCH, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denemarke se Paedogogiske Instituut, 1960.

6 SCHMIDT, F.L. The Urry method of approximating the item parameters of latent trait theory. *Ecucational and Psychological Measurement* 37, 1977, 3, 613-620.

7 URRY, V.W. Tailored Testing: a successful application of latent trait theory. *Journal of Educational Measurement* 14, 1977, 181-196.

8 WOOD, R.L., WINGERSKY, M.S. & LORD, F.M. *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6*. Princeton, Educational Testing Service, 1976.

9 WRIGHT, B.D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement* 14, 1977, 97-116.

RGN-PUBLIKASIELYS

'n Volledige lys van RGN-publikasies of 'n lys van publikasies van 'n besondere instituut van die RGN kan van die President van die Raad verkry word.