AANTEKENINGEN BIJ DISSERTATIE MELLENBERGH

1.1 M. onderscheidt 4 typen vragen: essay/short answer/completion/ multiple choice.
Dit is nogal een inperking van mogelijkheden: diverse vormen van mondelinge toetsen, en toetsen van practische vaardigheden vallen er buiten. Verder is de typering van M. willekeurig, in die zin dat hij kijkt naar de vorm van de vragen. M. kiest de vorm als kritische dimensie, en gaat dan kijken of diverse variabelen verschillen op deze vormdimensie: denkprocessen / rxx / rxT / operationalisatie / effect op studiegewoonten en studiesucces.
Ook deze opsomming is niet volledig, de latenttrait ontbreekt, de doelstellingenanalyse ontbreekt, het economisch rendement ontbreekt. de vraag naar de invloed op beoordelingsgewoonten en selectiepraktijken ontbreekt, om maar enkele punten te noemen.

M. bespreekt geen artikelen van propagandistische of beschouwende aard. Dat lijkt me interessant om nu juist wèl te doen. Bovendien is een en ander noodzakelijk om het hele technologische verhaal van M. in een kader te plaatsen van waaruit het aan te vallen of misschien wel te verdedigen is. Technologie is niet alleenzaligmakend.

1.2 Een denkpsychologische vergelijking.

M. vermeldt eers een onderzoek van De Jong, dat als triviaal resultaat had dat verschillende soorten ipateriaal blijkbaar tot verschillende oplossingsstrategieën leiden bij verschillende typen van vragen. Uit de samenvatting die M. van dit onderzoek geeft, blijkt dat het geen enkel significant resultaat opleverde (mijn conclusie).

Het tweede en laatste onderzoek dat M. geeft is van Veltman, met even nietszeggende resultaten als dat van De Jong. Met nietszeggend bedoel ik dat geen enkele kennis toegevoegd wordt aan hetgeen w apriori al kunnen veronderstellen over de aard van de werkwijze.

De duidelijke verwijzing naar de literatuur over problem solving die voor de hand ligt, wordt door M. niet vermeld, laat staan uitgewerkt. M. voldtaat met het zeer magere verslagje van deze twee zeer magere onderzoekjes. Er worden geen conclusies getrokken. De zin van deze paragraaf blijft volledig in het duister. Voor zover ik het nu al kan bekijken komt M. er ook nergens meer op terug.

Probeer in andere literatuur nog verwijzingen te vinden naar relevant onderzoek. Dit is toch wel een belangrijk onderwerp. Bloom zegt er in de le taxonomie wel het een en ander over dacht ik. Maak onderscheid tussen de oplossingswerkwijze (de De Jong en Veltman onderzochten) en denkprocessen wel iets heel anders kan zijn.

1.3 Vergelijking van de betrouwbaarheid.

Bij tests die door beoordelaar gescoord worden treden de twee
extra bronnen van foutenvariantie op: (Cureton 1958).
1) 'Ithat associated with sampling in the population of
possible judges";
2) "that associated with sampling in the universe of evaluations of the same papers bi the same judge on different occasions".

Welnu, voordat we verder gaan zouden we eigenlijk moeten weten
wat nu precies het belang van die beruchte betrouwbaarheid is.
Dat betekent een stukje beschouwelijk denkwerk, precies hetgeen
M. weggelaten heeft bij zijn literatuurkeuze.
1) Lord & Novick blz.72: De validiteit van een test met betrek
kina tot welk criterium dan ook, kan nooit hoger worden dan
de betrouwbaarheidsindex van de test: rho xy < √ rho xx'

W Wanneer we een toetsscore dus gaan gebruiken als voorspeller van bijv. toekomstig studiesucces, dan is een goede betrouwbaarheid niet geheel onbelangrijk. Ik zeg, niet geheel, omdat een hoge validiteit niet altijd noodzakelijk is om een toets als een goede voorspeller te mogen gebruiken (Taylor Russell bijvoorbeeld, of zie tabellen van Van Naerssen in: Bedrijfspsychologie).
2) Bij het voorgaande is wel verondersteld dat de beslissingen die genomen worden op grond van de scores op de toets in het belang zijn van de onderwijsinstelling. Het belang van de individuele student is iets heel anders.

We zullen dan ook de vraag of de onderwijsinstelling er is, voor de student, of wel de student voor de onderwijsinstelling, moeten beantwoorden. M. heeft al een aantwoord gegeven zij het ook impliciet: hij heeft gekozen voor het belang van de onderwijsinstelling boven dat van de student. Hij beschouwt de onderwijsinstelling als een fabriek met een reeks achtereenvolgende productie afdelingen, waarvoor de mensen geselecteerd kunnen worden op grond van de resultaten die ze in de voorgaande productieafdeling(en) geboekt worden (want deze vormen een goede

indicatie voor het te verwachten rendement in de volgende afdeling; en hoe hoger de betrouwbaarheid. hoe beter die voorspelling kan zijn).

3. Een summiere uitwerking van de andere mogelijkheid. de onderwijsinstelling ten dienste van de student, zou er zo uit kunnen zien: zeer frequent vindt er toetsing van de vorderingen plaats, waarbij geen cijfer aan de "prestaties" toegekend worden, maar waar deze prestaties van commentaar voorzien worden.

Dit commentaar is een analyse van fouten en tekortkomingen in termen van activiteiten die de student moet ondernemen om deze fouten bij te spijkeren (dit geldt voor werkelijk belangrijke punten uit het gegeven onderwijs. Op fouten van ondergeschiktk belang is een eenvoudige constatering van het feit van het foutzijn, met daarbij het goede antwoord, voldoende). Het "zakken" met overdoen van de hele stof komt niet meer voor.
Zakslaag beslissingen worden niet meer genomen. Meetbetrouwbaarheid die van belang is om een enigszins redelijk gefundeerde zakslaag beslissingsstrategie te kunnen hanteren is dan ook overbodig. Het enige dat van deze toetsnieuwestijl verwacht wordt is dat de opgaven representatief zijn voor het gegeven onderwijs, en dat het bepalen van kritische vragen berust op leerstofanalyse ('a la Gagné bijv.) Wanneer iemand dan een stuk stof over moet doen omdat hij een slechte prestatie leverde terwijl zijn inzicht toch wel voldoende was, dan kost hem dit overdoen vrijwel geen extra tijd. De hertoets vindt zo snel mogelijk plaats, en alleen op de punten die de betreffende student gevraagd is nog eens te bestuderen.

De toetsing zou zo frequent plaats moeten vinden dat de taken die op basis van de toetsresultaten gegeven worden in maximaal ee å` twee dagen uitgevoerd kunnen worden.
De studietoets verliest op deze wijze zijn testkarakter, en wordt een integraal onderdeel van de eigenlijke onderwijssituatie. Allerlei problemen van kat testtheoretische aard maken plaats voor problemen van leerstof analyse en van constructie van vragen z6danig dat studie en denkprocessen in overeenstemming met de doelstellingen gebaseerde leerstof verlopen.

In Graulund zijn nog enkele interessante punten die in dit kader passen. te vinden. Vooral het artikel van Cronbach.

De vraag bij een dergelijke opzet is ten eerste of deze werkwijze haalbaar is in termen van beschikbare docentenuren? ten tweede in verband met studiemotivatie. Dit laatste punt ligt, dacht ik, erg eenvoudig: in de geschetste opzet wordt de student maximaal geactiveerdp,door snelle feedback, die een positieve motivatie oplevert. De negatieve motivatie die de selectiedreiging oplevert is vervallen, maar we mogen aannemen dat bij een organisa4i torische juiste opzet de nieuwe werkwijze méér zal motiveren, en dat zonder een aantal vervelende invloeden die de selectie dreiging op onderwijssituatie en persoonlijk welbevinden van de student heeft.

4. Er is over die betrouwbaarheid nog wel meer te zeggen, maar misschien is het beter even af te wachten of die punten verdere studie van het rapport van M. niet vanzelf boven water komen. (bijv. fouten die verband houden met het trekken van een steekproef items uit de populatie van mogelijke items, e.d.).

1.3. M. vermeldt een onderzoek van Sims (1932). Bij discussievragen (zijn dit zgn. essayvragen?) treden grote verschillen in het beoordelen op.

Vraag: wat is groot in dit verband? M. geeft geen aanwijzing hier over, hij generaliseert wel op grond van één onderzoek dat de betrouwbaarheid van het beoordelen bij ongeveer 1/3 van de vragen van proefwerken en examens van "belang" is.

1.3.1. Betrouwbaarheid van beoordelingen: intersubjectiviteit

M. haalt resultaten van onderzoek van Edgeworth aan. Deze resultaten zijn, zoals M. ze beschrijft, volkomen betekenisloos: bijvo het slechts aangeven van een "range" voor de cijfers door div. beoordelaars gegeven, geeft vrijwel geen enkele relevanté informatie over de verdeling van de cijfers. De resultaten van Starch & Elliot zijn al een stuk beter: cijfers vertonen een normaalverdeling. Welke betekenis moeten we hier aan geven? Ik weet het niet, immers, wat verwacht je anders dan een normaalverdeling wanneer je het oordeel van 142 mensen over een bepaald opstel vraagt?
Wanneer M. stelt dat sterk significante verschillen tussen beoordelaars werden gevonden in nieuwere onderzoekingen, wat betekent significant dan? (Waar verwij_st de significantie naar?) M.a.w. hoe bepaal je "significantie" in dit verband?

1.3.1 vv. Bij geciteerde onderzoek van Childers geldt weer hetzelfde bezwaar als bij Edgeworth: alleen een range geeft nauwelijks informatie. Ook bij onderzoek van Valin vraag je je af: nou en ?

Na deze losse observaties die als "onderzoek" gepresenteerd werden, komen we toe aan correlatietoestanden (blz.9) eorrelaties worden bere1cend tussen verschillende docenten die dezelfde vragen nakeken. Wat bedoelt M. met correlatie co~ffici;~nt? De betrouwbaarheids co;~ffici;~nt, de betrouwbaarheidsindeX, of een andere maat?

Correlatiecoeff. zijn vaak "vrij laag". Wat is vrij laag? Waar meten we dat aan af? De tabel is nogal indrukwekkend op het eerste gezicht. Wat ontbreekt is de evaluatie van deze resultaten:

welke betekenis moeten we er aan hechten. Wat betekent een "gemiddelde correlatiecoefficient"? Is het "erg" wanneer deze maar .48 is?
Hoe zijn de docenten gekozen? Is er sprake van "sampling in the population of possible judges" zoals Cureton dat noemt (zie 1.3) of is de keuze niet random geweest? Zijn de populaties eigenlijk ooit wel omschrevent Mellenbergh blijkt toch wel verdomd weinig informatie door te geven. Misschien is deze informatie helemaal niet beschikbaar, maar laat M. in dat geval dan de door hem zo lukraak geciteerde "onderzoeken" liever op deze punten kritiseren dan ze zo zinledig in zijn rapport te vermelden..'

1.3.2 Betrouwbaarheid van beoordelingen: stabiliteit.

Ging het in de vorige paragraaf heimelijk om "sampling in the population of possible judges", dan komen we hier toe aan "sampling in the universe of evaluations of the same papers by the same judge on different occasions". Alleen. M. zegt dit niet expliciet, wat wel zo handig zou zijn geweest.
M. geeft ook hier een tabel van onderzoekresultaten.
M. zegt dat ook hier de corr. coeff. lager zijn dan Ten zou wensen.
Ik zou graag van M. willen horen wat "wenselijk" is.

Vergelijk de geciteerde onderzoek gegevens eens met die in Vijven en Zessen.

1.3.3 Bronnen van onbetrouwbaarheid bij het beoordelen.
De in 1.3.1 3n 1.3.2 besproken intersubjectiviteitscorrelaties en stabiliteitscorrelaties verschillen nogal eens van elkaar.
M. citeert Vernon & Millican die vier factoren noemen die deze verschillen kunnen "verklaren". Hier moet dan toch wel even gezegd worden dat "verklaren" wel iets anders is dan "beschrijven".
Het enige dat V.& M. doen is "beschrijven", en dan vinden ze vervolgens dat hun "beschrijving" "van invloed is" (?) op de waarden van de gevonden correlaties. Dit is wetenschappelijke rimram van de onderste plank.
De Groot (Methodologie blz.243) kan er ook wat van: hij somt vijf effecten op. En vervolgens vindt Mellenbergh in de literatuur
"experimentele bevestiging" van deze opsomming. Nu volgt een reeks van volkomen duistere samenvatting en van onderzoekingen van Cast (1939, 1940); French (1961), Britton (1963); Bonardel (1946) en Remondino (1959). Ook hier geen enkele evaluatie, conclusie of discussie van Mellenbergh n.a.v. deze onderzoekingen. Ze vallen wél allemaal onder het door De Groot genoemde signifisch effect, en dat is dat de opvattingen over de te beoordelen taak verschillend kunnen zijn. Tante Betje. Het tweede effect is het halo effect, "de uitstraling van bi, ci """" enz. op de beoordeling van ai. Ook hier weer een aantal onderzoekingen waar we het label "haloeffect" op kunnen plakken. Over de praktische betekenis van dit soort effect ook weer geen enkele aanwijzing. (Dat een bepaald effect aanwezig is, betekent niet vanzelf dat het ook "storend" is.'

1.3~3 Het derde De Groot effect: het sequentieeffect: de nawerking van' voorafgaande beoordelingen op het oordeel over de kwaliteit van de beantwoording.van de opgave. M. geeft ~én onderzoek weer waar gevonden werd dat de spreiding van de toegekende cijfers groter was voor de laatst dan voor de eerstbeoordeelden.

Het vierde effect. "De vrijheid in de beoordelingsschclcxl kan leiden tot oordeelsverdelingen, die algemeen menselijke of persoonlijke neigingen van de beoordelaar uitdrukken, bijvoorbeeld aanpassing aan de groep en persoonlijke vergelijking. Hieronder valt ook de wet van Posthumus. M. vermeldt onderzoekresultaten die tamelijk indrukwekkend zijn in hun demonstratie van aanpassing van de docent aan het niveau van de klas (Rot en Bujas). Jammer genoeg wijdt M. ook op dit punt geen woord aan een beschouwing over de betekenis van dit effect voor de onderwijssituatie, hoewel die betekenis immens groot is. (van de andere effecten is dat niet zo duidelijk. Effecten die differentieren over sociale klassen worden niet genoemd door De Groot of Mellenbergh, een reden te meer om ze bij een eigen publicatie wél boven water te halen). Tenslotte het contaminatieeffect: de vrijheid in de beoordelingsprocedure wordt gebruikt voor andere doeleinden dan die van de onbevooroordeelde beoordeling. Het is niet geheel duidelijk uit de tekst van M. wat nu het specifieke van dit effect is. Sociale discriminatie zou waarschijnlijk onder dit effect geplaatst moeten worden? (of onder het vierde?) (of onder het tweede?). Dan is nu de zaak opgesomd, en gaat M. over tot het volgende topic (dat hier wel mee in verband staat).

1.3.5 Afgezien van deze haarkloverijen, kleeft M. geen bevredigend antwoord op de vraag die hij min of meer ook zelf stelt: is het wel zinvol om van tests die toch behoorlijk uiteen kunnen lopen qua vorm, denkprocessen en dergelijke grappen, te gaan vergelijken op hun betrouwbaarheid? En dan nog, wanneer het mogelijk is aan te tonen dat meerkeuzetoetsen betrouwbaarder zijn als toets (en niet alleen op de scoringsprocedure, die hier immers buiten beschouwing gelaten wordt), welke betekenis moeten we dan aan dit resultaat hechten. Zoals altijd dus weer de vraag: Gut, wat interessant, Mellenbergh, maar wat voor betekenis moeten we nu aan die resultaten hechten?

1.3.6 SAMENVATTING

M. komt hier weer even terug op de scoringssleutel die in staat stelt opstellen m.b.v. de computer te beoordelen.
M. schrijft dat op deze wijze het misschien mogelijk is de betrouwbaarheid van de beoordeling van nietgeprecodeerde vragen te verhogen. Dat is best mogelijk, maar wat winnen we met een grotere betrouwbaarheid wanneer bijvoorbeeld relevante validiteitscoefficienten, gelijk blijven? (om maar iets te noemen).

1.4 Vergelijking van de criterium validiteit.

1.4.1 Criteriumvaliditeiten van open vragen en meerkeuzevragen.

M. signaleert hier gelukkig wel het probleem van het criterium. Hij gaat echter niet verder dan het uitputten van het vermoeden dat schoolcijfers die als criterium gebruikt worden waarschijnlijk bepaald zijn d.m.v. opsteltoetsen, zodat bij een vergelijking tussen meerkeuze en opstel toetsen met schoolcijfers als criterium de laatste in het voordeel zijn. Over het algemeen worden geen grote verschillen tussen beide typen toetsen gevonden. In veel gevallen vertonen de meerkeuzetoetsen wat hogere criteriumvaliditeits coefficienten.

Het wordt niet expliciet gezegd, maar waarschijnlijk is het zo dat in dit geval naast verschil in vorm (opstel of meerkeuze) ook de scoringsprocedure (objectieve vs. subjectieve beoordelings procedure) een rol speelt. Ik vind tenminste nergens een aanwijzing dat in de vermelde onderzoekingen deze scoringsprocedure constant gehouden is. Als dit vermoeden juist blijkt te zijn (wel even controleren met een paar steekproeven uit de vermelde onderzoekingen) dan zijn de resultaten van al dit onderzoek werk uiterst schamel en nauwelijks in het voordeel van de meerkeuze toets.

De bovengenoemde bedenkingim van M. tegen het gebruik van schoolcijfers als criterium is volslagen irrelevant. Er zijn wel forsere bedenkingen (daar moeten we dan maar weer een stukje over schrijven). Bovendien verzuimt hij aan te tonen (op psychometrische gronden bijv.) dat zijn bedenking re~;el kan zijn.

1.5 Open vragen en meerkeuze toetsen als operationalisaties van dezelfde variabele.

1.5.1 Meten de verschillende typen toetsen hetzelfde?

M. blijkt het begrip "meten" nogal beperkt op te vatten. Hij vermeldt hier alleen onderzoekingen waar "de" correctie voor attenuatie op de correlatieco;~fficignt tussen twee typen tests wordt toegepast. De procedure is vervolgens dat nagegaan wordt of deze gecorrigeerde correlatie significant (op 5% niveau) kleiner is dan 1.00. M. zegt dat voor de toetsing gebruik werd gemaakt van een formule vermeld in Forsyth en Feldt (1969), terwijl de onderzoekingen die hij citeert alle van vroegere datum zijn. Bedoelt hij dan dat hij, M. zelf. de toetsingen heeft uitgevoerd? Deze procedure heeft tauwelijks iets te maken met de vraag die gesteld is (meten de toetsen hetzelfde?). Alleen op basis van dit soort empirische analyse kom je daar niet achter. Misschien meten ze dezelfde antwoordtendenties., om maar iets te noemen. Of misschien meten ze dezelfde scoringstendenties van docenten. Het gaat namelijk om representaties. Voordat we test A met test B vergelijken mogen we eerst wel eens behoorlijk uitzoeken of test A zowel als test B eigenlijk wel datgene meten wat bedoeld wordt te meten; de beantwoording van deze vraag is verre van eenvoudig (maak een schets van een procedure die voor een dergelijke probleemstelling gehanteerd kan worden; kijk daarvoor eerst een bij Bezembinder).

1.5.2 Samenvatting.

M. noemt hier als bezwaar tegen geneemd soort onderzoek dat de correcties voor attenatie te hoge correlaties opleveren (en het is niet bekend hoeveel ze te hoog zijn).
M. verbaast vervolgens met de uitspraak dat op grond van de bestaande literatuur wij geneigd zijn om te concluderen dat in sommige gevallen geprecodeerde toetsen hetzelfde meten als nietgeprecodeerde toetsen en dat dit in andere gevallen niet zo is.
Tenslotte komt hij met een volkomen van de context losstaande opmerking waarvan het veel meer de moeite waard zou zijn dieverder uit te werken dan al het gezeur over attenuatie:
"Als een doelstelling van een bepaald soort onderwijs vooral is om de bestudeerde stof actief te produceren, dan lijkt het heel goed mogelijk dat geprecodeerde toetsen deze onderwijsdoelstellingen niet adequaat kunnen meten". (hier gebruikt hij het begrip "meten" dus in een andere betekenis dan bij het attenuatieonderzoek).

1.6.1 Wat zijn de effecten van verschillende vraagvormen?

Een dorre opsomming van enkele experimenten waar geen enkel inzicht in de problematiek uit verkregen kan worden.

1.6.2 Samenvatting.

M. concludeert in "enkele" experimenten (ni. door M. genoemde) in geen enkel geval een verschil in prestaties kan worden aangetoond bij verschillende manieren van voorbereiding (bijv. op meerkeuzetoets of voorbereiding op opstel). En daarmee is voor M. dan de kous af. Dit korte sokje kan denk ik toch nog een heel stuk "uitgebreid" worden dan M. doet. Om te beginnen zou het wel eens zo kunnen zijn dat mensen die zich voorbereiden op een opstel toets meer inzicht in de stof verwerven die het hun eveneens makkelijk maakt om bij meerkeuze vragen tot een goede keuze te komen bij detailvragen. Aan de andere kant kun je je afvragen of mensen die zich prepareren op een meerkeuze toets door de stof gedetailleerd te bestuderen, om die reden wel minder in staat ziuden zijn om wanneer dat gevraagd wordt een goed opstel te maken(m.a.w. op welke a priori gronden verwacht M, of verwachten de door hem geciteerde onderzoekers, dat dit wel eens niet het geval zou kunnen zijn? Met geen woord rept M. hier over). Bovendien missen we hier, evenäls trouwens overal, een evaluatie van methodologische en statistische opzet van de onderzoekingen, die de resultaten wat meer reli;;f zouden kunnen geven (of, ook niet onbelangrijk, het vertrouwen in de vermelde resultaten een behoorlijke trap kunnen geven).

1.6.3 Over de effecten van de vorm van toetsing of studiegedrag en studiesucces zijn door meerdere mensen opmerkingen gemaakt, al of niet gefundeerd op onderzoek.
Omdat M. ze niet vermeldt, zullen we dat zelf maar weer moeten uitzoeken. Ten eerste bij De Groot, vervolgens bij Gronlund, Ebel, Gage e.d.)*
Bijvoorbeeld is een belangrijke overweging dat niet een verschil voor een bepaalde toets van belang is, maar het effect van toetsmethode op het gehele complex van onderwijs dat iemand van zijn zesde tot god mag weten welke leeftijd doorloopt.

STUDIE II : Een onderzoek naar het beoordelen van open vragen

Laten we maar aannemen dat de statistiek zoals die hier gepleegd
is in orde is (hoewel ik daar graag wat meer van zou willen weten
dan hetgeen M. vermeldt).
Het eerst opgemerkt moet dan worden dat de hele ingewikkelde toestand die M. opbouwt een puur luchtkasteel is. Het is werkelijk volmaakt irrelevant of iemand nu 33 punten krijgt, of 40, zolang hij er maar tenminste 23 heeft, want dan is hij geslaagd. De schaal van 040 waar Mellenbergh van uitgaat, heeft geen re ele betekenis, de werkelijk gehanteerde schaal is een dichotomié,2, scores < 23 en scores > 23. Welnu, dan moeten we bij alle bere~ keningen en schattingen van correlaties en betrouwbaarheden uit gaan van deze dichotomie, waarmee de irrelevante nauwkeurigheid van de 41punts schaal die nu komt te vervallen zal resulteren in schrikbarende daling van de diverse betrouwbaarheidsco;~ffi ci;~nten (zie bijv, tabel 22 om een indruk te krijgen van de om vang van de rampzaligheden die we zo moeten constateren: bij 42% van de studenten hangt de beslissing zakken of slagen af van het feit welke beoordelaar de antwoorden van de student nakijkt (ervan uitgaande dat deze scores al gegeven zijn, anders moeten we de betrouwbaarheid van de individuele beoordelaar daar nog bijtellen).
Natuurlijk is het ook belangrijk om iets verder te gaan dan alleen een analyse van betrouwbaarheden. Jammer genoeg vindt Mellenbergh het al welletjes als hij een reeks geflatteerde betrouwbaarheidscoefficienten heeft geproduceerd.

STUDIE III : HET BEOORDELEN EN KLASSIFICEREN VAN ITEMS.

Ook hier zien we weer een grenzeloze vanzelfsprekendheid bij het presenteren van deze beoordelingsproblemen als dè beoordelingsproblematiek in het onderwijs. Deze impliciete keuze van Mellenbergh, en met hem bijna alle overige docenten, testpsychologen en iteminstructeurs, zal tegenwicht gegeven moeten worden; belangrijk hiervoor is bijv. zoiets als het artikel van Cronbach in Gronlund, (specifiek onderwijskundige overwegingen), het beslissingsmodel van Cronbach & Gleser (psychometrischstatistische overwegingen), en analyse van het maatschappelijk systeem dat bijv. concurrentieprincipes omvat, ongelijke kansen voor ongelijke klassen, eenzijdige keuze van het model van de beroepsvoorbereiding als kenmerk voor het onderwijs, e.d.

Het leveren van kritiek op het enorme aantal flutonderzoekjes door Mellenbergh aangehaald belooft een eindeloos vervelende en frustrerende zaak te worden. Ook hier is het weer een kritieken evaluatieloze presentatie van onderzoe3c opzetten en resultaten waarbij het aan de lezer overgelaten wordt om uit te maken of één en ander wel relevant is voor de probleemstelling, en of de statistieken en co;~ffici~~nten die aan de lopende band geproduceerd worden ook maar enige betekenis hebben. Immers, een hoge betrouwbaarheidsco~ffici~nt op zich betekent niets, het vermelden daarvan zonder enige evaluatie is het je onledig houden met noaal zinloze uitspraken.

3. Deze meer algemene bezwaren wil ik illustreren aan het door M. gepresenteerde analyse, waarbij allerlei detailkwesties buiten beschouwing blijven.

3.2 Het beoordelen van de moeilijkheid van items.

Het probleem wordt door Mellenbergh onmiddellijk ingeperkt tot beoordeling door de docenten. Een verborgen autoritair trekje dat kenmerkend is voor het hele rapport: beoordeling in het onderwijs is een oordeel over studenten uitgesrpoken door docenten.

Misschien is een dergelijke "mentaliteit" wel een onveranderlijk kenmerk van de studietoets technologie. (in dit geval dus: studietoetstechnocratie).

Onderzoek SISSINGH (1965). Vijf beoordelaars, vijftig items. De correlatie tussen geachatte en gevonden pwaarden is erg laag: .20, .20, .32, .35 en .40. Dit schijnt te betekenen dat een docent niet van te voren kan zeggen hoe moeilijk een bepaald item voor een bepaalde groep studenten is, een conclusie die M. ook trekt. Uit de geringe (maar positieve) correlatie tussen beoordelaars trekt M. de conclusie dat zij blijkbaar een andere opvatting hebben over wat moeilijk en makkelijk is; waarmee hij de voorkeur geeft aan een ingewfakkelde uitspraak waarbij de betekenis van "moeilijk" of "makkelijk" in het midden gelaten wordt, boven de meer voor de hand liggende conclusie dat de beoordelaars er gewoon een slag naar slaan, m.a.w. slechts een zeer flauwe notie hebben van wat met "makkelijk" of "moeilijk" bedoeld zou kunnen zijn. Nog eenvoudiger: bij lage correlatie tussen schattingen van de individuele beoordelaar en de gevonden pwaarden, kunnen de overeenkomsten tussen verschillende beoordelaars ook slechts minimaal dat zijn; tenzij het onwaarschijnlijke geval zich voordoet/de beoordelaars het onderling "eens" zijn, en de toets in flagrante afwijking daarmee blijkt te functioneren! En inderdaad, de onderzoekers hebben zich al heel eenvoudig van dit probleem afgemaakt door een moeilijkheidsdimensie te definieren als de ratio van het aantal goede antwoorden tot het totale aantal antwoorden, dowoze de pwaarde*

3.2 Dat bij het nemen van gemiddelden voor de beoordelaars de realiteit beter benaderd wordt. is waarschijnlijk alleen al op grond van de "wet van de grote aantallen" te verwachten, en heeft geen practische betekenis. (Immers, het is eenvoudiger de toets af te nemen, om steekproefschattingen van de pwaarden te kunnen verkrijgen, dan de toetsitems door een aantal docenten op moeilijkheid te laten beoordelen. (Het door M. geciteerde onderzoek geeft geen enkele qarantie dat ook in andere gevallen de gemiddelde schatting in de buurt van de empirische pwaarde moet liggen, hooguit geven extreme schattingen aanwijzing dat er wel eens iets mis kan zijn).

De generalisaties van M. uit een handjevol onderzoekresultaten (van Sissingh, en twee onderzoeken van DerksenMögelin) gaan mij toch wel te ver. Ik zou graag ondersteuning zien door andere onderzoekingen, die in de literatuur bij bosjes te vinden moeten zijn. M. noemt echter geen enkel ander onderzoek. Het lijkt mij ook nogal dwaas om op deze wijze 7 blz. van een dissertatie te vullen alleen om aannemelijk te maken dat het verstandig is bij toetsconstructie niet af te gaan op de geschatte itemmoeilijkheid, maar gewoon af te wachten hoe de items het in de praktijk blijken te doen; daarbij de meer gevinteresseerde lezer verwijzend naar enkele onderzoekingen die deze uitspraak ondersteunen.

3.3 Het beoordelen van de kwaliteit van items

Drie methoden werden toegepast (door M.?): 1. items laten beoordelen door docenten en toets ontwikkelaars 2. items analyseren aan de hand van de empirische gegevens 3. items laten beoordelen door de studenten, die ze opgelost hebben.

We hebben hier weer het bekende euvel: M. stort zich hals over kop in de experimentjes om antwoord op zijn probleemstelling te vinden, zonder eerst te bespreken wat hij bedoelt met zoiets als "kwaliteit" van items. Uit zijn verslag blijkt dat kwaliteit zeer eng opgevat wordt: allereerst dat er geen elementaire tekortkomingen in de formulering van de items mogen voorkomen, ten tweede dat een aantal coefficienten een redelijke grootte bereikt. Het eerste is vanzelfsprekend, het tweede wordt als vanzelfsprekend gepresenteerd waar vanzelfsprekendheid geenszins vanzelfsprekend is. Daarover gaat de volgende paragraaf.

3.3.1 Statistische itemgegevens

rit itemtest correlatie co~ffici~nt

rio item validiteits coeffici;nt (in dit geval correlatie tussen item scores en extern criterium)

fi bijdrage aan de signaalruis verhouding van de test (positief betekent dat het item de betrouwbaarheids coëfficiënt verhoogt, negatief dat de betrouwbaarheids ccoëfficiënt verhoogd wordt).

p : de fractie personen die het item goed heeft beantwoord.

3.3.1 Van Naerssen (1967) vond dat r it en f i weinig stabiel zijn van steekproef tot steekproef. Dit is niet zo best, zoals Klaas opmerkt, omdat vooral de r it een belangrijk criterium is voor het al of niet handhaven van een item. (en de f i blijkt zeer sterk met r it gecogrdeerd te zijn, volgens hetzelfde onderzoe3c van Van Naerssen). M. trekt geen enkele conclusie uit deze ge gevens (die op zijn minst voor de nodige "verontrusting" onder de toetsboeren zou moeten zorgen), en komt ook in de samenvatting en conclusie niet op deze resultaten terug. Dat betekent o.a. dat de lezer zelf maar het verband moet leggen tussen deze re sultaten en de waarde van al hetgeen er verder in deze paragraaf met r it gegevens gegoocheld wordt.

Over pwaarden wordt een zeer verwarrend relaas gedaan. Items met lage pwaarden blijken volgens

M. vaak items te zijn die gebreken vertonen (zoals: het als juist bedoelde alternatief is niet één'duidig juist). De bijdrage aan toetsbetrouwbaarheid is negatief (v.Naerssen, 1966 a), waarbij M. volstaat met het citeren van dit resultaat en er geen verdere analyse aan wijdt. Wanneer een item, dat goed samengesteld is, een lage p waarde heeft dan is het zo moeilijk dat slechts een onbetekenende fractie van de studenten het item goed maakt door eigen inzicht, de overigen hebben het item toevallig goed. Dit betekent dat het item vrijwel geen informatie geeft. De vraag is nu, en daar geeft M. geen antwoord op, of meerdere van deze items niet in staat zijn om te discrimineren tussen de allerbesten en de iets mindere goden in de groep. Ik dacht dat dit wel mogelijk was, daarvoor moeten we Lord & Novick maar eens raadplegen. Een andere vraag is of, ook al blijkt dit mogelijk, we dan ook maar van deze mogelijkheid gebruik moeten maken. Ten eerste gaat het om eenvoudig onderscheid tussen zakkers en slagers, waarbij verschillen in de groep van slagers eigenlijk niet interessant zijn (tenminste, dit is wat in de onderwijspraktijk het doel is van beoordeling). Ten tweede, zouden we toetsen willen gebruiken als feedback naar de student over zijn beheersing van de stof (en niet over zijn relatieve positie op de klasse ranglijst) dan kunnen we waarschijnlijk veel beter van klassieke opsteltoetsen gebruik maken.

Ten derde. het opnemen van items die slechts door een enkeling bewust goed gemaakt worden, betekent een forse frustratie voor alle overige studenten. Ten vierde. zou het volgens Klaas wel eens zo kunnen zijn dat de items die gé6n lage pwaarde hebben, juist de "brave" items zijn, items die naar domme kennis e.d. informeren. Met dit laatste ben ik het niet zonder meer eens. het is namelijk niet noodzakelijk zo; een goede toets is goed afgestemd op het voorafgaande onderwijs, en wanneer dit onderwijs van voldoende kwaliteit is geweest (zowel inhoudelijk als naar de mate waarin de studenten gemotiveerd meegedaan hebben), dan moet het grootste deel van de studenten ook bijna alle vragen in de toets goed kunnen beantwoorden. Er is geen enkele reden om te veronderstellen dat beoordelingscijfers altijd een normaalverdeling of een rechthoekige verdeling zouden moeten volgen. Over pwaarden is, op grond ook van andere literatuur, waarschijnlijk nog wel iets meer te zeggen. Dat komt dan een andere keer maar*

3.3.1 Het gescharrel met itemvaliditeiten is mijns inziens een kwalijke zaak. Ten eerste lijkt de vraag naar item validiteiten mij weinig zinvol, het gaat tenslotte om de toets als geheel, en daarover
geeft M. geen enkele uitspraak. (Dit zal ik natuurlijk wel moeten beargumenteren, dat stellen we dan maar weer even uit).
Ten tweede wordt alleen gesproken over predictieve validiteit, waarmee het juist voor studietoetsen primaire belang van inhoudsvaliditeit in de doofpit gestopt wordt.
Ten derde wordt gehanteerd de predictie van zoiets belachelijks als het aantal tentamina dat de studenten in één jaar gehaald hadden. Hierover valt op te merken dat een zekere superieure correlatie hiermee bestempeld wordt tot validiteit (immers, alle volgzame figuren die een behoorlijke studietoetsinstelling hebben ontwikkeld, zullen dit soort correlatie lekker opfokken), vervolgens dat er nog wel iets meer te voorspellen zou kunnen zijn dan dit soort belachelijkheden (maar ja, Van Naerssen heeft al laten zien, in het kielzog van Cronbach en Gleser, dat in dat geval je maar beter helemaal op kunt houden met toetsen), en tenslotte, dat waar Van Naerssen al aantoonde (overigens in slechts één onderzoek), dat r it een nogal "onbetrouwbaar" gegeven is, mogen we dit ook van r iv verwachten, immers het is, dacht ik, nog steeds zo, dat de betrouwbaarheid een bovengrens vormt voor de validiteit (r xt < √r xx, d.w.z. correlatie van waargenon en "ware" score is kleiner of gelijk aan de betrouwbaarheidsindex).

3.3.2 en
3.3.3 Het beoordelen van de kwaliteit van items door docenten, itemconstructeurs en studenten.

Dat na de ongelooflijk magere gegevens uit de vorige paragraaf, en het al eerder gebleken onvermogen van beoordelaars om de pwaarden te schatten, het nog enige zin heeft om de titelvraag te gaan onderzoeken, dat betwijfel ik. En de resultaten blijken er dan ook wel naar te zijn (nog afgezien van de vraag wat nu eigenlijk de practische relevantie van deze probleemstelling is; stel bijv. dat het mogelijk is een aantal indices een heel klein beetje goed te schatten, dan is het nog altijd zo: the proof on the pudding is in the eating).

Het enige significante resultaat (dat echter niet als zodanig vermeld wordt) is dat van de 53 studenten die gevraagd werden een vragenlijst over items in te vullen, na afloop van de toets, er 21 de enquéte blanco inleverden (dit onder voorbehoud dat de genoemde aantallen, zie blz.97, betrekking hebben op hetzelfde onderzoek als in 3.3.3 waar helemaal geen opgave bij vermeld is van het aantal studenten dat meedeed en/of het aantal weigeraars).

3.3.4 Wat maakt een item slecht?

Een dorre opsomming van wat er allemaal aan een item zou kunnen mankeren. Een checklist voor een itemtechnd)keuring. Wat je mist is een vergelijkbare lijst, maar dan voor complete studie toetsen. Een dergelijke lijst zal er waarschijnlijk nogal anders uitzien dan voor enkelvoudige items. En lijkt me bovendien veel meer relevant dan het gezeur over punten, komma's en grammatica. Bijvoorbeeld de relevantie: in hoeverre is de toets relevant ten opzichte van wat men wil meten met de toets? De beantwoording van deze vraag is heel wat minder simpel en vanzelfsprekend als de vraag over de relevantie van een item (en ook hier geldt dat de som van relevante items geenszins een relevante toets hoeft op te leveren).

3.2.5 Het verkriigen van meer gedetailleerde informatie over de kwaliteit van items.

Hier worden weer onderzoeJcjes van het bekende lullige soort besproken. Betrouwbaarheid van beoordelingen van items aan de hand van enkele vragen (over formulering, alternatieven e.d.) blijken verre van ideaal te zijn (lijkt mij, M. spreekt geen oordeel uit). Hieruit zou dan de conclusie moeten volgen dat je van een beeritiseerd item mag aannemen dat er wel eens iets mee aan de hand zou kunnen zijn, en dat je van items waar geen kritiek op komt nog niet weet of ze goed zijn of niet. (Of die kritiek nu van de testconstructeur is of van anderen is dan niet relevant) .

3.3.6 Samenvatting en conclusies.

Mellenbergh besluit met de mijns inziens volstrekt niet door de door hem aangedragen gegevens ondersteunde conclusie: "De algehele indruk van deze onderzoekingen is dat beoordelaars betrouwbare en zinvolle informatie kunnen leveren over items". Een uitspraak die bij direct weer probeert te relativeren: "De resultaten van de onderzoekingen zijn echter van dien aard dat men niet een te groot vertrouwen mag hebben in beoordelingen van items". Wat moet de lezer nu geloven?

3.4 Het klassificeren van items naar een onderwijskundig gezichtspunt.

Vooral de categorie~n kennis, inzicht en toepassing van Bloom c.s. De inperking tot de eerste drie categorieen voorspelt al niet veel goeds. Immers, volgens Bloom c.s. is het zo dat de complete hiërarchie op alle niveaus van onderwijs terug tevinden is. Daarom is het merkwaardig dat er dan ineens belangrijke onderwijsonderdelen blijken te bestaan die volgens de docenten/ toetsconstructeurs beoordeeld kunnen worden door alleen op de eerste drie categorie~n te letten. (Zoek de betreffende uitspraak van Bloom op).
Dat beoordelaars in staat zijn te onderscheiden tussen kennnis, begrips, en toepassingsitems, is nauwelijks verwonderlijk.
Bloom rapporteert al dat het klassificeren van items in het gehele categorieënsysteem (mét subcategorieën) aardig lukt.

(Ook deze passage opzoeken). Hieruit zou je kunnen concluderen dat de door M. geciteerde onderzoekjes nogal irrelevant zijn. Dan volgt ineens, in afwijking van de tot nu toe in de dissertatie gevolgde procedure, een bespreking van een onderzoek van Van Dam met een uitgebreide kritische bespreken, gevolgd door een enorm uitgebreid verhaal over een replicatie van dit onderzoek (ook door Van Dam uitgevoerd?), dat afgesloten wordt met een aantal vragen die het gebrek aan resultaat zouden kunnen verklaren. waarbij je je meteen afvraagt of de onderzoekers soms onnozel waren om een onderzoek te doen zonder zich van te voren bezig te houden met de nu achteraf geciteerde problemen. Het onderwerp van onderzoek, en de resultaten, zijn dus gewoon het vermelden niet waard.

5. Evaluatie.

Mellenbergh: "In de eerste plaats hebben wij aangetoond dat het beoordelen van openvvragen een hachelijke zaak is". Nee Mellenbergh, jongen, dat heb jij niet aangetoond. Je hebt wel aannemelijk gemaakt dat de scoring betrouwbaarder gebeurt bij meerkeuze toetsen dan bij opstellen; en dat ligt voor de hand nietwaar, het zou al heel vreemd zijn als dit niet zo was. Maar om daar dan de conclusie uit te trekken dat het beoordelen van open vragen een hachelijke zaak is, is het bedrijven van propaganda (en met propagandistische en beschouwende zaken zou M. zich niet bezig houdenrp zie 1.1). Dezelfde fout maakt hij waar gesteld wordt dat daar waar grote belangen voor de student de inzet van de beoordeling zijn, deze zo betrouwbaar mogelijk moeten zijn, daarbij vergetend dat betrouwbaarheid vaak wel een noodzakelijke, maar absoluut geen voldoende voorwaarde is voor verantwoorde selectieprocedures. Bovendien maakt hij de zeer fundamentele fout door te veronderstellen, dat een selectieprocedure die gericht is op groepen ook voor individu~n rechtvaardig zou zijn. Zie hierover Cronbach & Gleser. Maximalisering van rendement voor een fabriek (in dit geval "universiteit", "school" e.d. genoemd) heeft niets te maken met optimalisering voor het individu. Dit vereist een volkomen ander statistisch model, wat voor zover ik weet, nog niet in enige zinvolle mate uitgewerkt is.

(Misschien komt voor sommige gevallen het model van Rosch in de buurt).
De rest van zijn "evaluatie" is obligaat geouwehoerg of een her haling van de spaarzame opmerkingen die hij al eerder gaf.

uÑ≈13%'+.£∏›ÓßS_§∞Δ”–C/3!–"$Ì$˜*ò* ,Ù–'1Δ1Œ4Ø4·7ü7¥===Ÿ>>

R123]^≤onoºΩæËÈ67 ? @
«
»

7
8
v
≠
Ê
˘
y
z
π
¯8f¶„û~:;rÛÙù

%‘%‘%‘
%‘%‘%‘%‘%‘%‘%‘%‘M,',k,†,°,Ó,Ô–'–k–®–„.i/B/
√33404P9t9u9v:Í<<<<*<+<û=∫=‘=’=÷>>>G>H@b@cB¶BßC€C‹C›D2D3DiDjIuIvIwIåIçJ5K0K,M*M+M,MfMgMÿMŸMÓMÔSS
SßT5U(U)U*UhUiU™UÏVZP[9[:[;
[s°°°ˆ°°°°°ÒÏˆÁ°,°›°°,ÿ°°°°ˆÿ°°°°°°ˆ°”°Œ°ÿ°°ˆ°°°›°°°°ÒÏÒÿ°°°°..
%‘ %‘%‘%‘%‘%‘%‘%‘%‘%‘O[s[t^Y^Z` `ÒaàaâaªaºbπbʃcRcShhÎhÏkÇkÉn?
n@nnnooZo[q«q»qÈqÍrrrJrsr™r‡ss:svswsxv\v]|P|
QÄIÄJÅÅéÇ6Ü0Ü1Ü2Ü;ÜöÜúà∏äxäyäöäõçjçkçlç¨ç¬ç√êêê$ê‰ëˇííGíIîPîŒïHïIïJòMòNò\ò]ù¬ù√ûû
ûMûàûâûäûãûåûç°°
°°°°°°°°°°°]
$$
ﬁùçûç™
πıÉ"Ü)ë0–8=HC‰KgRYP_ÒgnZt |ﬁÖ0åíôùçˇˇåˇˇ?ˇˇ}ˇˇèˇˇœˇˇˇˇ„ˇˇ^ˇˇ ÷ˇˇ
ëˇˇ√@ˇˇ
pPˇˇ
È@ˇˇóPˇˇ©@ˇˇPˇˇí@ˇˇkPˇˇ@ˇˇLPˇˇE@ˇˇØPˇˇûìP,'[sûçQRSU–“”‘’÷◊/€›,˜¯ ”°æ#/–/
ã7Ö?ˇHP–Xˇ‹ˇÿÙÄxﬁHP ÄP Ä
Ä'=‡/–`†≈U–LPPalatino¿ùâùâÄÄùâ(Uå(6å