Aantekeningen bij Wim van der Linden Decision models for mastery testing (APM 1980). 80-5-22_aantekeningen_vdlinden.rtf 80-5-22_aantekeningen_vdlinden.pdf in map literature op de website

Linden, W. J. van der (1980). Decision Models for Use with Criterion-Referenced Tests 469-492 *Applied Psychological Measurement* Volume *4*, Number 4 Fall, 1980 SPECIAL ISSUE Contributions to Criterion - Referenced Testing Technology Guest Editor: Ronald K. Hambleton

22-5-80

De aantekeningen zullen nogal uitgebreid zijn, omdat Wim een uitstekend overzicht van de stand van zaken heeft geschreven en er op diezelfde stand van zaken wel het een en ander aaan te merken is. Ik wil in deze becommentariëring een aantal van de issues nog eens duidelijk stellen, duidelijker misschien dan ik in mijn beide TOR artikelen heb gedaan: die zijn zo weinig mogelijk polemisch geschreven om de teksten goed op zichzelf leesbaar te doen zijn, maar ik ben

bang dat een aantal van de pointes die ik er op vriendelijke wijze heb ingewoven aan Wims aandacht ontsnapt zijn. Dat wijst dan op de noodzaak het materiaal uit beide artikelen in de vorm van een engelstalige presentatie te gieten: enerzijds in meer gecomprimeerde vorm, anderzijds met een aantal van de eerder uitgezette lijnen wat verder doorgetrokken.

17 1e helft. Bij mastery testing wordt op het onderliggende continuum T de cutting score aangewezen. Dat klopt, zo gebeurt dat vaak. Dat betekent niet dat het noodzakelijk zou zijn om zoiets te doen (wie met utiliteitsfunkties werkt heeft het niet nodig, zoals ik later nog zal laten zien), of dat er geen andere mogelijkheden zijn dan het aanwijzen van een punt. Wim signaleert dit niet, zodat deze stilzwijgende veronderstelling dat het aanwijzen van zo'n cutting score noodzakelijk zou zijn hem in het vervolg hinderlijk blijft achtervolgen. Dat zal nog blijken.

17-20 Een verbazende passage. Er wordt een onderscheid gemaakt tussen 2 grensskoreproblemen, op de ware en de

waargenomen schaal. O.K., maar die hebben wel iets met elkaar te maken. M.b.v. decision theory (o.a. houdt dat het specificeren van je verliesfunctie, of liever: je utiliteitsfunkties over de doelvariabelen in) kun je de laatste aanpakken. De eerste moet je dan maar zien te vinden op de natte vinger (i.e. op basis van'value judgments and norms')

Hier breekt Wim zijn geringe kennis van de besliskunde **buiten** de decisie theorie a la Ferguson op: wat hier geponeerd wordt impliceert dat de 'value judgments and norms' aan de ene kant iets anders zouden zijn dan de utiliteitsfunkties aan de andere kant wat uiteraard niet zo is.

Wanneer je zorgvuldig je decisie analytische benadering uitwerkt, zoals in mijn 'Enkele radikale oplossingen etc.' dan wordt ook duidelijk welke schijntegenstelling Wim hier oproept. Je zou ook kunnen zeggen dat zich hier het gebrek aan interesse voor wat utiliteitsfunkties inhoudelijk voorstellen, wreekt; niet alleen bij v.d. Linden (en Mellenbergh), maar in vrijwel alle publicaties gepleegd over krit. geref. grensskores, inclusief geloof ik die van Davis Hickman en Novick. En dat is een ernstige zaak, want je

komt op deze wijze maar al te makkelijk tot onzinnige methoden, en onzinnige conclusies.

De eigen waardeoordelen en normen over T druk je uit in je utiliteitsfunktie over T. Hetzelfde doe je met je oordeel over de kosten van remedial treatment. Tn combinatie met de waarschijnlijkheidsverdeling over T, gegeven X=c, voor beide verschillende treatments, kom je dan tot de cuttingscore op T als het snijpunt van de beide zo verkregen **uitkomst**utiliteitsfunkties. Deze **uitkomst**utiliteitsfunkties zijn nimmer uit het blote hoofd te specificeren; in sommige gevallen zijn **uitkomst** veriiezen wel direct te becijferen.

Optimaliseer je c voor een gegeven waarde voor d, dan is dat slechts een zinvolle uitspraak wanneer om te beginnen d bepaald is op decisie analytische wijze, via het juist kort geschetste model, dat ik in 'Enkele radikale oplossingen etc.' gespecificeerd heb (zie ook figuur 4 daarin).

19 De opmerking van Wim over de methode

Nedelsky terecht. Hetzelfde kan bij de kernitemmethoden opgemerkt worden, maar Wim vindt nederlandse aangelegenehden te parochiaal om te bespreken. Het laatste is overigens verbazingwekkend, en teleurstellend. Be besproken literatuur bevat geen nederlandse bronnen behalve die waarin Wim als (mede) auteur fungeert. Is wat hier gebeurt verder zo armzalig?

20 Vergelijk de opmerking over Glass met voetnoot in TOR, 5, p. 52.

21 Met de state space omega bedoel je gewoon T, dus. Tenzij je utiliteitsfunktie over T een drempelfunktie is, dan zou je kunnen spreken over een state space omega met twee states. Het zou dan ook heel wat eenvoudiger zijn wanneer je deze verdubbeling van symboliek weglaat, en over individuele 'states' spreekt als t i.p.v. kleine omega.

21 De verliesfunktie l(t,a) is blijkens de formulering een 'negatieve uitkomstutiliteits'funktie (geen gelegenheidsverlies; uit de voetnoot is dat niet duidelijk, Wim zegt daarin alleen dat hij

de term gelegenheidsverlies niet gebruikt).

21-22 Het staat er niet met zoveel woorden, maar Wim stevent in zijn specificatie van het probability model rechtstreeks af op de impliciete opvatting van mastery testing als selective testing. Ik heb in 'Enkele radikale etc.' nu juist laten zien dat er bij kriterium. gerefereerd beslissen **twee** states of nature relevant zijn, de ene is gewoon T, de andere is T-na-bijspijkeren. Tenminste een **twee-dimensionale** statespace dus, hoewel twee 'dimensionaal' niet precies het goede woord is.

En er is dan ook een **tweevoudig probability model** nodig

Het punt is dan ook niet dat ik op enigerlei wijze bezwaar zou hebben tegen een decisie theoretische benadering in de stijl van Ferguson: ik denk dat je er met zo'n benadering misschien ook nog wel uitkomt (hoewei de nromal form analyse dan toch wel wat extra problemen zal gaan opleveren), maar dan moet het probleem wel in de juiste termen gesteld worden, tenminste als 44n waarbij een **dubbel**

waarschijnlijkheidsmodel aan de orde is.

22 Met het verwachte verlies (1) is iets opmerkelijks aan de hand, dat waarschijnlijk ook te maken heeft met het verschil tussen normal form en extensive form analyse. Het is het **verwachte veriies voor een gegeven t**. Dat zegt op zich nog niets over de optimale actie a. Het zegt wél iets over de statistische decisie theorie: die werkt vooral graag met dít verwachte verlies. In de decisie analyse (zeg: de Harvard school van Raiffa) speelt een ander verwacht verlies de hoofdrol: **de verwachte negatieve utiliteit voor een gegeven a**. (of: verwachte gelegenheids verlies, of verwachte utiliteit, dat maakt niet uit).(Raiffa & Schlaifer 1961 formule 1-1. Het verwachte verlies voor gegeven t correspondeert dan, als ik het zo snei goed heb bekeken, aan Raiffa en Schlaifer's (1-5). Inderdaad het verschil tussen extensive en normal form analyse.

24-26 Deze paragraaf is hinderlijk pedant in overdreven mathematiek. Dat contrasteert bovendien nogal met de ongenuanceerde opmerking aan het eind van de par. 'Then it may prudent etc' : welke reden er zou zijn om

voorzichtig te zijn, daar laat Wim de lezer over in het ongewisse, en eveneens over wat een 'optimal decision **rule**' dan wel is en waarom dergelijke frules' voorbehouden zouden zijn aan de nromal form analyse. Overigens, Davis, Hickman en Novick zijn niet de enigen die bepaalde voordelen aan de extensive form analyse toeschrijven: deze heeft in mijn ogen aanzienlijke wiskundige en heuristische voordelen bij het kriterium gerefereerde grensskore prohleem; Raiffa en Schlaifer zijn evenmin enthousiast over de normal form analyse. Bovendien: bij beslissingsproblemen buiten de mathematische statistiek à la Ferguson (management, bestuur) wordt uitsluitend de extensive form analyse gebruikt (vgl. Schlaifer's Probability and statistics for business decisions; an introduction to managerial economics under uncertainty. McGraw-Hill, 1959, en Keeney & Raiffa 1976).

Er is een klein probleempje met die aanname van monotonie: dat hoeft ook bij criterium gerefereerde grensskores helemaal niet reëel te zijn. Het is niet uitgesloten dat er deelnemers zijn die geen baat hebben bij

bijspijkeren en hertoetsen: ten aanzien van hen zou je dan ook bij voorkeur dat extra verlies aan onproductieve tijd willen vermijden, en hen ófwel doorlaten, ófwel definitief afwijzen, ófwel terugverwijzen naar een voorafgaand (voorbereidend) stuk onderwijs dat kennelijk niet behoorlijk gedaan of verwerkt is. Maar ja, dat is binnen het **dubbele** probabilistic model .....

Tenslotte: het is een beetje hinderlijk wanneer er gesproken wordt over 'series beslissingen' die gemaakt moeten worden: dat is een beetje onzinnig, je maakt geen series beslissingen maar wijst een grensskore aan. Het beslissingsprobleem is het aanwijzen van de 'optimale' grensskore. Alleen wanneer je slechts t.a.v. een enkele student een beslissing doorgaan-of-niet hoeft te nemen, kun je volstaan met het in die termen te gieten.

Het taalgebruik 'series van beslissingen' stamt waarschijnlijk uit zoiets als quality control settings, waar je dat inderdaad letterlijk moet doen: telkens weer opnieuw. Nou ja, ik weet het ook niet precies, maar de terminologie van Wim is hier een beetje glibberig, terwijl er wél de afwijzing van de extensive form

analyse aan opgehangen wordt, en dat bevalt me niet.

25 (midden in de blz.) Voor een goed begrip zou het geen kwaad gekund hebben aan te geven dat de verliesfunkties $l_0(t)$ en $l_1(t)$ betrekking hebben op een **gegeven** t, en niet over de verdeling van t getotaliseerd zijn.

35 (bovenaan) Dat herschalen gaat niet zomaar: wat Wim hier doet is geen herschalen, maar overgaan op gelegenheids verlies (dat komt er hier op neer dat je voor iedere state of nature **afzonderlijk** herschaalt.

35 (form. 22) Deze formule is uiteraard equivalent aan de bekende selectieratio
P = (B-A) / (B-A+C-D)

36 Wim raadt aan alleen een grensskore te bepalen wanneer beschikbare steekproeven groot genoeg zijn om stabiele oplossingen te garanderen. Dat is een vreemd advies: het punt is nu juist dat er een beslissing (t.a.v. de grensskore) genomen **moet** worden, ook al heb je vrijwel geen empirische gegevens. Wat je in zo'n situatie doet is dan ook de

optimale grensskore aanhouden, gebaseerd op maximale verwachte utiliteit, ook ai zijn je data maar zwak. Vergelijk ook het aardige artikel van Simon in zijn Models of man (Hoofdstuk 2). En bedenk dat dit het principe van decisie analyse is. Er is meestal immers niet de optie om maar geen besluit te nemen .....

39 (form. 24) Ik heb Wim en Don er al eens eerder op gewezen dat de formule zoals hij hier staat een rechtstreekse 'vertaling' is van de bedoeling om de utiliteit proportioneel aan d-t te maken (en dan is het ook inzichtelijker om in de tweede functie d-t te nemen i.p.v. d+t). Het is natuurlijk erg veel handiger om die hele drempel d er uit te halen, **want die speelt geen enkele rol** (is een constante die er gewoon uitvalt).

Een andere opmerking die ik eerder al maakte is dat Wim hier echt in de probiemen komt omdat hij verzuimt zich te realiseren dat noodzakelijkerwijs beide functies eikaar snijden boven de aangewezen cuttingscore d. Zou dat namelijk niet het geval zijn, dan zijn verliesfunkties en aangewezen cutting score niet congruent. Per definitie is de

cutting score het punt waar je onverschillig bent of de klassificatie 'mastery' is, dan wel 'non-mastery' (bij perfecte kennis).

Niets nieuws, Davis, Hickman en Novick wezen er ook al op, p. 59. Daar moet één belangrijke opmerking bij gemaakt worden: dit geldt alleen wanneer beide verliesfunkties met dezelfde waarschijnlijkheidsverdeling over t gewogen worden. Ga je daarentegenover op het meer correcte model met twee waarschijnlijkheids verdelingen (contingent op beide beslissingsalternatieven), dan is de relatie iets minder eenvoudig, maar is de plaats van d nog steed uniek (tenminste: bij monotonie) bepaald op basis van verliesfunkties en waarschijnlijkheidsmodellen (merk op dat in dit geval het niet noodzakelijk is dat de verliesfunkties elkaar snijden om een niet-triviale d op te leveren).

39 wanneer dit (figuur 2) negatieve utiliteitsfunkties over t zijn, hoe moet ik me dat dan in vredesnaam voorstellen?

a. Ik kan me heel goed voorstellen dat je over t een utiliteitsfunktie neemt die stijgt (lineair) in t; dus een verliesfunktie die daalt.

Omdat je bij de beslissing 'doorlaten' alleen met de 'state' t te maken hebt, is dat dan ook de uitkomstutiliteitsfunktie voor 'doorlaten dus voor X groter gelijk c. Oké, prachtig. Maar nu:

b. Voor de treatment 'afwijzen, bijspijkeren, herkansen' heb je met dezelfde variabele T te maken. Welnu, voor één en dezelfde variabele kun je maar één verliesfunktie specificeren. Zo is dat nu eenmaal: die ware beheersing ga je nu niet ineens anders waarderen. Dus voor 'afwijzen' heb je exact dezelfde lineaire verliesfunktie over t als bij 'doorlaten'. Wim tekent er eentje die bijna haaks op de eerste staat Kennelijk is er nog iets anders dan de ware beheersing t die een rol speelt. De enige kandidaat daarvoor zijn de **kosten** verbonden aan de treatment 'afwijzen'. Het moet al heel raar lopen wanneer die kosten niet ongeveer constant zijn over het hele bereik van t (misschien iets hoger bij lage t, iets lager bij hoge t). Dan moet die tweede lineaire uitkomstverliesfunktie dus ongeveer **parallel** aan de eerste (van 'doorlaten'), óf wat meer verticaal dan de eerste, ligt hij in ieder geval **boven** de eerste, en snijden beide elkaar

niet (in het bereik van t).

c. Wim heeft dus nog te verklaren hoe híj dan aan zijn uitkomstverliesfunktie voor X kleiner dan c komt Trouwens, dat zal voor Davis, Hickman, en Novick ook een probleem zijn.

Hetzelfde verhaal geldt natuurlijk voor wat verderop nog aan andere vormen voor deze funkties gepresenteerd wordt. En hetzelfde natuurlijk voor de drempel toestanden in figuur 1.

40 Het zou zo aardig geweest zijn wanneer Wim hier ook had laten zien hoe de extensive form analyse in elkaar steekt (en hoe die op eindeloos veel eenvoudiger wijze tot dezelfde resultaten komt .....).

41 Als de T coordinate of the intersection niet gelijk is aan d zit je in moeilijkheden .....
Het resultaat op deze b1z. had je ook zonder halsbrekende mathematiek kunnen verkrijgen door te bedenken dat inderdaad de 'T coordinate of the intersection' gelijk is aan d, zodat je optimale grensskore c te kiezen is rekening houdend met meetfouten

en dan kom je vanzelf op het hier gevonden resultaat uit. (formule 29).

Je kunt aan deze constante misvatting van Wim (en Don) niet genoeg aandacht schenken, omdat ze leidt tot verdere misvattingen: zoals de 'practical meaning' die aan (29) wordt toegekend. Maar daarover straks, wanneer ook Wim er aan toe is.

46 De filosofie op deze blz. is krakkemikkig, en ook overbodig. Al was het alleen maar vanwege onbegrijpelijk bijvoeglijke naamwoorden 'hard werkende' en 'langzame'. Voor hetzelde geld kun je dat precies omdraaien: 'Luie begaafde' (voor de hoog scorende groep), en 'hardwerkende hoog gemotiveerde' voor de laag scorende groep. Varianten iaten zich zo bedenken.
De filosofie die Wim geeft is geïnspireerd op de algebra van (29), en heeft niets met de onderwijspraktijk te maken.

par. 2-5 geeft een goede samenvatting. Maar de vraag waar ik mee blijf zitten is: Wat kan ik in de praktijk met zo'n coëfficiënt delta uitrichten? Welke betekenis kun je daaraan geven? Mijn vraag is ook: hoe blijven de

resultaten in deze par. overeind wanneer mijn kritiek op de besliskundige aanpak van Wim (en Don) tot een verbeterd model omgewerkt wordt? Komt er dan ook nog zoiets aardigs als (42) uit?

51 Wat Wim hier over ATI zegt, berust op een misverstand. Ik heb dat een aantai keren in de werkgroep ingebracht, en daarmee kennelijk onvoldoende cognitieve dissonantie weten op te roepen om het inzicht in hoe Cronbach & Snow ATI opvatten, recht te zetten. Zie de par. over Cronbach & Snow's ATI in mijn 'Enkeie radikale etc.'.

Wim verwijt: "*it takes the success variable as continuous instead of using the dichotomy mastery non-mastery for this purpose*". Twee dingen: wat is dan wel het voordeel van het gebruiken van die waanzinnig gekunstelde dichotomie mastery non-mastery? De succes-variabele is de **geëvalueerde** succesvariabeieg en dat is hetzelfde wat v.d.Linden en Mellenbergh doen met hun iineaire verliesfunktie over t .....

52 Er is maar weinig ruimte voor het kiezen van andere dan de best passende

utiliteitsfunktie; Keeney & Raiffa leren je al snel dat de vorm van de utiliteitsfunktie tamelijk onwrikbaar is. Wat je na het bepalen van een utiliteitsfunktie nog kunt doen is proberen deze te vervangen door een funktie die prettige wiskundige eigenschappen heeft en de werkelijkheid niet teveel geweld aandoet.

Wat hier over 'robuustheid' van utiliteitsfunkties geschreven wordt, ja zelfs over 'maximale robuustheid', is onzin. Wim wordt er toe verleid omdat hij niet wil inzien dat de cuttingscore d uniek bepaald wordt door de utiliteitsfunktie (de uitkomst verliesfunkties). Het punt is dat d het resultaat is van de specificatie van de verliesfunkties; er is geen sprake van dat d onafhankelijk van die funkties, ook niet van hun parameters, zou kunnen zijn. Jazeker, als je die lineaire funkties laat draaien, terwijl hun snijpunt gelijk blijft, dan blijft ook d staan. Maar dat is zinledige algebra, dat heeft geen enkele relatie meer met waar die funkties voor pretenderen te staan. Hetzelfde uiteraard voor normaal ogief verlies.

54 You can't have your cake and eat it. Er

wordt in het paper een zware nadruk gelegd op het specificeren van die mastery score d, waarzonder je het niet zou kunnen, waar je evaluatief en normerend maar aan moet zien te komen, m.b.t. welke score verliesfunkties maximaal robuust zouden zijn.... En dan zou dat allemaal in de laatste zinnen overboord gezet worden omdat strict formeel gesproken het inderdaad zonder die d kan? Kom nou. Ik vind het uiteraard een correcte conclusie, maar het is een conclusie op basis waarvan je een behoorlijk aantal passages in je artikel op niet-triviale wijze moet omwerken.