Vergelijking van VJTJ en SMVO met behulp van niet-lineaire multivariate technieken

Jan de Leeuw
Eeke vd Burg
Bert Bettonvil
Vakgroep Datatheorie FSW/RUL
Breestraat 70
2311 CS Leiden

Diskussiestuk SISWO-werkgroep 'Longitudinaal'

1: Inleiding

Dit rapport is een tussentijdse rapportage over onze sekundaire analyses van de VJTJ en SMVO data, bedoeld als diskussiestuk voor de werkgroep Longitudinaal van de SISWO. Het sluit aan bij eerdere sekundaire analyses over VJTJ alleen, gerapporteerd in De Leeuw en Stoop (1979), of over SMVO alleen, gerapporteerd in Gifi (1980, 1981). Een eerdere vergelijkende sekundaire analyse, die gebruik maakte van andere technieken, werd gepresenteerd in mei van dit jaar op een data-analyse dag georganiseerd door het CBS.

We gaan er in dit stuk vanuit dat de lezer bekend is met het VJTJ en SMVO-materiaal. We werken met het volledige VJTJ bestand van 1845 respondenten, met dien verstande dat we 57 respondenten die geen sekondair onderwijs volgden uit het bestand verwijderd hebben (om dezelfde redenen als De Jong e.a., 1981). Uit het SMVO bestand trokken we een steekproef van vergelijkbare grootte, en we verwijderden respondenten die een buitengewone vorm van voortgezet onderwijs volgden en respondenten uit gezinnen waarin het gezinshoofd zonder beroep was (zie alweer De Jong e.a., 1981). Dit leidde uiteindelijk tot een VJTJ-steekproef van 1788 individuen en een SMVO-steekproef van 1519 individuen.

We gebruiken zes variabelen uit de twee databestanden. Het selektie-kriterium is hier in de eerste plaats dat de variabelen in beide bestanden moeten voorkomen (en van min of meer vergelijkbare betekenis zijn), en in de tweede plaats dat de variabelen in eerdere analyses samenhang vertoonden met andere variabelen in het bestand. Op basis van deze kriteria kwamen we tot de variabelen:

TON: eerste schoolkeuze na LO,

PRE: skore op schooltoets,

ADV: advies onderwijzer,

OPV: opleiding vader,

OPM: opleiding moeder,

BVA: beroep vader.

De kodering van deze zes variabelen in beide bestanden is in het algemeen verschillend, en (ernstiger) de gebruikte indikatoren zijn voor BVA en PRE nogal verschillend (zie De Jong e.a., 1981). Door middel van herkodering hebben we in sommige gevallen geprobeerd de kategorieen van de variabelen in beide bestanden zo vergelijkbaar mogelijk te maken. De kategorieen zijn als volgt:

TON: 1: VGLO
2: LBO
3: ULO/MAVO
4: VHMO
5: onbekend

PRE: 1: laag
2: minder laag
3: gemiddeld
4: beter dan gemiddeld
5: hoog
6: onbekend

ADV: 1: VGLO
2: LBO
3: ULO/MAVO
4: VHMO
5: onbekend

OPV: 1: LO
2: LBO
3: ULO
4: MBO
5: VHMO
6: HBO
7: onbekend

OPM: 1: LO
2: LBO
3: ULO
4: MBO
5: VHMO
6: HBO
7: onbekend

BVA: 1: arbeider
2: boer
3: middenstander
4: lagere employé
5: middelbare employé
6: hogere employé
7: ontbrekend

Behalve de vergelijking van SMVO en VJTJ heeft dit paper nog een aantal andere bedoelingen. We willen de konklusies van Van Herpen en Smulders (1980), die op grond van tabellaire analyse op de komplete kohorten aanzienlijke verschillen vinden tussen 64/65 en 77/78, proberen te rijmen met die van De Jong e.a. (1981), die op grond van korrelarionele analyse alleen maar kleine verschillen vinden. We proberen met onze analyse ook enige tekortkomingen te illustreren van korrelarionele/ kausale analyse toegeapst bij dit en soortgelijk onderzoek. Deze tekortkomingen zijn niet noodzakelijkerwijs ernstig, maar nopen wel tot een zekere voorzichtigheid bij het interpreteren van de resultaten en het trekken van de konklusies.

2: Marginalen

De univariate marginalen (percentages) staan in tabel 1. We merken hierbij op dat in de SMVO-steekproef slechts één individu naar het VGLO ging, en dat we de skore van dit individu op TON van VGLO in MAVO veranderd hebben. De technieken die we willen gebruiken hebben weliswaar geen last van individuen met unieke patronen, maar om statistische redenen is het beter ze te verwijderen. Wat direkt opvalt in tabel 1 (grafisch weergegeven in figuur 1) is een verschuiving naar de hogere kategorieen. In figuur 1 staan de kumulatieve percentages, berekend over de niet-ontbrekende kategoriën. Merk op dat de interpretatie van deze resultaten niet noodzakelijk is dat iedereen sinds 1965 beter, slimmer, en rijker geworden is. In de eerste plaats hebben we gezien dat de indicatoren voor BVA en PRE verschillen in de twee kohorten, in de

tweede plaats is er vooral in SMVO bewust sprake van een grote selektieve uitval (ongeveer 20%) doordat we alle individuen met BVA onbekend weggelaten hebben. Uit eerdere analyses (bv Gifi, 1980, 256-258) blijkt dat deze groep van uitvallers uit de laagste sociaal economische milieus afkomstig is, en de slechtste schoolprestaties heeft (gemiddeld genomen natuurlijk). Voor meer informatie over deze groep verwijzen we naar Van Herpen en Smulders (1980, 121-122). "Het verschil in omvang van de groep 'overig/onbekend' wordt ten dele verklaard door de verschillende codering van de sociale beroepsgroep in de twee cohorten, ten dele ook doordat in de tussenliggende periode het aantal echtscheidingen per jaar is verdrievoudigd, het aantal invaliditeitsuitkeringen meer dan verdubbeld is en het aantal werkloosheidsuitkeringen bijna acht keer zo groot geworden is." (lc, p 122). Evenals Van Herpen en Smulders en als De Jong, Dronkers, en Saris hebben we de groep weggelaten, hoewel dit dus ongetwijfeld tot belangrijke vertekeningen aanleiding geeft, waarvan we de eerste al tegengekomen zijn. Een korrektie voor deze weglatingen zou wel eens tot de konklusie kunnen leiden dat iedereen sinds 1965 slechter, dommer, en armer geworden is. Zoals bekend wordt er zowel op het CBS als door leden van deze werkgroep apart onderzoek voorgesteld of gedaan naar deze groep. Dat is hard nodig, want deze verschuiving zou wel eens de meest belangrijke verandering in een vergelijking van de twee cohorten kunnen zijn. Alle overige konklusies uit de analyses moeten daarom met een fikse korrel zout genomen worden, ze gaan er in feite van uit dat er in Nederland geen werklozen, gescheiden vrouwen, internaatskinderen, en arbeidsongeschikten bestaan.

3: Kruistabellen

In tabel 2 geven we de kruistabellen voor alle paren variabelen. Dit vooral voor de volledigheid, deze kruistabellen kunnen nuttig zijn bij eventuele verdere analyse. Het is moeilijk om konklusies te trekken

uit deze ruwe tabellen, we zouden op diverse manieren kunnen percenteren, en diverse associatiematen kunnen berekenen, maar deze weg willen we hier niet volgen. We zullen de kruistabellen gebruiken voor het berekenen van korrelatiekoefficienten.

4: Methodologisch intermezzo

Er is weinig twijfel mogelijk dat algemene log-lineaire technieken voor dit soort gegevens ideaal zijn. Ze berusten op een minimum aan aannamen, en maken het mogelijk op eenvoudige wijze statistische informatie over de stabiliteit van de oplossingen te berekenen. In ons geval is echter het aantal cellen in de zes-dimensionale tabel gelijk aan 5 x 6 x 5 x 7 x 7 x 7 = 51430, terwijl de steekproef-grootte tussen de 1000 en de 2000 ligt. Er zijn dus tenminste 50000 lege cellen, en de asymptotische theorie waarop log-lineaire modellen gebaseerd zijn gaat volkomen de mist in. Met name is het onmogelijk hogere orde interakties ook maar enigszins betrouwbaar te schatten. We beperken ons daarom tot bivariate technieken, dat wil zeggen technieken die alleen maar gebruik maken van de informatie in de kruitabellen (bivariate marginalen) in tabel 2. In het geval dat hogere dan eerste orde interakties verdwijnen, veroorzaakt dit geen verlies van informatie. Dat we bivariate technieken gebruiken wil dus niet zeggen dat we veronderstellen dat hogere orde interakties niet bestaan, we bestuderen ze alleen niet omdat het op basis van het materiaal dat we hebben niet op verantwoorde wijze kan gebeuren.

In de tweede plaats gebruiken we onze kruistabellen om korrelaties te berekenen. Ook hierbij moeten enkele kanttekeningen geplaatst worden. We kunnen korrelaties berekenen als we een of ander skorings-systeem voor de kategorieen van een variabele kiezen. Omdat we dat op veel manieren kunnen doen, kunnen we veel soorten korrelaties uitrekenen, die helemaal niet hoeven overeen te stemmen. In principe

kan ieder skoringssyteem, hoe idioot ook, gebruikt worden, en kunnen we de berekende korrelaties verantwoord statistisch bestuderen. Dit is een versie van de gouden regel van de data analyse: alles mag, maar de keuzen die men maakt hebben wel invloed op de konklusies die men kan trekken. Over het algemeen is de korrelatiekoefficient een maat van samenhang tussen twee variabelen die goed geinterpreteerd kan worden wanneer de regressie lineair en homoscedastisch is, en die op een bekende schaal varieert wanneer de verdeling van de variabelen bivariaat normaal is. Alweer geldt: we hoeven geen normaalverdelingen aan te nemen om statistisch verantwoorde uitspraken over korrelatiekoefficienten te doen, maar als we geen normaliteit aannemen dan moeten we de korrelatiekoefficient anders interpreteren als we gewend zijn, en moeten we andere statistische technieken gebruiken dan we gewend zijn.

Laten we daarom eens aannemen dat de verdelingen in tabel 2 steekproeven zijn uit gediskretiseerde bivariate normaalverdelingen. De Jong e.a. (1981) nemen aanzienlijk meer aan, zij nemen in feite ook nog aan dat de diskretisatiepunten lineair zijn met de kategorienummers. Daardoor kunnen zij korrelatiekoefficienten berekenen door de scores 1, 2, 3, ... te gebruiken voor ieder van de variabelen. Alles mag, maar de aanname van lineariteit is in deze kontekst wel héél moeilijk vol te houden. En is bovendien onnodig (zie verderop). En is bovendien pertinent onjuist (zie de univariate marginalen in tabel 1). Merk overigens op dat de aanname dat we een steekproef hebben uit een gediskretiseerde normaalverdeling impliceert dat er geen hogere orde interakties zijn (behalve missschien door diskretisatie-effekten), en dat we eigenlijk geinteresseerd zijn in de onderliggende kontinue normaalverdeelde variabelen.

Hoe schatten we korrelatiekoefficienten in gediskretiseerde normaal-verdelingen zonder de lineariteitsaanname te maken. Daar zijn een groot aantal technieken voor beschikbaar. Voor de soort van toepassing

die De Jong e.a. op het oog hebben (schatten van de parameters van een kausaal model) ligt het voor de hand een variant van de maximum likelihood methode toe te passen die kategoriegrenzen, korrelaties, en padcoefficienten tegelijktijd schat. Hoewel het schrijven van een algemeen programma langs deze lijnen een grote klus is (het is in feite LISREL met nog iets eromheen) is het schrijven van een specifiek ad-hoc programma voor een bepaalde dataset, een bepaald model, en een bepaald aantal variabelen binnen het kader van een groot projekt een zeer kleine investering. In APL bijvoorbeeld kan men zo'n programma in één dag schrijven en korrigeren. Konklusie: gebruik nooit LISREL op dit soort gegevens, de statistische informatie die eruit komt is niet waardevol. Met een beetje extra inspanning kan dit verholpen worden. Tweede konklusie: een aangepaste maximum likelihood techniek zal in het algemeen verschillende schattingen van de korrelaties voor verschillende modellen opleveren. Derde konklusie: zelfs zo'n aangepaste techniek blijft parametrisch, dat wil zeggen levert alleen goed interpreteerbare resultaten op wanneer de gegevens inderdaad gediskretiseerd normaal zijn.

Als we noch de lineariteitsaanname, noch de normaliteitsaanname willen maken, dan blijft er nog een klasse technieken over die gebruikt kunnen worden om korrelatiekoefficienten te berekenen. Deze technieken gaan uit van een bepaalde funktie f(R), met R de korrelatiematrix, en kiezen de skores voor de kategorieën op zo'n manier dat f(R) zo groot mogelijk wordt. Een voorbeeld verduidelijkt dit misschien: een mogelijke keuze voor f(R) is de multipele korrelatie van de eerste variabele met de overige variabelen. Kies de skores (of: transformeer de variabelen) op zo'n manier dat deze multipele korrelatie zo groot mogelijk wordt. Een andere f(R) is de grootste eigenwaarde van de korrelatiematrix. We kunnen ook zo transformeren dat deze grootste eigenwaarde zo groot mogelijk wordt (d.w.z. de korrelatiematrix wordt zo één-dimensionaal mogelijk).

Verschillende keuzen voor f(R) leveren verschillende transformaties van de variabelen, en daardoor ook verschillende korrelaties tussen de getransformeerde variabelen op (zoals bij de maximum likihood methode verschillende modellen voor de korrelaties tot verschillende schattingen van die korrelaties leidden). In dit stuk bekijken we de transformaties in meer detail die we vinden als we de multipele korrelatie tussen TON en de overige vijf variabelen als uitgangspunt kiezen. In de eerder genoemde CBS-presentatie gebruikten we als kriterium de grootste eigenwaarde en als alternatief kriterium de grootste kanonische korrelatie tussen TON + PRE + ADV enerzijds en OPV + OPM + BVA anderzijds. In het geval van de gediskretiseerde normaalverdeling geldt voor al deze technieken dat ze konsistente schattingen van de populatie-korrelaties opleveren, dus als de steekproef maar groot genoeg is en de diskretisatie maar fijn genoeg, dan vinden we de uniek gedefinieerde goede korrelaties terug. Dit geldt ook voor de maximum likelihood methode die we eerder kort besproken. De f(R)-optimalisatie technieken leveren echter ook interpreteerbare resultaten op als er geen sprake is van binormaliteit, we vinden immers een bovengrens voor alle mogelijke multipele korrelaties die we kunnen berekenen met alle mogelijke skoringssystemen voor de kategorieen. Het voordeel van de maximum likelihood methode is dat in het geval dat er inderdaad sprake is van gediskretiseerde binormaliteit de schatters grotere stabiliteit zullen hebben dan de f(R)-schatters (ze zijn dan 'efficient'). Genoeg gepraat, het wordt weer tijd voor een aantal getallen.

5: Korrelaties

In tabel 3 staan de korrelaties zoals berekend door De Jong e.a. Zoals boven besproken is deze keuze van korrelaties nogal willekeurig, en moeilijk te verdedigen. We gebruiken tabel 3 dan ook uitsluitend voor vergelijking met later volgende tabellen. Met behulp van een

globale chi-kwadraat toets vinden De Jong e.a. dat tabel 3a (VJTJ) en tabel 3b (SMVO) niet signifikant van elkaar verschillen. We tekenen hierbij aan dat een dergelijke globale toets weinig onderscheidend vermogen heeft, en dat de aannamen waarop de chi-kwadraat verdeling gebaseerd is in deze toepassing zeker niet opgaan. De resultaten geven eerder aanleiding om op te merken dat ADV en OPM in het SMVO-cohort belangrijker variabelen zijn dan in het VJTJ-cohort. Met name de toename van het belang van ADV is ook al door Van Herpen en Smulders gekonstateerd.

In tabel 5 staan de korrelatiematrices die we gevonden hebben door transformaties zo te kiezen dat de multipele korrelatiekoefficient zo groot mogelijk wordt. Als we tabel 3 en tabel 5 vergelijken, moeten we bedenken dat de multipele korrelatiekoefficient zo groot mogelijk wordt als we de korrelaties tussen TON en de vijf prediktoren zo groot mogelijk maken, en als we de korrelaties tussen de vijf prediktoren onderling zo klein mogelijk maken. Dat is wat de techniek probeert, en het lukt aardig. De multipele korrelaties zijn vrij aanzienlijk omhoog gegaan, ADV is nog duidelijker de belangrijkste prediktor geworden, en ADV is belangrijker geworden als we VJTJ met SMVO vergelijken. Er is geen verschil meer tussen OPM in VJTJ en SMVO, het lijkt erop dat bij deze korrelatiematrix BVA in VJTJ belangrijker is dan in SMVO (maar deze variabele is in beide cohorten dan ook anders gemeten).

Optimaliseren van de multipele korrelatiekoefficient heeft als bijkomend voordeel dat de techniek ook de optimale transformaties van de variabelen oplevert, en deze kunnen nu ook voor de twee kohorten vergeleken worden. We doen dit in figuur 2. In figuur 2 merken we op dat de transformaties voor de twee cohorten grofweg overeenkomen, met als belangrijkste verschil dat de SMVO transformaties meer naar rechts liggen. Dit wordt verklaard door de afwijkende verdelingen

in figuur 1. Een tweede verschuiving (te zien vooral in ADV) is dat LBO in SMVO lager getransformeerd wordt dan in VJTJ. Een effekt dat constant blijft, is de lage transformatie van MBO en HBO bij OPM, en de relatief hoge transformatie van ULO en VHMO. Vanwege de lage gewichten van OPM in de regressie is dit een weinig belangrijk effekt, maar de stabiliteit tussen cohorten is nogal opvallend. Opvallend is ook de ongeveer gelijke transformatie van de vier laagste BVA-categorieën bij SMVO. Merk ook op dat de transformaties het meest verschillen bij BVA en OPV, en dit zijn de variabelen die het meest te lijden hebben van de selektie op BVA in onze SMVO-steekproef.

In de eerder genoemde CBS-voordracht hebben we ook transformaties afgeleid uit HOMALS (maksimaliseer de eerste eigenwaarde) en een andere versie van CANALS (die de eerste kanonische korrelatie maksimaliseert). De transformaties van TON, ADV, en PRE zijn in alle analyses min of meer hetzelfde, in het algemeen is voor alle variabelen de HOMALS-transformatie gladder en meer regelmatig. Voor OPV, OPM, BVA zijn ze zelfs veel regelmatiger, we kunnen hier grofweg zeggen dat CANALS de onregelmatigheden in de HOMALS transformaties (zoals de plaats van MBO bij OPM of LBO bij ADV) als het ware uitvergroot. We geven de plots hier niet, voor belangstellenden zijn ze ter inzage.

De konklusies tot zover uit onze analyses geven we hier kort weer. Het belangrijkste effekt is ongetwijfeld de verschuiving in figuur 1 en 2, die een artefakt is van onze selektie van individuen. Een tweede belangrijk effekt is dat het belang van ADV toegenomen is, meer mensen houden zich dus aan het advies of VO-scholen letten tegenwoordig meer op ADV dan op PRE dan in de VJTJ-tijd. Er is een indikatie dat een LBO-advies bij SMVO negatiever is dan bij VJTJ (mensen houden zich relatief meer aan LBO-adviezen, ze worden minder vaak gegeven), bovendien lijkt er iets vreemds aan de hand met de MBO en HBO kategorieen van OPM, waarvoor we vooralsnog geen

interpretatie hebben. De belangrijkste methodologische konklusie is dat het zinloos is te praten over de korrelatiematrix van de variabelen, er zijn er oneindig veel, en ze kunnen behoorlijk verschillen. Keuze uit de korrelatiematrices is alleen mogelijk op basis van een model, en dit model moet realistisch zijn, of op basis van een kriterium, en dit kriterium moet interpreteerbaar zijn.

6: Stabiliteit

Het zou kunnen dat onze oplossingen voor de transformaties en de regressie-statististieken sterk afhankelijk zijn van steekproef-fluktuaties. Om dit na te gaan gebruiken we de bootstrap (Efron, 1979, Gifi, 1981, p 326-336). Het principe van de bootstrap is gemakkelijk uit te leggen: als we een steekproef van grootte N hebben, trekken we uit deze steekproef M nieuwe steekproeven van grootte N, met teruglegging, en we passen onze techniek toe op deze M nieuwe steekproeven. Het gemiddelde van de resultaten uit de analyse van de M bootstrap-samples kan gebruik worden voor bias-korrektie, de variantie kan gebruikt worden om de variantie van de resultaten te schatten, dat wil zeggen hun stabiliteit onder onafhankelijke replikaties. Als illustratie staat in tabel 7 de CANALS analyse op 5 bootstrap samples uit het VJTJ-sample, met ter vergelijking in de eerste kolom de oorspronkelijk VJTJ-resultaten. De multipele korrelaties blijken zeer stabiel, de regressiegewichten zijn ook stabiel, de transformaties van TON, PRE, ADV zijn redelijk stabiel, de transformaties van OPV, OPM, BVA zijn nogal instabiel, met name de extreme skores, die dikwijls overeenkomen met matig gevulde kategorieën. Deze stabiliteitsanalyse maakt ogenblikkelijk duidelijk dat er geen konklusie te trekken valt uit het MBO-HBO effekt op OPM, en dat meer algemeen er geen konklusies te trekken zijn uit de transformaties van OPV, OPM, BVA, behalve

met de grootst mogelijke voorzichtigheid.

7: Representatie van korrelatiematrices

We willen in deze laatste paragraaf illustreren, dat het probleem van de modelkeuze verre van triviaal is. Anders gezegd: naast het probleem om een korrelatiematrix uit de vele mogelijke korrelatiematrices te kiezen, is er ook nog het probleem de korrelatiematrix inzichtelijk weer te geven via een van de vele mogelijke kausale modellen. Dit probleem geldt vanzelfsprekend voor zowel VJTJ als voor SMVO, we illustreren het daarom met één enkele korrelatiematrix, de CANALS-korrelaties van VJTJ uit tabel 5a. We beperken ons hierbij tot volledige (juist geidentificeerde) modellen, dat wil zeggen modellen die de korrelatiematrix exact reconstrueren.

Multipele regressie op zichzelf is het eerste voorbeeld van een volledig kausaal model. We geven in dit stuk geen pijldiagrammen (die worden nogal vol voor volledige modellen), maar wel de strukturele vergelijkingen. Voor multipele regressie is dit de enkele vergelijking

$$TON = \beta_1 PRE + \beta_2 ADV + \beta_3 OPV + \beta_4 OPM + \beta_5 BVA + RES,$$

waarbij we aannemen dat RES ongekorreleerd is met de vijf exogene variabelen, die onderling wel gekorreleerd zijn. De schattingen van de beta's hebben we al berekend, die staan in de eerste kolom van tabel 6, de schatting van de variantie van RES is $1 - R^2$, dat wil zeggen .30.

We bekijken nu een tweede model, eveneens volledig. De variabelen zijn ingedeeld in drie groepen. De eerste groep is (OPV,OPM,BVA), deze variabelen zijn exogeen. De tweede groep is (PRE, ADV), en de derde is TON. We nemen aan

$$PRE = \alpha_{11} OPV + \alpha_{12} OPM + \alpha_{13} BVA + RES_1,$$

$$ADV = \alpha_{21} OPV + \alpha_{22} OPM + \alpha_{23} BVA + RES_2,$$

$TON = \beta_1 PRE + \beta_2 ADV + \beta_3 OPV + \beta_4 OPM + \beta_5 BVA + RES_3$,
waarbij de residuen RES_1, RES_2, RES_3 ongekorreleerd zijn met OPV, OPM, BVA. RES_3 is bovendien ongekorreleerd met RES_1 en RES_2, maar RES_1 en RES_2 mogen onderling wel korreleren. Schattingen van de parameters van dit model staan in tabel 8, we geven niet alle schattingen omdat de schattingen van de beta's hetzelfde zijn als in het multipele regressie model, terwijl de variantie van RES_3 weer gelijk is aan $1 - R^2$. We kunnen zeggen dat dit tweede model een nadere specifikatie van het eerste is, met een bepaalde decompositie van de matrix van korrelaties tussen de exogene variabelen, het model is echter nog steeds geheel tautologisch en voegt niets aan het eerste model toe, behalve de mogelijkheid tot een wat preciesere interpretatie. Opmerkelijk is verder dat de korrelatie tussen RES_1 en RES_2 gelijk is aan .64. Dit is rijkelijk hoog, het toont aan dat de in eerdere analyses gerezen vraag naar kausale prioriteit van PRE of ADV of naar het bestaan van reciproke interaktie tussen de twee variabelen empirisch niet te beantwoorden is.

In een derde volledig model voeren we een latente variabele LAT in, die exogeen is. We nemen aan

$TON = \alpha_{11} OPV + \alpha_{12} OPM + \alpha_{13} BVA + \gamma_1 LAT + RES_1$,
$PRE = \alpha_{21} OPV + \alpha_{22} OPM + \alpha_{23} BVA + \gamma_2 LAT + RES_2$,
$ADV = \alpha_{31} OPV + \alpha_{32} OPM + \alpha_{33} BVA + \gamma_3 LAT + RES_3$.

De residuen RES_1, RES_2, RES_3 zijn ongekorreleerd met OPV, OPM, BVA, LAT. De latente variabele LAT is ongekorreleerd met OPV, OPM, BVA. Schattingen van de parameters staan in tabel 9. Opmerkelijk is dat LAT tussen de 50% en de 60% van de variantie in TON, PRE, en ADV 'verklaart', en dat terwijl LAT ongekorreleerd is met OPV, OPM, BVA. Als LAT dat deel van de genotypische intelligentie is, dat orthogonaal staat op OPV, OPM, BVA, dat volgt uit dit model dat genotypische intelligentie als geheel zeker 80% van de variantie in schoolsukses 'verklaart'. Aan de andere kant kan LAT ook best 'puur geluk' zijn, onafhankelijk van OPV, OPM,

en BVA, en verantwoordelijk voor 60% van de variantie in schoolsukses. Dit illustreert het onbetwistbare nut van latente variabelen voor vindingrijke sociaal wetenschappelijke onderzoekers.

In het laatste model (van het Hauser-Goldberger-Jöreskog type) is LAT endogeen geworden. We nemen aan

$$TON = \alpha_{11}OPV + \alpha_{12}OPM + \alpha_{13}BVA + \gamma_1 LAT + RES_1,$$

$$PRE = \alpha_{21}OPV + \alpha_{22}OPM + \alpha_{23}BVA + \gamma_2 LAT + RES_2,$$

$$ADV = \alpha_{31}OPV + \alpha_{32}OPM + \alpha_{33}BVA + \gamma_3 LAT + RES_3.$$

Dit stuk is hetzelfde als bij het vorige model. Maar bovendien nemen we aan

$$LAT = \eta_1 OPV + \eta_2 OPM + \eta_3 BVA + RES_4.$$

Dit model heeft te veel parameters. Om het te identificeren nemen we aan dat RES_4 variantie gelijk aan één heeft, en dat we uit alle mogelijke oplossingen diegene kiezen waarvoor de drie bij drie matrix met de alpha's zo klein mogelijk is (in kleinste kwadraten zin). In tabel 10 staan de parameters. Duidelijk is dat de alpha's zo klein zijn dat we ze gevoeglijk weg kunnen laten. LAT (nu te identificeren met 'schoolvorderingen' of 'VHMO-geschiktheid') verklaart weer 60% van de variantie in TON, PRE, ADV, terwijl van LAT zelf nog geen 15% verklaart wordt door OPV, OPM, BVA samen. Alweer die genotypisch vastgelegde intelligentie ? Of misschien lichaamslengte. Dit laatste model heeft veel te maken met het model van De Jong, Dronkers, en Saris, de mysterieuze latente variabele speelt dezelfde rol, en die rol is om variantie uit de residuen te halen zodat de 'verklaarde' variantie toeneemt.

De boodschap van deze laatste paragraaf is duidelijk, en heeft niets te maken met vergelijking van VJTJ en SMVO. We hebben alle vier onze modellen ook op de SMVO-korrelaties losgelaten, met relatief weinig verschil, omdat de korrelaties op zichzelf weinig verschillen. Dit laatste resultaat blijft opmerkelijk: ondanks de grote verschillen

in de marginalen en ondanks het feit dat verschillende variabelen in de beide cohorten verschillend gedfinieerd zijn, blijven de korrelaties opmerkelijk konstant. Hetzelfde geldt ook voor de optimale transformaties, voor zover ze stabiel zijn. Hiervoor is geen goede verklaring, althans wij weten er geen een. In eerste instantie zou men vermoeden dat de moeilijkheden bij konstruktie van vergelijkbare steekproeven er toe leiden dat er geen behoorlijke vergelijking van de cohorten mogelijk is. We zijn dan ook niet verbaasd over de resultaten van Van Herpen en Smulders, die uit de tabellen redelijk aanzienlijke verschuivingen konstateren. Dat onvergelijkbaarheid tot verschillen leidt is invoelbaar. Dat onvergelijkbaarheid op korrelationeel en dus op 'kausaal' niveau niet voor systematische verschillen zorgt is wat moeilijker te begrijpen.

Literatuur

U. de Jong, J. Dronkers, W.E. Saris
Veranderingen in de schoolloopbanen tussen 1965 en 1977/8.
Konsept artikel, 1981.

L.W. van Herpen, R.H.M. Smulders
Sociale beroepsgroep en schoolkeuze.
C.B.S. Select 1, 1980, p 117-133

J. de Leeuw, I. Stoop
Secondaire analyse 'Van Jaar tot Jaar' met behulp van niet-lineaire multivariate technieken.
Van achteren naar voren, 1979, 118-158.

A. Gifi
Niet-lineaire multivariate analyse
Afd. Datatheorie FSW/RUL, 1980

A.Gifi
Non-linear multivariate analysis.
Afd. Datatheorie FSW/RUL, 1981.

B. Efron
Bootstrap methods: another look at the jackknife.
Ann. Statist., 7, 1979, 1-26.

TON	VJTJ	SMVO
1	8.8	0.0
2	38.6	25.9
3	33.8	37.4
4	18.8	36.3
5	0.0	0.5

PRE	VJTJ	SMVO
1	9.6	5.1
2	29.8	20.6
3	36.2	30.3
4	19.6	23.4
5	4.8	6.2
6	0.0	14.4

ADV	VJTJ	SMVO
1	12.0	0.9
2	40.8	22.8
3	29.1	37.8
4	16.4	36.5
5	1.7	2.0

OPV	VJTJ	SMVO
1	40.1	23.0
2	37.5	24.7
3	8.2	4.9
4	4.8	25.8
5	3.1	3.3
6	5.3	16.5
7	1.0	1.8

OPM	VJTJ	SMVO
1	65.8	39.4
2	20.0	24.6
3	7.3	10.6
4	2.6	12.2
5	2.0	1.7
6	1.7	5.9
7	0.6	5.6

BVA	VJTJ	SMVO
1	41.1	36.5
2	10.9	5.9
3	12.9	8.9
4	9.3	14.4
5	18.4	21.8
6	5.7	12.5
7	1.8	0.0

tabel 1: univariate marginalen (percentages)

	1	2	3	4	5	6
1	30	66	55	6	0	0
2	136	335	196	23	0	0
3	6	126	312	153	8	0
4	0	6	84	168	78	0
5	0	0	0	0	0	0

	1	2	3	4	5	6
1	0	0	0	0	0	0
2	68	177	76	16	0	56
3	9	121	258	101	1	77
4	1	13	122	238	93	85
5	0	2	4	1	0	0

tabel 2a: TON(rijen) maal PRE(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5
1	84	50	15	3	5
2	87	542	43	2	16
3	38	131	383	47	6
4	5	6	80	242	3
5	0	0	0	0	0

	1	2	3	4	5
1	0	0	0	0	0
2	4	315	58	2	14
3	5	26	445	83	8
4	4	4	66	469	9
5	0	1	5	1	0

tabel 2b: TON(rijen) maal ADV(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5	6	7
1	83	55	9	6	2	2	0
2	374	265	28	10	4	5	4
3	209	257	54	33	15	29	8
4	51	94	55	37	35	59	5
5	0	0	0	0	0	0	0

	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	162	116	12	76	3	11	13
3	121	151	32	167	19	66	11
4	67	106	29	147	28	172	3
5	0	2	2	2	0	1	0

tabel 2c: TON(rijen) maal OPV(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5	6	7
1	112	30	8	6	1	0	0
2	540	117	16	8	3	3	3
3	384	136	44	17	9	10	5
4	141	75	62	16	22	17	3
5	0	0	0	0	0	0	0

	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	229	86	17	27	1	2	31
3	214	160	64	66	10	18	35
4	152	127	79	91	15	69	19
5	4	0	1	2	0	0	0

tabel 2d: TON(rijen) maal OPM(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5	6	7
1	76	16	24	18	17	0	6
2	397	84	77	50	65	5	12
3	214	68	83	70	133	31	6
4	48	26	46	28	114	66	8
5	0	0	0	0	0	0	0

	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	214	25	50	49	41	41	0
3	225	35	49	80	122	56	0
4	115	29	36	87	166	119	0
5	1	0	0	3	2	1	0

tabel 2e: TON(rijen) maal BVA(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5
1	45	117	3	0	7
2	107	339	71	5	11
3	53	249	273	63	9
4	9	24	166	150	1
5	0	0	8	76	2
6	0	0	0	0	0

	1	2	3	4	5
1	0	60	16	0	2
2	2	157	130	10	14
3	3	59	272	120	6
4	0	16	81	254	5
5	0	0	2	92	0
6	8	54	73	79	4

tabel 2f: PRE(rijen) maal ADV(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5	6	7
1	109	56	1	3	0	1	2
2	250	213	35	12	9	11	3
3	272	229	57	33	14	34	8
4	74	143	43	28	23	35	4
5	12	30	10	10	10	14	0
6	0	0	0	0	0	0	0

	1	2	3	4	5	6	7
1	41	16	2	18	0	0	1
2	103	87	13	69	10	24	7
3	88	127	23	126	12	73	11
4	60	74	17	104	13	86	2
5	7	14	4	21	8	39	1
6	51	57	16	54	7	28	5

tabel 2g: PRE(rijen) maal OPV(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5	6	7
1	143	21	3	2	0	2	1
2	395	91	25	9	5	2	6
3	421	138	44	19	12	12	1
4	183	88	42	13	15	7	2
5	35	20	16	4	3	7	1
6	0	0	0	0	0	0	0

	1	2	3	4	5	6	7
1	52	13	2	2	1	0	8
2	148	82	31	22	3	5	22
3	172	125	52	57	9	22	23
4	104	87	37	65	5	38	20
5	18	24	14	13	5	19	1
6	105	42	25	27	3	5	11

tabel 2h: PRE(rijen) maal OPM(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5	6	7
1	100	28	18	15	8	2	1
2	269	60	68	51	61	14	10
3	264	67	88	55	131	27	15
4	91	35	46	34	100	39	5
5	11	4	10	11	29	20	1
6	0	0	0	0	0	0	0

	1	2	3	4	5	6	7
1	44	4	10	10	7	3	0
2	155	17	32	43	42	24	0
3	165	25	41	65	113	51	0
4	100	26	25	52	95	58	0
5	13	3	3	14	30	31	0
6	78	14	24	35	44	23	0

tabel 2i: PRE(rijen) maal BVA(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5	6	7
1	113	74	12	4	3	6	2
2	365	288	34	18	6	13	5
3	178	208	54	33	16	26	6
4	49	92	43	29	29	48	4
5	12	9	3	2	2	2	0

	1	2	3	4	5	6	7
1	0	4	0	5	0	3	1
2	141	106	9	66	4	9	11
3	137	143	33	162	14	75	10
4	65	117	30	146	31	162	4
5	7	5	3	13	1	1	1

tabel 2j: ADV(rijen) maal OPV(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5	6	7
1	149	43	9	4	3	3	3
2	557	128	20	10	6	4	4
3	315	124	50	17	7	7	1
4	135	61	47	13	19	16	3
5	21	2	4	3	0	0	0

	1	2	3	4	5	6	7
1	6	4	2	1	0	0	0
2	208	75	17	18	2	1	25
3	221	150	57	74	8	23	41
4	152	133	82	89	16	65	18
5	12	11	3	4	0	0	1

tabel 2k: ADV(rijen) maal OPM(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5	6	7
1	105	21	24	20	26	10	8
2	380	87	97	56	85	12	12
3	190	59	70	58	115	24	5
4	48	27	35	30	93	54	7
5	12	0	4	2	10	2	0

	1	2	3	4	5	6	7
1	2	1	1	1	2	6	0
2	195	23	43	38	38	9	0
3	222	31	53	80	131	57	0
4	125	34	34	93	154	115	0
5	11	0	4	7	6	3	0

tabel 2l: ADV(rijen) maal BVA(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5	6	7
1	589	100	13	7	2	2	4
2	433	177	35	14	4	5	3
3	69	35	26	11	3	2	0
4	34	22	19	4	5	2	0
5	12	13	16	2	9	4	0
6	28	10	20	9	12	14	2
7	12	1	1	0	0	1	2

	1	2	3	4	5	6	7
1	212	80	15	18	2	1	22
2	162	111	33	39	2	5	23
3	33	18	14	6	1	0	3
4	129	110	45	69	4	13	22
5	11	9	14	7	1	7	1
6	45	38	36	46	16	63	6
7	7	7	4	1	0	0	8

tabel 2m: OPV(rijen) maal OPM(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5	6	7
1	452	66	84	57	44	3	11
2	255	108	108	61	120	7	12
3	19	1	21	27	65	9	4
4	1	14	8	12	41	8	2
5	0	0	7	4	27	16	2
6	3	0	2	2	29	58	1
7	5	5	0	3	3	1	0

	1	2	3	4	5	6	7
1	239	11	31	43	17	9	0
2	171	43	42	54	51	14	0
3	14	1	4	37	17	2	0
4	110	28	45	56	113	40	0
5	0	1	0	13	19	17	0
6	9	3	9	15	109	105	0
7	12	2	4	1	5	3	0

tabel 2n: OPV(rijen) maal BVA(kolommen) voor VJTJ(links) en SMVO(rechts).

	1	2	3	4	5	6	7
1	583	133	146	107	158	27	23
2	121	52	53	35	82	10	5
3	17	4	17	18	59	14	1
4	6	2	5	5	18	9	2
5	1	1	3	1	6	23	0
6	3	0	3	0	6	17	1
7	4	2	3	0	0	2	0

	1	2	3	4	5	6	7
1	291	25	57	92	90	44	0
2	144	33	40	45	86	25	0
3	32	5	12	32	48	32	0
4	46	7	17	29	55	32	0
5	2	0	2	3	8	11	0
6	4	4	1	8	33	39	0
7	36	15	6	10	11	7	0

tabel 2o: OPM(rijen) maal BVA(kolommen) voor VJTJ(links) en SMVO(rechts).

	TON	PRE	ADV	OPV	OPM	BVA
TON	1.00					
PRE	.63	1.00				
ADV	.74	.68	1.00			
OPV	.37	.30	.34	1.00		
OPM	.24	.19	.24	.45	1.00	
BVA	.31	.28	.29	.60	.35	1.00

tabel 3a: korrelaties VJTJ uit De Jong e.a.

	TON	PRE	ADV	OPV	OPM	BVA
TON	1.00					
PRE	.63	1.00				
ADV	.80	.67	1.00			
OPV	.37	.27	.36	1.00		
OPM	.36	.21	.31	.48	1.00	
BVA	.33	.26	.33	.59	.40	1.00

tabel 3b: korrelaties SMVO uit De Jong e.a.

	VJTJ	SMVO
PRE	.22	.17
ADV	.55	.64
OPV	.09	.04
OPM	.01	.10
BVA	.03	.01
R^2	.59	.67

tabel 4: beta-gewichten en multipele korrelaties uit tabel 3.

	TON	PRE	ADV	OPV	OPM	BVA
TON	1.00					
PRE	.68	1.00				
ADV	.80	.69	1.00			
OPV	.44	.30	.36	1.00		
OPM	.27	.18	.22	.34	1.00	
BVA	.42	.31	.32	.56	.27	1.00

tabel 5a: korrelaties VJTJ uit CANALS.

	TON	PRE	ADV	OPV	OPM	BVA
TON	1.00					
PRE	.61	1.00				
ADV	.85	.59	1.00			
OPV	.38	.27	.35	1.00		
OPM	.23	.14	.22	.26	1.00	
BVA	.28	.20	.25	.47	.21	1.00

tabel 5b: korrelaties SMVO uit CANALS.

	VJTJ	SMVO
PRE	.21	.15
ADV	.57	.72
OPV	.10	.06
OPM	.05	.03
BVA	.10	.04
R^2	.70	.74

tabel 6: beta-gewichten en multipele korrelaties uit CANALS.

		VJTJ	BS1	BS2	BS3	BS4	BS5
TON:	VGLO:	-0.74	-0.79	-0.76	-0.75	-0.67	-0.88
	LBO :	-0.93	-0.89	-0.89	-0.83	-0.88	-0.98
	ULO :	0.24	0.20	0.23	0.13	0.12	0.32
	VHMO:	1.79	1.88	1.84	1.92	1.83	1.64
PRE:	1 :	-1.39	-1.32	-1.49	-1.32	-1.40	-1.40
	2 :	-0.89	-0.86	-0.89	-0.82	-0.90	-0.86
	3 :	0.09	-0.01	0.21	0.06	0.13	0.04
	4 :	1.27	1.42	1.31	1.30	1.14	1.24
	5 :	2.34	2.26	2.02	2.43	2.58	2.34
ADV:	VGLO:	-0.69	-0.72	-0.64	-0.70	-0.62	-0.95
	LBO :	-0.86	-0.80	-0.83	-0.76	-0.79	-0.89
	ULO :	0.45	0.39	0.41	0.25	0.25	0.60
	VHMO:	1.85	1.99	1.94	2.04	1.93	1.60
OPV:	LO :	-0.64	-0.83	-0.49	-0.49	-0.58	-0.71
	LBO :	-0.38	-0.09	-0.33	-0.55	-0.40	-0.19
	ULO :	1.23	1.51	1.12	1.61	1.56	0.23
	MBO :	1.72	0.31	0.65	1.15	2.20	2.60
	VHMO:	2.20	2.20	0.25	2.61	3.02	2.27
	HBO :	2.42	2.68	3.93	2.22	1.65	1.98
OPM:	LO :	-0.53	-0.41	-0.37	-0.52	-0.37	-0.47
	LBO :	0.36	0.01	-0.28	0.03	0.61	0.08
	ULO :	2.72	0.53	2.97	2.36	2.93	2.39
	MBO :	0.32	4.11	0.93	2.32	-2.35	-0.21
	VHMO:	1.83	3.34	2.11	2.85	-0.27	3.43
	HBO :	0.09	2.34	-0.09	1.95	-1.21	1.23
BVA:	1 :	-0.90	-0.88	-0.83	-0.88	-0.76	-0.93
	2 :	-0.35	-0.03	0.01	-0.37	-0.48	-0.27
	3 :	0.96	1.08	0.63	1.73	1.20	0.71
	4 :	-0.40	-0.60	-1.04	-0.12	-0.17	-0.40
	5 :	1.00	1.30	1.54	0.83	0.16	1.41
	6 :	2.43	1.54	1.48	1.34	2.94	1.79
β	PRE :	.24	.26	.21	.22	.20	.25
	ADV :	.69	.68	.70	.72	.74	.68
	OPV :	.13	.13	.14	.12	.12	.10
	OPM :	.06	.06	.08	.07	.07	.07
	BVA :	.11	.11	.13	.07	.12	.14
Rmult	:	.84	.84	.85	.86	.86	.85

tabel 7: CANALS analyse op VJTJ en op vijf bootstrap-samples uit VJTJ.

	PRE	ADV
PRE	.87	.55
ADV	.55	.84

	OPV	OPM	BVA
PRE	.17	.07	.20
ADV	.25	.09	.16

tabel 8: kausale keten met drie blokken, varianties en covarianties van RES_1 en RES_2 (links) en padkoefficienten (rechts).

	γ	ε^2
TON	.74	.20
PRE	.68	.42
ADV	.81	.18

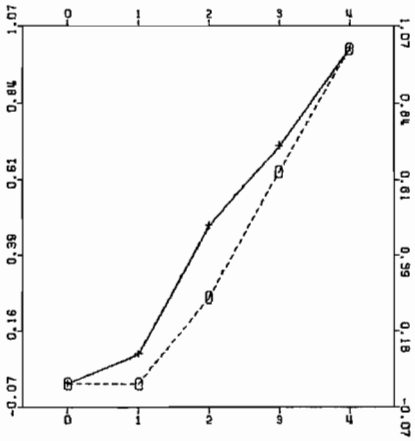
	OPV	OPM	BVA
TON	.27	.11	.23
PRE	.17	.07	.20
ADV	.25	.09	.16

tabel 9: model met exogene latente variabele, padkoefficienten van latente variabele en varianties van residuen (links) en padkoefficienten van manifeste variabelen (rechts).

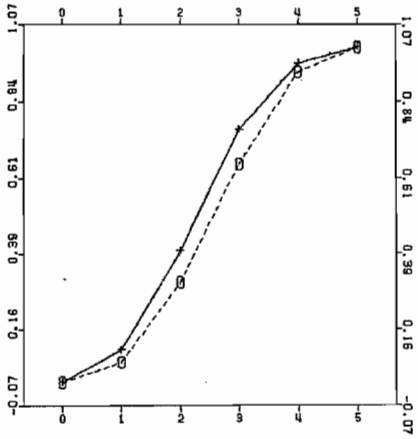
	η	γ	ε^2
TON	.31	.74	.20
PRE	.13	.68	.42
ADV	.26	.81	.18

	OPV	OPM	BVA
TON	.04	.02	.04
PRE	-.04	-.01	.02
ADV	-.01	-.01	-.06

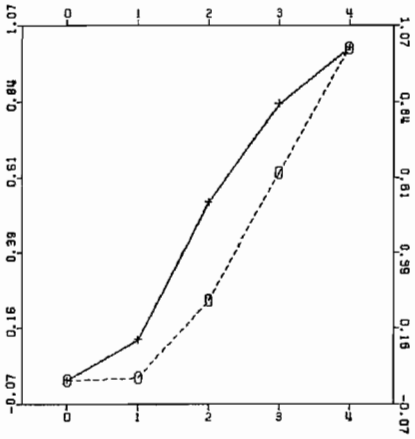
tabel 10: model met endogene latente variabele, padkoefficienten naar en van latente variabele en varianties van residuen (links) en padkoefficienten van manifeste variabelen (rechts).



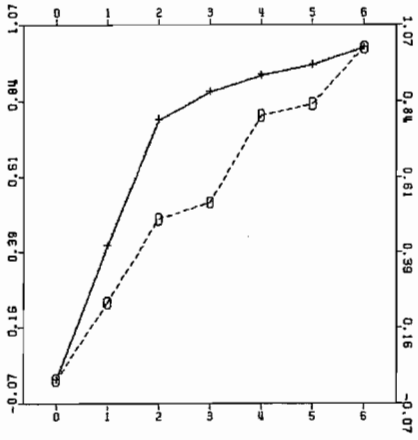
TON



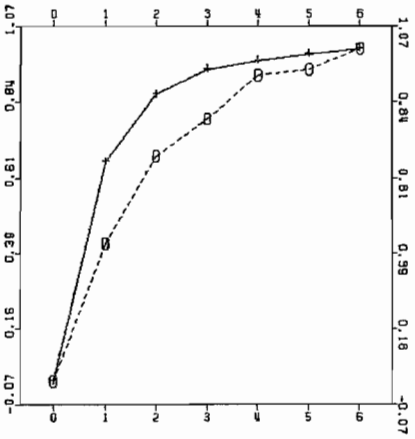
PRE



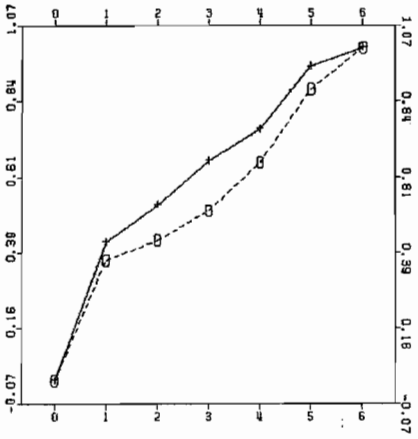
ADV



OPV

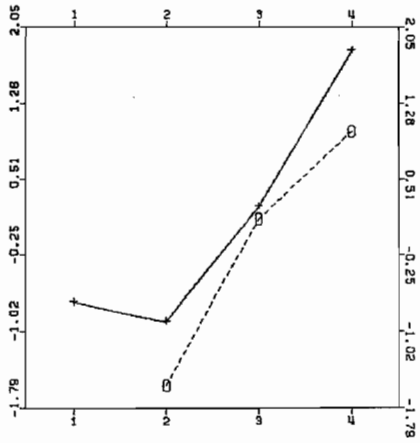


OPM

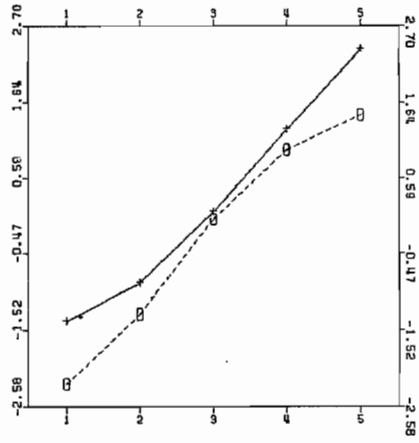


BVA

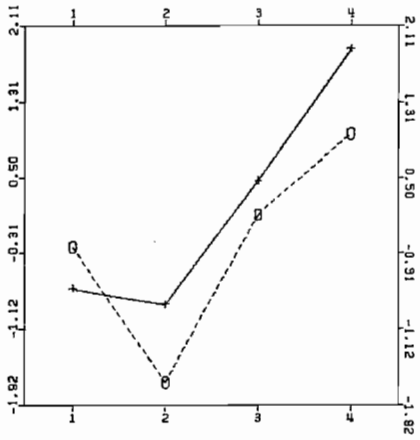
figuur 1: kumulatieve verdelingen VJTJ(+) en SMVO(0).



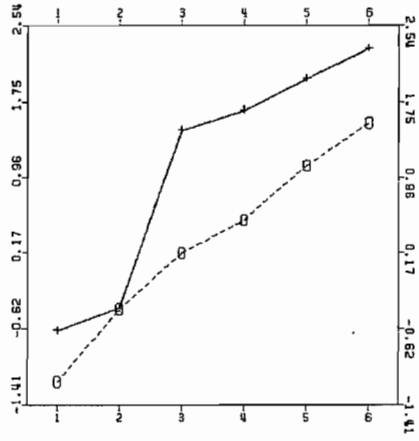
TON



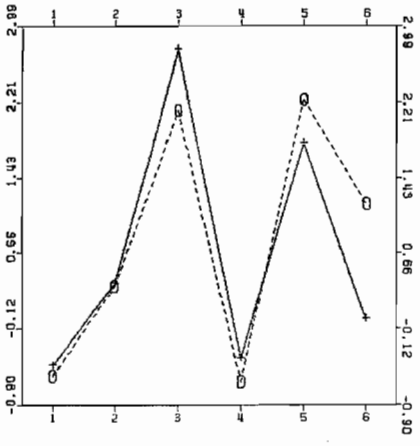
PRE



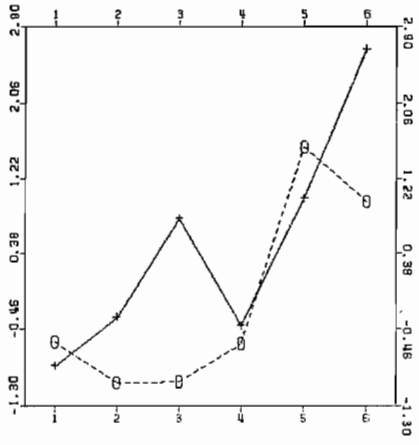
ADV



OPV



OPM



BVA

figuur 2: CANALS-transformaties VJTJ(+) en SMVO(O).