Dr. A. Hazewinkel

Werkclassificatie: een wetenschappelijk instrument?

Uitgegeven voor
het Nederlands Instituut voor Praeventieve Geneeskunde TNO
door
J. B. Wolters Groningen

Werkclassificatie: een wetenschappelijk instrument?

Werkclassificatie:
een wetenschappelijk instrument?

EEN THEORETISCHE EN EXPERIMENTELE STUDIE

DR. A. HAZEWINKEL

Uitgegeven voor het Nederlands Instituut voor
Praeventieve Geneeskunde TNO, te Leiden

J. B. WOLTERS · GRONINGEN · 1967

Als dissertatie verscheen dit boekwerk onder de titel:

De genormaliseerde methode van werkclassificatie als meetinstrument, een theoretische en experimentele studie.

Opgedragen aan
mijn Moeder
de nagedachtenis van mijn Vader

Inhoud

HOOFDSTUK I

Inleiding

1.1 TOEPASSING VAN WETENSCHAPPELIJKE KENNIS IN DE PRAKTIJK

In een gesprek met vrienden werd mij de vraag gesteld: "Is het niet gevaarlijk als een psycholoog een oordeel geeft over de geschiktheid van een sollicitant? Een afwijzende beslissing kan immers niet alleen een carrière-mogelijkheid bij dit bepaalde bedrijf uitsluiten, het kan ook – en hier ligt het werkelijke gevaar – de beoordeelde een verkeerde indruk geven over zichzelf, waardoor hijzelf zekere ontwikkelingsmogelijkheden uit gaat sluiten".

De vragenstelster ging er – terecht – van uit dat psychologen in hun uitspraken feilbaar zijn, zelfs wanneer deze uitspraken gebaseerd zijn op een grondig onderzoek met behulp van langs wetenschappelijke weg ontwikkelde methoden.

Het door haar gesignaleerde probleem heeft een duidelijke parallel in de toepassing van een werkclassificatie-methode. Ook daar kan men zich afvragen of het niet "gevaarlijk" of op zijn minst onverantwoord is om uitspraken te doen omtrent loonverhoudingen, uitspraken die belangrijke loon-politieke konsekwenties kunnen hebben en die – evenals het oordeel van een psycholoog – feilbaar zijn.

Van deze overeenkomst tussen het psychologisch advieswerk en de toepassing van werkclassificatie-methoden zal hier gebruik gemaakt worden bij een poging om dit onderzoek met betrekking tot de Genormaliseerde Methode van Werkclassificatie in te leiden en te rechtvaardigen. Eerst zal daartoe een antwoord gegeven worden op de vraag, een aanklacht bijna, naar het gevaar van psychologische adviezen.

Het probleem heeft minstens twee aspecten:

- de psycholoog kan falen;
- door zijn wetenschappelijke pretentie wordt zijn feilbaarheid op gevaarlijke wijze onderschat.

Het eerste probleem is betrekkelijk eenvoudig op te lossen. De psycholoog kan falen, maar de personeelsfunctionaris, chef of wie ook bij afwezigheid van een psycholoog sollicitanten beoordeelt en selecteert, kan ook falen. Het gaat er maar om wie er minder vaak naast is, of beter gezegd: welke selectieprocedure de beste resultaten levert, een procedure waarbij een psycholoog ingeschakeld is of de traditionele. "Beslissend is dan alleen of men de vraag beter dan een ander kan beantwoorden" (SNIJDERS 1965, p. 566). Stelt men de vraag op deze wijze dan is het probleem gereduceerd tot een vraagstelling die in principe oplosbaar is door wetenschappelijk onderzoek. Dit onderzoek is bij selectie voor functies die weinig voorkomen, en dus voor de meeste hogere functies, in de praktijk niet of zeer moeilijk uitvoerbaar, maar dat doet aan het principe niets af.

Het tweede probleem is veel gecompliceerder en men kan zich hiervan niet afmaken door er op te wijzen dat het gevaar slechts in de pretentie schuilt en dat het dus afwezig is als deze pretentie van wetenschappelijkheid en dus onfeilbaarheid (een duidelijk 'non sequitur' trouwens) vermeden wordt. KOUWER (1955) heeft dit met nadruk naar voren gebracht als "een gewetensprobleem van de toegepaste psychologie". Het helpt vaak niet of de psycholoog waarschuwt voor overschatting van zijn kunde: het publiek wil zekerheid en zoekt die bij de psycholoog of deze wil of niet. Zoals KOUWER (op. cit., p. 10) het uitdrukt: "men wil de stem van het noodlot horen en het noodlot twijfelt niet". Vanuit de eigen onzekerheid zoekt men bij de ander onfeilbaarheid. Het opvallende hierbij is, dat het juist de term 'wetenschap' of 'wetenschappelijk' is, die deze associatie van absolute zekerheid en onfeilbaarheid oproept. Het prestige van de wetenschap (speciaal de natuurwetenschappen) samen met het wegvallen van oude religieuze zekerheden zal hier wel mede schuld aan zijn. Hoe dit ook zij: de wetenschappelijke pretentie van een methode of van een advies is in veel gevallen gevaarlijk, niet omdat ten onrechte een wetenschappelijk karakter gesuggereerd zou worden, maar omdat het begrip wetenschap misverstaan wordt. Dit gevaar kan dan ook alleen bestreden worden door een begripsverheldering gevolgd door een duidelijk stellen van waar de verantwoordelijkheden liggen. Wetenschap immers impliceert geenszins zekerheid en onfeilbaarheid. Gevoelens van zekerheid en onfeilbaarheid vinden juist in de wetenschappelijke instelling hun grootste vijand, want deze instelling is voor alles een kritische. Men zie in dit verband LIJFTOGT (1966, Hfdst. VI, paragraaf d) en ook NAGEL (1955, p. 345) die er op wijst dat de opvatting

van zekerheid als een kenmerk van wetenschap verouderd is: "The long history of science and philosophy is in large measure the history of the progressive emancipation of men's minds from the theory of self-evident truths and from the postulate of complete certainty as the mark of scientific knowledge."

Het zoeken naar zekerheid heeft in de wetenschap geleid tot een steeds weer opnieuw ondergraven van oude waarheden, waardoor ook de nieuwste kennis een voorlopig karakter krijgt. Alleen de historische continuïteit van de wetenschapsontwikkeling, gevolg van het feit dat wetenschapsbeoefenaars behalve kritisch ook constructief zijn, garandeert dat de oude verworvenheden niet als afval weggeworpen worden, maar meestal na een zekere relativering als bouwstenen in een groter gebouw blijven meefunctioneren. De paradox blijft echter bestaan: wetenschap zoekt naar zekerheid door te proberen om alle zekerheden systematisch en konsekwent-kritisch aan te tasten, of met andere woorden: "Een goed opgezet wetenschappelijk toetsingsonderzoek is in feite *altijd op falsificatie gericht*". (DE GROOT 1961, p. 105; cursivering DE GROOT)

Ook al kan deze stelling tot eenzijdigheid leiden, zij is noodzakelijk om te voorkomen dat 'wetenschappelijk' gebruikt wordt in de zin van 'absoluut zeker' en 'onbetwijfelbaar' hetgeen leidt tot schromelijke overschatting van de rol van deze wetenschap. Dit gevaar is meer dan alleen maar theoretisch wanneer het gaat om de gedragswetenschappen.

Een vooraanstaand psycholoog als SIEGEL was zelfs van mening dat: "Our generation of psychologists needs to stay in the laboratory, pursuing systematic and theoretically based programs of experimentation, and leaving the solution of immediate practical problems in the hands of responsible men experienced with them" (SIEGEL 1964, p. 21). Men hoeft het niet geheel met hem eens te zijn om nochtans te zien, dat toepassing van de gedragswetenschappen op praktijksituaties een hachelijke zaak kan zijn en dat uiterste voorzichtigheid hierbij geboden is. Die voorzichtigheid verplicht in de eerste plaats tot een relevantie-onderzoek: is, wat de wetenschap ons te bieden heeft, relevant voor ons probleem? SIEGEL's antwoord was, als we bovenstaande weergave van zijn mening letterlijk mogen nemen, in alle gevallen: "neen". Weinigen zullen zover met hem mee willen gaan. Hier wordt gepleit voor een handelwijze waar wél de oplossing van praktijkproblemen overgelaten wordt aan de verantwoordelijke personen die ervaring op het desbetreffende gebied bezitten, maar waar deze personen tevens de beschikking krijgen over wetenschappelijk verkregen

informatie om hun eigen ervaring aan te vullen en zo nodig te corrigeren.

Een belangrijk punt blijft dan nog de verdeling der verantwoordelijkheden. KOUWER (1955) wijst er terecht op dat wanneer iemand advies vraagt in verband met een te nemen beslissing, dit niet mag leiden tot het afschuiven van de verantwoordelijkheid *voor de uiteindelijke keuze* op de adviseur, i.c. de psycholoog. Dit neemt echter niet weg dat de psycholoog ook een verantwoordelijkheid heeft die hij op zijn beurt niet mag afschuiven op zijn cliënt: de verantwoordelijkheid voor het ontstaan van een helder inzicht bij de adviesvrager in de betekenis èn de betrekkelijkheid van de hem verschafte wetenschappelijke gegevens, een betrekkelijkheid die zowel ligt in het voorlopige karakter van alle wetenschappelijke kennis, als ook in de nooit volledige toepasbaarheid van algemene kennis op individuele gevallen.

Deze toepasbaarheid is in sommige gevallen nauwelijks een probleem, terwijl ze in andere situaties het advieswerk bijna onmogelijk maakt. Als een psycholoog voor een bedrijf een selectieprocedure heeft ontwikkeld en gevalideerd (door middel van kruisvalidatie!) en daarna deze methode toepast op hetzelfde bedrijf is de toepasbaarheid van zijn methode redelijk gegarandeerd. Elk nieuw individu waarover beslist moet worden is, voor het bedrijf, vergelijkbaar met de reeds behandelde individuen. CRONBACH en GLESER (1957) spreken in zo'n geval van "institutional decisions", waarbij zij de nadruk leggen op het feit dat alle beslissingen binnen hetzelfde selectieproces steeds weer gebaseerd worden op hetzelfde waardesysteem.

Zij stellen dit tegenover de "individual decisions", waar de waardeoverwegingen van geval tot geval verschillen en waar men dientengevolge niet alle individuen meer over één kam kan scheren. Door hun besliskundige preoccupatie hebben zij vooral de nadruk gelegd op de toepasbaarheid van waardesystemen, die in individuele beslissingssituaties tot moeilijkheden leiden. Er is echter ook nog een verschil in toepasbaarheid van onderzoekresultaten die geheel los staat van waardeoverwegingen. Gedoeld wordt op het feit dat een sociaal-wetenschappelijk onderzoek meestal uitgevoerd wordt op een steekproef van personen. Wat gevonden wordt heeft slechts betrekking op de populatie waar de steekproef uit getrokken werd en niet op individuele gevallen. Dit is een speciaal geval van het bekende adagium: "scientia non est individuorum" (ALLPORT 1937), hetgeen zoveel zeggen wil als: wetenschappelijke uitspraken hebben nooit betrekking op, en zijn dus ook nooit volledig toepasbaar op, individuele gevallen. Deze

schijnbaar eenvoudige stelling raakt aan een problematiek van wetenschaps-theoretische aard, die in het kader van deze studie niet verder behandeld kan worden. De lezer zij verwezen naar de uitgebreide literatuur die op dit gebied verschenen is (bijvoorbeeld ALLPORT 1937, SNYGG en COMBS 1949, Hfdst. XV, DE GROOT 1954a, DE GROOT 1961, p. 363 e.v.). Wel zal hier iets gezegd worden over een veel voorkomende misvatting met betrekking tot het begrip 'waarschijnlijkheid' (of 'kans').[1] Met name in de testpsychologie wordt dit begrip vaak onvoldoende zorgvuldig gebruikt doordat de bovengenoemde problematiek niet gezien of tenminste niet dóórzien wordt. Het betreft hier bij uitstek praktische situaties waar een zekere verheldering van de begrippen zeker geen overbodige luxe is.

Een score op een (als voorspeller gebruikte) test of de uitkomst van een selectieprocedure heeft op zichzelf geen enkele betekenis. Betekenis ontleent ze pas aan vergelijking met andere gegevens, bijvoorbeeld met praktijkresultaten. De functie van een zogenaamd validatieonderzoek is nu juist om aan deze scores betekenis toe te kennen. Als het gaat om een voorspellende test zoals we die bij selectieprocedures gebruiken, wordt ernaar gestreefd om aan elke score of groep van scores een waarschijnlijkheid toe te kennen, die bijvoorbeeld aangeeft wat de kans is dat een individu met zo'n score, als hij door het bedrijf aangenomen wordt, als werknemer zal bevallen.

Deze waarschijnlijkheden hebben echter niet betrekking op individuen als zodanig, maar slechts op individuen *voor zover zij elementen zijn van populaties van personen* (vergelijk NAGEL 1955, p. 362). Wil men dit in feite uitermate abstracte waarschijnlijkheidsbegrip wat concreter maken, dan stelle men zich het volgende, uiteraard niet werkelijk uitvoerbare, experiment voor. Definieer de populatie waarop het onderzoek betrekking zal hebben: bijvoorbeeld alle arbeiders die tussen 1960 en 1970 bij bedrijf A gesolliciteerd hebben of dat nog zullen doen. Neem de test in kwestie af telkens als een arbeider uit deze populatie solliciteert. Kies alle personen onder hen die een score hebben van bijvoorbeeld hoger dan S. Stel dit zijn er 100 in aantal. Ga na hoeveel van deze 100 in bedrijf A slagen. Stel dit zijn er 60. Dan kan men zeggen dat de waarschijnlijkheid dat iemand uit deze populatie, die een score hoger dan S heeft, zal slagen, 60% is. Het is deze kans die in

[1] De term 'waarschijnlijkheid' wordt op vele manieren geïnterpreteerd. Hier wordt deze term uitsluitend gebruikt om de frekwentie-opvatting van het waarschijnlijkheidsbegrip aan te duiden. Voor andere opvattingen zie NAGEL (1955, p. 359 e.v.).

het psychologisch onderzoek geschat wordt. Het is duidelijk dat deze 60% iets zegt over de groep van 100 arbeiders met een score boven S. Het zegt niets over een willekeurig individu in die groep. "No meaning can be attached to any expression which, taken literally, assigns a probability to a single individual as having a specified property" (NAGEL 1955, p. 365). Het is niet juist om te concluderen uit het voorgaande dat voor ieder van deze 100 arbeiders geldt dat zij een kans hebben van 60% om te slagen; voor elk van hen geldt immers dat zij òf slagen – en dan was hun 'kans' kennelijk 100% – òf niet slagen en dan was hun 'kans' kennelijk 0% (vergelijk ook ALLPORT, zoals geciteerd in SNYGG en COMBS 1949, p. 347).

Mocht deze redenering nog niet overtuigend zijn, dan stelle men zich voor dat een ander zo'n experiment doet met een iets anders gedefinieerde populatie, bijvoorbeeld alle arbeiders die tussen 1960 en 1965 solliciteerden. Wat nu, als een arbeider in beide populatie-definities valt en als van de groep in de tweede populatie die boven S scoort er slechts 40% slagen? Heeft deze arbeider dan tegelijk 60% en 40% kans om te slagen? Inderdaad, want hij heeft deze kansen *alleen als lid van een populatie*, en hij is lid van deze twee – en oneindig veel andere – populaties. Als individu heeft hij echter geen kans in deze strikte, waarschijnlijkheids-theoretische betekenis van het woord.

Deze schijnbaar te theoretische redenering heeft belangrijke konsekwenties voor de praktijk van het advieswerk, konsekwenties die echter veel te weinig getrokken worden. Hoewel dit onderscheid door CRONBACH en GLESER (1957) niet expressis verbis gemaakt wordt en bovenstaande redenering bij hen dan ook niet terug te vinden is, hebben zij toch wel de konsekwenties hiervan getrokken, zij het in wat verhulde vorm. Zij schrijven over een student namelijk als "having confidence in his ability to take advantage of the one chance in ten which would give him a very good record", waarmee zij verklaren waarom een student soms het risico van een laag cijfer zal accepteren als er een kleine kans (10%) bestaat op een heel hoog cijfer. De hier genoemde 'ability' is niet gemeten en de student meent dus dat hij over extra informatie beschikt betreffende zijn persoonlijkheid (ability to take advantage of . . . etc.), waardoor hij ingedeeld kan worden in een subpopulatie waar de kans op succes, geassocieerd aan zijn testscore, veel hoger is dan één op tien. Het is daarom van het grootste belang om in adviesgesprekken, ten eerste, kansen van slagen te noemen en daarbij, ten tweede, een duidelijke populatie-definitie te geven, zodat de persoon die uiteindelijk de beslissing moet nemen, zelf kan uitmaken in

hoeverre hij zich bij dit beslissingsproces wil beschouwen als lid van deze populatie.

Wij hebben gezien dat er minstens twee aspecten aan het relevantieprobleem onderscheiden kunnen worden:

– is de onderzoekpopulatie relevant? Met andere woorden: in hoeverre is het individu bereid zichzelf te beschouwen als lid van deze populatie (voor zover het deze beslissing betreft)?

– is het waarde-systeem relevant? Met andere woorden: in hoeverre komt het waarde-systeem, waarop de beslissingsprocedure gebaseerd is, overeen met de voor het individu belangrijke waarde-overwegingen?

1.2 RELEVANTIE EN FUNCTIE VAN DE GENORMALISEERDE METHODE

Bij de Genormaliseerde Methode van Werkclassificatie[2] [3] kan men deze twee aspecten van het relevantieprobleem terugvinden. De graderingscijfers worden opgesteld op grond van vergelijkingen van functies. De feitelijke uitkomsten van deze vergelijkingen, de hoogte van de gegeven cijfers en zeker ook het belang van het gezichtspunt zoals dat onder meer in de spreiding tot uiting komt, zijn sterk afhankelijk van de gekozen populatie, dat wil zeggen van de groep functies waarmee men de gegradeerde functies in principe vergelijkbaar acht.

De keuze van gezichtspunten en de relatieve hoogte van de afweegfactoren zijn gebaseerd op waarde-overwegingen, of behoren dat althans te zijn. Gelden de gebruikte waarde-overwegingen voor alle (groepen van) functies op gelijke wijze? Dit is het waarde-aspect van de relevantie-vraag. De moeilijkheid is dat het waarde-aspect bij de GM onontwarbaar verweven is met het graderingsaspect, hetgeen ook voor het afzonderlijk beschouwen van de twee relevantie-vragen zijn konsekwenties heeft. Op deze verwarring komen wij in het tweede hoofdstuk nog uitvoerig terug. Hier willen wij slechts verwijzen naar wat LIJFTOGT (1966, p. 51) geschreven heeft over de onderlinge samenhang van graderingscijfers met afweegfactoren.

Verdiept men zich in het relevantieprobleem van de GM dan valt

[2] Een historisch overzicht van het ontstaan en de ontwikkeling van deze methode wordt gegeven door LIJFTOGT (1966). Technische gegevens betreffende de methode en haar toepassing vindt men in de twee publikaties van de Deskundigencommissie (1952 en 1959). De geschiedenis van deze Commissie wordt beschreven door LIJFTOGT (1966).

[3] De Genormaliseerde Methode wordt hier verder aangeduid met GM.

op dat het uitgangspunt van de G M is geweest dat men aan één populatie-definitie genoeg had. Buiten Nederland is zo'n standpunt, voor zover ik weet, nooit verdedigd. Waarschuwingen dat de toepassing van een methode steeds aangepast moet zijn aan de specifieke bedrijfssituatie vindt men daarentegen in overvloed. Om twee recente voorbeelden aan te halen: LIJFTOGT (1966, p. 89) schrijft heel duidelijk: "De vergelijkbaarheid der functies uit verschillende bedrijfstakken was echter een der zwakke punten uit de methodiek (van de G M)" en HUNDT (1965, p. 86) deelt mee dat een algemeen geldend voorstel voor de bepaling van afweegfactoren nooit gedaan is, "Dies ist aber auch gar nicht möglich, da die Festlegung der Gewichte und der Stufenwertverläufe von einer Vielzahl von Bestimmungsfaktoren, wie Betriebsstruktur, Produktionsrichtung, Arbeitsmarktlage, bisherige Lohnstruktur, Tradition usw. abhängt, . . .". Deze waarschuwingen dienen au sérieux genomen te worden, als men wenst te vermijden, dat de informatie die de G M levert als te absoluut opgevat wordt, doordat de grenzen van haar toepasbaarheid niet bekend zijn. Te absoluut wordt de G M ook opgevat op nog andere, veel ingrijpender, wijze. De keuze van de gezichtspunten wordt nergens discutabel gesteld, de graderingsschalen zijn op vrij willekeurige wijze samengesteld, vragen naar betrouwbaarheid worden niet gesteld. Voor een psycholoog liggen zulke vragen voor de hand en met betrekking tot de psychologische praktijk zijn ze vaak even moeilijk te beantwoorden. Dat is dan ook het motief voor de hier naar voren gebrachte vergelijking tussen psychologisch advieswerk en de toepassing van werkclassificatie. Deze vergelijking gaat natuurlijk voor een deel mank, maar als belangrijke punten van overeenkomst blijven, dat ook bij de toepassing van werkclassificatie het begrip 'wetenschappelijk' vaak gehanteerd wordt om zekerheid te suggereren (LIJFTOGT 1966, Hfdst. VI) en dat er een sterke behoefte aan zekerheid bestaat omtrent de 'juiste' loonverhoudingen.

Deze behoefte aan zekerheid behoeft geen verwondering te wekken. Loonvorming immers is een uiterst delicate zaak, waarin tegenstrijdige belangen en een veelheid van onduidelijke normen een in de praktijk nauwelijks te ontwarren rol spelen. *Niemand* weet op welke gronden werk hoog of laag beloond wordt of zelfs maar zou moeten worden: moet men inspanning belonen die aan de functie-beoefening voorafging (moeilijkheid en duur van opleiding) of financiële investering (kosten van opleiding) of is inspanning bij de uitoefening van de functie zelf het belangrijkst? Hoe kunnen we onderscheid maken tussen geestelijke belasting (verantwoordelijkheid, risico's), intellectuele belasting

(kennis), motorische eisen (handvaardigheid, houding) en hoe moeten we deze verschillende soorten van inspanning, als we ze al kunnen onderscheiden, ten opzichte van elkaar waarderen? Men kan met evenveel recht verdedigen dat de man die het meest onaangename werk doet het best betaald moet worden (compensatie-gedachte) als dat de man die het psychisch meest eisende werk verricht het hoogste loon moet ontvangen, ondanks het feit dat zijn werk meestal interessant is en dus in zichzelf zijn beloning met zich draagt. Er is zelfs gepleit voor een volledig uniforme beloning van iedere arbeid, bijvoorbeeld door BELLAMY (zie WIEGERSMA 1954, p. 84).

Tot nu toe zijn we er van uit gegaan dat het juist is om bij de loonbepaling slechts met de functie als zodanig rekening te houden. Er is dus even aangenomen dat RITTER (1958) gelijk heeft als hij zegt, dat de functie zelf, of, zoals hij het uitdrukt, de "intrinsieke waarde van de functie" een "belangrijke, zo niet de belangrijkste factor voor de vaststelling van de beloning is" (p. 196). Er zijn echter velen die er anders over denken en bijvoorbeeld aan economische factoren als vraag en aanbod niet alleen een belangrijke maar ook een rechtmatige rol toekennen in de loonvorming. Volgens WIEGERSMA (1954) is de GM opgesteld uitgaande van de gedachte dat de schaarste van eigenschappen die voor een functie-uitoefening nodig zijn, betaald moet worden. LIJFTOGT (1966, Hfdst. V) komt tot dezelfde conclusie wat betreft de overtuiging van de voorzitter van de Deskundigencommissie, GEVERS DEYNOOT.

Als andere factoren die het loon mede bepalen worden door WIEGERSMA (1954, p. 81-90) genoemd 'sociale status' en 'machtspositie'. Het is duidelijk dat al deze factoren loonbepalend werken al weet niemand precies hoe. Het is echter niet duidelijk of dit juist is of niet, of zelfs maar: of er belangengroepen zijn die dit juist vinden.

Het ontbreken van duidelijke normen heeft er onder andere toe geleid, dat men bij het zoeken naar methoden om meer bewust en systematisch het probleem van de loonverhoudingen te benaderen, zich steeds op het standpunt heeft gesteld dat het resultaat, van welke methode ook, nooit te veel van de bestaande verhoudingen mocht afwijken.[4] Het is in dit verband interessant dat de Deskundigencommissie

[4] Een extreem voorbeeld van een werkclassificatie-methode die van dit standpunt uitgaat, is de methode toegepast bij de Nederlandse Spoorwegen (VAN WESSEM 1952). Daar is namelijk op grond van multipele correlatierekening een aantal gezichtspunten gekozen, zodanig dat de overeenkomst met de bestaande loonstructuur maximaal werd.

(1952, p. 10) spreekt van een "verantwoorde rangorde" en niet van een rechtvaardige rangorde. Men zou kunnen zeggen, dat de methode ertoe dient om nieuwe functies zo in de bestaande structuur in te passen dat het zo min mogelijk weerstand oproept. HUNDT (1965, p. 86) zegt het als volgt: "Die sozialpsychologische Forderung für die Bestimmung der Gewichtung ist, dass die landesüblichen Einschätzungen und Meinungen über den Wert und die Entlöhnung der verschiedensten Tätigkeiten, . . ., respektiert werden".

Ook RITTER (1958, p. 196) laat zich in deze geest uit: "Welke motieven bij de waardering als relevant zullen worden aangemerkt en welke invloed zij op de waardering (dus op de beloning) zullen hebben, wordt niet door de enkeling bepaald, doch veeleer door de opvatting (het rechtsgevoel) van de gemeenschap". Het is wellicht in deze zin ook, dat TIFFIN (1951, p. 361) spreekt van "a fair solution". Wij moeten ons bij de wijze waarop men in feite rekening houdt met deze "landesüblichen Einschätzungen" en bij dit 'rechtsgevoel van de gemeenschap' niet te veel voorstellen. Het geeft een soort evenwichtstoestand aan waarbij de meesten zich voorlopig hebben neergelegd.

Het is zelfs niet onwaarschijnlijk dat de feitelijke functie van werkclassificatie eenzelfde is als die van stereotypen die, sociaal gezien, een stabiliserende invloed hebben. Werkclassificatie is, in deze zin opgevat, een differentiatie en rationalisatie van bestaande opvattingen omtrent loonverhoudingen. DE VISSER (1963, p. 1152) drukt het als volgt uit: "Een andere reden waarom taakwaarderingssystemen met 'succes' worden toegepast, schuilt in hun schijnobjectiviteit en schijnexactheid. Zo lang een bedrijf gelooft in de objectiviteit van het systeem, werkt het gedurende enige tijd bevredigend. Dogma's kunnen vals zijn, de zielerust die zij geven is soms echt". Zo'n methode zal goed werken, dat wil zeggen zijn stabiliserende functie goed kunnen uitoefenen, als hij in grote lijnen overeenkomt met de reeds bestaande loonstructuur. Deze overeenkomst wordt bereikt als aan de volgende voorwaarden is voldaan:

- Factoren die bij het tot stand komen van de bestaande structuur doorslaggevend waren, zijn op een of andere wijze in de methode verdisconteerd. Zo zal bij iedere goed werkende methode 'opleiding' op een of andere wijze (bijvoorbeeld via een gezichtspunt Kennis) vertegenwoordigd zijn en een belangrijke invloed uitoefenen.
- Factoren die in feite belangrijk zijn maar waarvan de invloed ethisch moeilijk te verdedigen valt, zijn verhuld of indirect in de methode opgenomen. Zulke factoren zijn bijvoorbeeld de door WIEGERSMA

(1954) genoemde 'sociale status' en 'machtspositie'. Deze factoren worden automatisch in de methode opgenomen als de bestaande loonstructuur als richtlijn voor het vaststellen van afweegfactoren (en graderingscijfers, LIJFTOGT 1966) dient.

De genoemde voorwaarden hebben betrekking op eigenschappen van de G M zoals deze is ontwikkeld en in voorschriften (Deskundigencommissie 1952 en 1959) is vastgelegd. Het is daarnaast zeer wel mogelijk dat er in de toepassing van de methode in de praktijk, door het gebrek aan exactheid, zoveel ruimte overblijft, dat de taakanalist naar een bestaande loonstructuur kan toewerken. Dat deze mogelijkheid niet alleen maar een theoretische is, zal uit de resultaten van het experimentele gedeelte van deze studie blijken.

Als, door welke oorzaak ook, de G M beschouwd wordt als een methode met enige wetenschappelijke pretentie, bestaat het gevaar dat haar mogelijkheden overschat worden. Er is hier in meer algemene zin gewezen op de betrekkelijkheid van 'zekerheid' en 'onfeilbaarheid' in de wetenschappelijke kennis. Deze studie betreffende de Genormaliseerde Methode van Werkclassificatie wil bovendien meer specifiek de 'zin en onzin' in de methode ontwarren. Daartoe wordt in de twee volgende hoofdstukken ingegaan op fundamentele problemen, zoals ook WIEGERSMA (1954) tot op zekere hoogte reeds gedaan heeft. Dit is een vooral theoretische benadering met vragen zoals: indien de graderingscijfers die per gezichtspunt gegeven worden, slechts rangordegetallen zijn, mag men deze dan vermenigvuldigen met afweegfactoren en optellen? En wat is dan het karakter, schaaltechnisch gesproken, van de einduitkomst?

Daarnaast is een eerste poging gedaan om de methode als praktisch gebruiksinstrument experimenteel te onderzoeken. Speciale aandacht is daarbij besteed aan de mogelijkheid dat de taakanalist toewerkt naar een einduitkomst, die slechts gedeeltelijk gebaseerd is op de functie-inhoud en gedeeltelijk op factoren die met functie-eisen als zodanig niets te maken hebben. Een dergelijk verschijnsel zou van belang zijn voor de interpretatie van de grote overeenkomst die in het algemeen geconstateerd wordt tussen werkclassificatie-resultaten en de bestaande loonstructuur. Het experiment dat met dit doel is uitgevoerd, wordt besproken in hoofdstuk IV. In een laatste hoofdstuk worden tenslotte de resultaten en bevindingen samengevat en afgesloten met een voorlopige conclusie.

HOOFDSTUK II

Meettheoretische kritiek

In dit hoofdstuk wordt eerst de vraag gesteld of schaaltechnische overwegingen voor de Genormaliseerde Methode van Werkclassificatie relevant zijn. Nadat deze in positieve zin beantwoord is – de Genormaliseerde Methode wordt opgevat als een meetinstrument, niet als een voorspeller – wordt een overzicht gegeven van de vraagstukken samenhangende met de verschillende typen van meetschalen, zoals vooral door STEVENS in psychologische kringen geïntroduceerd. Daarna worden de algemene beschouwingen toegespitst op het speciale geval van de Genormaliseerde Methode, opgevat als meetinstrument. Het hoofdstuk wordt afgesloten met een kritiek op de werkwijze van de Deskundigencommissie.

2.1 INLEIDING

Werkclassificatie is een methode om functies te ordenen met het doel de zo verkregen rangorde te gebruiken als hulpmiddel bij de loonvorming.

De problematiek van de Genormaliseerde Methode van Werkclassificatie heeft minstens twee belangrijke aspecten:
– *een praktisch loonpolitiek* aspect en
– *een theoretisch meettechnisch* aspect.

LIJFTOGT (1966) besteedt in zijn studie vooral aandacht aan het eerste aspect. Hij beschrijft de wordingsgeschiedenis en de verdere ontwikkeling en houdt zich meer in het algemeen bezig met een waardering van de methode als loonpolitiek instrument. Hier wordt de methode vrijwel uitsluitend bezien vanuit een strikt meettechnisch standpunt.

De Genormaliseerde Methode van Werkclassificatie is een variant op wat in de Amerikaanse literatuur bekend staat als een 'point rating system' of 'point evaluation system' (zie bijvoorbeeld SATTER 1949). Dat wil zeggen, dat een functie van verschillende gezichtspunten uit, afzonderlijk, beoordeeld wordt. Deze oordelen worden in cijfers, bij de G M graderingscijfers genoemd, uitgedrukt; deze worden vervolgens elk, al naar gelang van het gezichtspunt, met een 'afweegfactor' van

een voor het gezichtspunt constante grootte, vermenigvuldigd. De produkten worden tenslotte opgeteld om zo tot een totaalwaardering te komen.

Een oppervlakkige beschouwing van deze werkwijze werpt reeds vele vragen op. De belangrijkste zijn wel de vragen naar de schaaleigenschappen van de zo gekwantificeerde oordelen (ordinale schalen, intervalschalen of zelf ratioschalen?) en naar de rechtvaardiging van de keuze van gezichtspunten en bijbehorende afweegfactoren. Hierbij wordt dan nog opzettelijk afgezien van vergelijking van het 'point system' met andere methoden zoals de 'factor comparison method' (OTIS en LEUKART 1954) en tevens wordt hier aangenomen dat een lineaire combinatie van de gezichtspunten om verschillende redenen te verkiezen is boven niet-lineaire combinaties.

Wat betreft de schaaltechnische eigenschappen en hun konsekwenties voor de methode als meetinstrument, blijkt dat de meeste auteurs deze ofwel negeren ofwel zeer gemakkelijk over de problemen heenlopen. Zo lezen we bij OTIS en LEUKART (1954, p. 130): "... the differences between degrees must be equidistant either in an absolute (that is, arithmetic) or geometric progression", maar de auteurs vragen zich niet af of zij met hun methode ook inderdaad gelijke intervallen krijgen. Een onderzoeker als LAWSHE, die zeker acht "Studies in job evaluation" gepubliceerd heeft in de Journal of Applied Psychology negeert het probleem volkomen (zie bijvoorbeeld LAWSHE en SATTER 1944). SATTER (1949) geeft wel een beschrijving van een methode waarvan men kan verwachten dat ze tot intervalschaalwaarden leidt, maar noemt dit niet als argument voor de gebruikte methode (die van de paarsgewijze vergelijkingen). Er wordt ook geen statistische toets vermeld om na te gaan of het model wel bruikbaar is. Ook DE GROOT (1953, p. 405) vermeldt als voordelen van deze methode niet het ontstaan van een intervalschaal. Hij adviseert het opstellen van een *rangorde* op grond van paarsgewijze vergelijkingen.

Voor het merkwaardigste geval behoeven we echter niet zo ver van huis te gaan. In het rapport van de Deskundigencommissie voor Werkclassificatie[1] (1959, p. 25) lezen we: "Met behulp van graderingsschema's en voorbeelden van gradering kan op vrij nauwkeurige wijze een *rangorde* van functies met betrekking tot elk gezichtspunt worden vastgesteld" (cursivering van de schrijver). Dit wijst op een ordinale schaal en inderdaad – en terecht – vervolgt men met: "De aan een

[1] Deskundigencommissie wordt hier verder aangeduid met DC.

functie voor de verschillende gezichtspunten gegeven cijfers kunnen echter niet zonder meer worden opgeteld, aangezien het ongelijkwaardige grootheden betreft". Maar nu bederft men de zaak weer door te stellen dat optelling wel geoorloofd is na weging. Dit zou alleen dan het geval zijn als de schalen reeds intervalschalen waren. Een ordinale schaal wordt door een lineaire transformatie (vermenigvuldiging met een factor) geen intervalschaal!

Er zijn eigenlijk maar twee auteurs die met de kwestie van de schaaleigenschappen ernst maken; hun conclusies zijn echter met elkaar in strijd. WIEGERSMA (1954) komt op grond van zijn meettheoretische beschouwingen tot de conclusie dat de uitkomsten van de werkclassificatie-methode zonder betekenis zijn. Slechts een zeer globale indeling van functies is hiermee mogelijk; dit kan echter ook bereikt worden zonder de werkclassificatie en dan sneller en goedkoper.

DE GROOT (1954b) merkt naar aanleiding van deze conclusie op dat een meettheoretisch aanvechtbare methode zeer wel te verdedigen valt en zinvol kan zijn als deze maar aan een criterium getoetst kan worden. Het verschilpunt tussen deze twee opvattingen lijkt vooral hierin te schuilen dat WIEGERSMA twijfelt aan de mogelijkheid om ooit tot een goed criterium te komen en dat DE GROOT zo'n criterium wel denkt te kunnen ontwikkelen.

2.2 GM – VOORSPELLER OF MEETINSTRUMENT?

Het gaat hier in feite om een meer fundamenteel onderscheid. DEGROOT (1961, p. 261) onderscheidt, in zijn discussie van het begrip 'instrumentele utiliteit van een variabele', twee functies die variabelen kunnen hebben: een *voorspellende* en een *metende*. De voorspellende functie houdt in dat de variabele in strikt logische zin een vervangingsmiddel is, een surrogaat, voor een andere variabele waarvan men de waarde minder gemakkelijk of op dat moment nog in het geheel niet kan vaststellen. Die andere variabele wordt in de psychologische literatuur meestal het *criterium* genoemd. Als een variabele een metende functie heeft is dit onderscheid in voorspeller en criterium niet langer zinvol: de variabele is in zekere zin zijn eigen criterium. Naast deze (criterium-) variabele moet men dan nog wel, aldus DE GROOT, een 'begrip-zoals-bedoeld' onderscheiden, maar dat kan alleen in de fantasie de status van variabele bereiken. DE GROOT (1962) spreekt daarom van een "critère virtuel". De konsekwenties van dit onderscheid voor de validi-

teitsbepaling is een onderwerp op zichzelf; men zij hiervoor verwezen naar DE GROOT (1961, Hfdst. 8).

Het standpunt van WIEGERSMA (1954) is dat de werkclassificatie als criteriummaat moet dienen, dat wil zeggen een metende functie behoort te hebben. Het standpunt van DE GROOT blijkt duidelijk uit zijn commentaar op SATTER. SATTER stelt dat "the problem is not only one of *measuring* the criterion, but, in the first place, of defining one" (geciteerd in DE GROOT 1953, p. 407; zie onze literatuurlijst: SATTER 1949, p. 215). DE GROOT merkt naar aanleiding hiervan op: "Het belang van deze vraagstelling is in de eerste plaats, dat de *slechts benaderende, hypothetische waarde van de werkclassificatie* erin wordt geïmpliceerd; wat dus *toetsing aan een 'eigenlijker criterium'* noodzakelijk maakt" (cursivering DE GROOT). Voor hem heeft dus de werkclassificatie-methode een voorspellende functie. Hij beschouwt de vraag naar de (predictieve) validiteit van de methode dan ook als een zinvolle: hoe goed voorspelt de methode dat wat voorspeld moet worden, namelijk het 'eigenlijker criterium'? DE GROOT (1953, p. 408) zoekt dit in de richting van "de menselijke (op rechtsbewustzijn gefundeerde) beoordeling" – een formulering die hij ontleent aan VAN SUSANTE – en noemt de methode van paarsgewijze vergelijkingen (van functies in hun geheel) als een operationalisatie hiervan die "directer ons rechtsgevoel" weerspiegelt "en daardoor psychologisch aanvaardbaarder" is dan de GM.

Er zijn enkele bezwaren hiertegen in te brengen. Ten eerste is dit een zeer tijdrovende, arbeidsintensieve procedure. Dat is dan ook de reden waarom de GM als voorspeller van dit criterium, dat wil zeggen als praktisch bruikbaar vervangingsmiddel, moet dienen. Men zal echter niet zo snel geneigd zijn om het criterium voldoende frekwent te herzien in het licht van een veranderde sociaal-economische situatie. Ten tweede is een zodanige operationalisatie van het rechtsbewustzijn moeilijk in loononderhandelingen in te passen, terwijl dat toch de plaats is waar het waardesysteem, dat aan een werkclassificatie ten grondslag behoort te liggen, geformuleerd moet worden. Bovendien wordt het resultaat van de loononderhandelingen niet discutabel. Men kan niet goed discussiëren over een rangorde van een aantal sleutelfuncties en eerst recht niet over afweegfactoren die slechts ten doel hebben van de methode een goede voorspeller te maken. Dit laatste punt zal nog ter sprake komen. Tenslotte valt nauwelijks te verwachten dat een op deze wijze ontwikkeld criterium werkelijk nieuwe informatie geeft: de overeenkomst met de bestaande loonstructuur zal onge-

twijfeld zeer hoog zijn. DE GROOT (1953, p. 408) ziet dit niet als een bezwaar. Hij realiseert zich dat de praktijksituatie een "zeer twijfelachtig criterium" is, maar trekt een interessante analogie met de gang van zaken in de testpsychologie. Ook daar heeft men zich in het begin moeten behelpen met zeer onzuivere criteria als schoolsucces of beoordeling door een onderwijzer, maar langzamerhand heeft het begrip zich van de afhankelijkheid van deze criteria losgemaakt en is het eerder zo dat het oordeel van een onderwijzer aan de testresultaten getoetst kan worden.

Er zijn echter twee belangrijke verschillen tussen intelligentietests en de GM. Ten eerste kunnen testscores onafhankelijk van een criterium vastgesteld worden en op die wijze een eigen leven gaan leiden met een eigen ontwikkeling. Bij de toepassing van de GM komt men echter zelden tot resultaten die belangrijk van de bestaande loonverhoudingen afwijken: als dit zou gebeuren zou men al spoedig geneigd zijn om de methode af te keuren. Een tweede, meer fundamenteel, onderscheid is dat door middel van de GM *waarde-oordelen* gegeven worden. Zolang men de methode *als geheel* te veel als een voorspeller in de zin van een test beschouwt loopt men het gevaar dat dit belangrijke aspect van een werkclassificatiemethode uit het oog verloren wordt.

Hier wordt derhalve gepleit voor een duidelijke splitsing van het *waarderende* en het *metende* element in de GM. Zo'n splitsing is niet mogelijk als men de GM als voorspeller beschouwt. Het waarderende element is dan namelijk naar het (moeilijk voor discussie vatbare) criterium verschoven. Het hier voorgestelde alternatief voor de voorspeller-opvatting is de opvatting dat de GM een meetinstrument zou moeten zijn met expliciet gemaakte waarde-elementen. Deze beschouwingswijze is ook beter aan de perceptie van de gebruiker aangepast. De taakanalist en de werknemer proberen met de werkclassificatie tot een schatting van de waarde van een functie te komen en niet tot een voorspelling van de uitkomst van een andere (criterium-)meting.

De vraag doet zich voor, speciaal in het licht van het bovenstaande, hoe het komt dat meettheoretische bezwaren tegen de in de methode gebruikelijke rekenkundige bewerkingen zo weinig gehoord zijn. Men is in feite niet verder gekomen dan een zekere 'lip service' aan de meettheorie, zoals wel het duidelijkst blijkt uit de in paragraaf 2.1 aangehaalde passages van de DC. Men moet zich dan echter realiseren dat een aanloop tot een wetenschappelijke discussie over werkclassificatie reeds in het begin van de vijftiger jaren vastliep (vergelijk LIJFTOGT 1966, Hfdst. I), ondanks herhaalde pogingen van de zijde van

psychologen om deze op gang te brengen. Met name DE GROOT heeft hiervoor sterk geijverd, maar zou er pas een jaar of tien later in slagen om een onderzoek van de grond te krijgen. Ook door de voorzitter van de DC werd herhaaldelijk de mening uitgesproken dat wetenschappelijk onderzoek gewenst is (vergelijk LIJFTOGT 1966, Hfdst. VI); niettemin kwam er niets tot stand.

Ook internationaal bezien, en dan wel speciaal in de Amerikaanse literatuur, vindt men zeer weinig discussie over het probleem van het schaalkarakter der graderingscijfers. Wel is er een uitgebreide literatuur over toepassing van de methode, waarbij gebruik gemaakt werd van technieken als factoranalyse en multipele regressie ter bepaling van gezichtspunten en afweegfactoren (zie bijvoorbeeld LAWSHE en SATTER 1944, LAWSHE 1945, CHESLER 1949).

Het fundamentele vraagstuk dat hier aan de orde gesteld wordt en door de DC slechts schijnbaar werd opgelost, werd ook in de Amerikaanse literatuur hoogstens in een enkele volzin achteloos afgedaan (zie OTIS en LEUKART 1954) of geheel genegeerd.

Een belangrijke factor hierbij is dat in de periode, waarin nog wel van wetenschappelijke zijde daadwerkelijk aan de discussie werd deelgenomen, althans in psychologische kringen, nog weinig algemeen bekend was op het gebied van schaaltypen. Toen deze kennis vooral na de bekende publikaties van STEVENS (1946, 1951) in grotere kring toegankelijk werd, was de werkclassificatie al bijna geheel overgegaan in handen van praktijkmensen en zeker in Nederland op korte termijn niet voor wijziging vatbaar (LIJFTOGT 1966, Hfdst. VI).

In het volgende zal nu een poging gedaan worden om de G M alsnog te bezien vanuit meettheoretisch standpunt. Met opzet wordt hier gesproken van een poging, niet uit bescheidenheid, maar omdat, zoals zal blijken, de methode in feite zo weinig doorzichtig is dat een analyse die uitgaat van wat officieel bij de toepassing plaats zou vinden (bepaling van graderingscijfers, vermenigvuldiging met afweegfactoren en tenslotte berekening van het eindcijfer, T) voor de werkelijke toepassing van de methode in de praktijk weinig relevant is.

2.3 MEETSCHALEN

De vraag naar de betekenis van getallen in de psychologie, de vraag naar wat men met deze getallen mag doen en wat de zin is van de uitkomsten van de uitgevoerde bewerkingen, is al vroeg gesteld. THORN-

DIKE schreef reeds in de twintiger jaren: "How far it is proper to add, subtract, multiply, divide and compute ratios with the measures obtained (by existing instruments for measuring intellect) is not known . . ." (geciteerd in KOHNSTAMM 1942, p. 341).

In Nederland maakte KOHNSTAMM reeds in 1942 onderscheid tussen *eigenlijke metingen* waar het gaat om "werkelijke 'hoeveelheden' die dus uit onderling vergelijkbare en vervangbare delen bestaan" en *oneigenlijke metingen* waar het gaat om "intensiteiten, waarbij objectief slechts van rangorde gesproken kan worden" (op. cit., p. 334).

De bovengenoemde publikaties van STEVENS (1946 en 1951) luidden echter eerst goed een nieuw tijdperk van meettheoretische 'sophistication' in. In psychologische kringen behoren deze nog steeds tot de belangrijkste literatuurbronnen op dit gebied. Keer op keer wordt STEVENS geciteerd, bestreden of als uitgangspunt voor de discussie genomen als er sprake is van meetschalen en de konsekwenties daarvan voor "wat men met meetuitkomsten mag doen". (Men zie bijvoorbeeld JAHODA, DEUTSCH en COOK 1951, COOMBS 1953, TORGERSON 1958, DE GROOT 1961, SUPPES en ZINNES 1963, BRODBECK 1963, ADAMS, FAGOT en ROBINSON 1965.)

Wat is nu het standpunt van STEVENS? In 'Mathematics, Measurement, and Psychophysics', het eerste hoofdstuk van het 'Handbook' begint STEVENS (p. 23) zijn verhandeling over meetschalen met de – uiterst ruime – definitie: "A rule for the assignment of numerals (numbers) to aspects of objects or events creates a scale". Vervolgens wijst hij erop dat schalen mogelijk zijn dank zij een isomorfie[2] tussen de eigenschappen van de objecten en de getallen. Kunnen wij de objecten alleen maar onderscheiden, kunnen wij dus alleen maar bepalen of twee objecten voor wat betreft een bepaald attribuut gelijk zijn (tot dezelfde klasse behoren) of niet, dan is er een isomorfie tussen deze objecten en een *nominale schaal.* De getallen bij een nominale schaal worden alleen als klasse-namen gebruikt: twee objecten met hetzelfde getal behoren tot dezelfde klassen. In feite wordt hier niet gebruik gemaakt van de eigenschappen van de getallen; men zou ook

[2] De term 'isomorfie' wordt door DE GROOT (1961, p. 233 e.v.) wat nader toegelicht en krijgt in SUPPES en ZINNES (1963) een formele definitie. Het begrip verwijst naar een structurele overeenkomst tussen twee systemen ongeveer zo dat met elke relatie tussen elementen in het ene systeem een relatie (niet noodzakelijkerwijs dezelfde) van evenveel elementen in het andere systeem correspondeert (zie ook BRODBECK 1963, pp. 88 en 90-93).

letters of namen kunnen gebruiken en de toekenning van de getallen is volkomen willekeurig (TORGERSON (1958) noemt dit dan ook geen meetschalen). Bij een *ordinale schaal* kan men de objecten ten opzichte van een bepaald kenmerk op een rangorde zetten en toekenning van de getallen is nu minder willekeurig: de rangorde van de getallen moet overeenkomen met die van de objecten. Bij een *interval-schaal* kan men ook bepalen in welke mate objecten verschillen en de verschillen tussen de getallen – de afstand tussen de schaalpunten – hebben reële betekenis en liggen, op een constante factor na, vast. Bij een *verhoudingsschaal* (ratio scale) is er bovendien een, niet meer willekeurig, (zgn. 'natuurlijk') nulpunt.

Het schaaltype waartoe meetuitkomsten behoren heeft belangrijke konsekwenties, zo stelt STEVENS, voor wat men met deze getallen doen mag. Men kan de verschillende schalen zelfs definiëren door middel van de transformaties die de structuur onveranderd laten. De nominale schaal heeft vrijwel geen structuur en elke transformatie, waarbij één getal vervangen wordt door één ander getal en waarbij nooit twee verschillende getallen vervangen worden door hetzelfde getal, laat deze structuur in tact. Bij de ordinale schaal is iedere monotoon toenemende transformatie, iedere rangorde-bewarende transformatie, toegestaan: 1, 2, 3 wordt bijvoorbeeld door de transformatie $y = x^2$ tot 1, 4 en 9, met dezelfde rangorde. Bij de intervalschaal is iedere lineaire transformatie, $y = ax + b$, toegestaan en bij de verhoudingsschaal alleen een transformatie van de vorm $y = ax$, dat wil zeggen een scalaire transformatie (similarity transformation). Deze toegestane transformaties bepalen ook welke statistische bewerkingen op het materiaal uitgevoerd mogen worden: een gemiddelde bijvoorbeeld kan berekend worden op een intervalschaal maar is zinloos bij een ordinale schaal. STEVENS gebruikt het begrip invariantie om duidelijk te maken waarom dit zinloos is. Als drie objecten A, B en C respectievelijk de posities 1, 2 en 3 op een ordinale schaal innemen, dan is de 'score' van het object B het gemiddelde van de drie scores. Na een transformatie, bijvoorbeeld volgens $y = x^2$, die niets essentieels verandert, omdat de rangorde ongewijzigd blijft, krijgen de objecten de posities 1, 4 en 9. Het gemiddelde, 4 2/3, correspondeert nu niet meer met B's positie. Het 'gemiddelde object' is dus niet invariant onder toegestane schaaltransformaties. Dan is het echter ook niet meer zinvol om zo'n gemiddelde te berekenen[3].

[3] Verleidelijk is het wel! LUCE en RAIFFA (1957, p. 16) formuleren dit op tref-

2.4 KRITIEK OP STEVENS VAN STATISTISCHE ZIJDE

Overtuigend als deze conclusie lijkt, velen hebben kritiek geleverd op STEVENS' standpunt. Deze kritiek is globaal in te delen in twee klassen: De eerste betreft vooral de toepasbaarheid van statistische bewerkingsmethoden op testscores en dergelijke. Hierop komen we nog nader terug. De tweede groep van bezwaren tegen STEVENS' voorschriften komen ongeveer neer op de hierna geformuleerde mening.

Men kan altijd met cijfers doen wat men wil en het toetsen van statistische hypothesen is gerechtvaardigd als maar voldaan is aan de voorwaarden gesteld door het statistische model.

LORD (1953) is hierin het verst gegaan. In een artikel gesteld in de vorm van een humoristisch verhaal geeft hij als zijn mening te kennen, dat men zelfs op getallen met een nominaal karakter gemiddelden en overschrijdingskansen van verschillen tussen gemiddelden mag berekenen. (Men zie voor het begrip 'overschrijdingskansen' bijvoorbeeld DE JONGE en WIELENGA 1963.) In zijn verhaal zijn de getallen nummers die studenten uit een automaat (zo iets als de ook in Nederland veel gebruikte koffie-automaat) trekken. De professor die de automaat gevuld had, had er zorg voor gedragen dat de nummers in volkomen willekeurige volgorde in de automaat zaten. Achteraf bleek echter dat de eerste-jaarsstudenten lagere nummers kregen dan de tweede-jaarsstudenten. De professor, een psycholoog, vroeg zich af of iemand met de machine geknoeid had of dat dit verschil ook aan het toeval toegeschreven kon worden. Een statisticus, die te hulp geroepen werd, berekende gemiddelden en zelfs de overschrijdingskans van het verschil van het eerste-jaarsgemiddelde met het populatiegemiddelde (het gemiddelde van alle nummers in de automaat). De professor, een overtuigd aanhanger van STEVENS, was geschokt: "But you can't multiply 'football numbers' ", the professor wailed. "Why, they aren't even ordinal numbers, like test scores". "The numbers don't know that, said the statistician. "Since the numbers don't remember where they came from, they always behave just the same way, regardless" (LORD 1953, p. 751).

Terecht merken BEHAN en BEHAN (1954) in hun kritiek op LORD op, dat de bewerkingen – als er geen isomorfie bestaat tussen deze

fende wijze, sprekende over getallen toegekend aan gedrags- of andere alternatieven om een persoonlijke voorkeursrangorde aan te geven: "... their very manipulative convenience is a source of trouble, for one must develop an almost inhuman self-control not to read into these numbers those properties which numbers usually enjoy".

getallen en de objecten – slechts betrekking kunnen hebben op de getallen. En het zijn geen getallen waarin wij geïnteresseerd zijn. ADAMS et al. (1965) wijzen er nog op dat de door LORD geïntroduceerde fictieve statisticus in feite geen hypothese toetste omtrent eigenschappen van voetballers, maar omtrent een methode om uit een nummerautomaat nummers te krijgen. LORD's statisticus was dus zinvol bezig, maar dat feit is voor het probleem onder discussie niet relevant.

Deze toelichting laat zien dat het verhaaltje van LORD niet alleen geestig, maar ook ondoorzichtig is en daarom op het eerste gezicht zo moeilijk te weerleggen schijnt: de statisticus had gelijk en toch ook weer niet. Zelfs STEVENS (1959) heeft gesteld dat men met getallen mag doen wat men wil: ". . . as I see this issue, there can surely be no objection to anyone computing any statistic to suit his fancy . . .". Of men iets met de resultaten van die bewerkingen kan doen, of ze nog zinvol zijn, is echter een andere kwestie.

Het voorbeeld van LORD was misleidend omdat er niet duidelijk in gespecificeerd was welke functie de getallen hadden. Meetgetallen hebben in het algemeen de functie om bepaalde aspecten van de werkelijkheid in een wiskundig systeem af te beelden. Deze afbeelding heeft alleen dan zin wanneer de structuur van de getallen overeenkomt met die van bepaalde eigenschappen van de afgebeelde objecten. Als deze overeenkomst bestaat spreekt men van isomorfie. De structuur van de meetgetallen kan volgens STEVENS tot uitdrukking gebracht worden door aan te geven welke transformaties toegestaan zijn.

Het spreekt vanzelf dat aan dezelfde getallen soms geheel verschillende structuren toegekend moeten worden als zij beschouwd worden als afbeeldingen van geheel verschillende (aspecten van) objecten. De getallen als afbeeldingen van de getal-aspecten van LORD's voetbalnummers hebben een geheel andere structuur dan getallen als afbeeldingen van voetballers. De eerste structuur laat vrijwel geen enkele transformatie toe, de tweede vrijwel elke.

De belangrijkste conclusie die wij hieruit kunnen trekken is deze: men kan slechts bepalen wat men met meetgetallen (zinvol) kan doen, als men iets weet van de structuur van de gemeten object-eigenschappen. Men moet dus steeds meetgetallen beschouwen *binnen een meetsysteem.* LORD's statisticus had natuurlijk gelijk met erop te wijzen, dat de getallen zelf niet wisten welk schaalkarakter ze hadden. Los van een meetsysteem (een begrip waarop we dadelijk nader in zullen gaan) hebben zij ook geen schaalkarakter in de zin van STEVENS. Los van een meetsysteem mag men dus ook met deze getallen spelen "to suit

his fancy". Binnen een meetsysteem is het echter wel zinvol om zich beperkingen op te leggen. Binnen zo'n systeem kan men namelijk van bewerkingen in termen van statistische grootheden bepalen of ze zinvol zijn of niet. Zijn beweringen omtrent statistische grootheden, zoals gemiddelden, niet zinvol dan mag men deze grootheden wel berekenen maar het wordt dan wel een wat nutteloze zaak.

2.5 DE BEGRIPPEN 'MEETSYSTEEM' EN '(EMPIRISCH) ZINVOL'

We hebben in het voorgaande gesproken over *meetsystemen* en het begrip *zinvol* gebruikt. Met meetsystemen duiden wij aan het begrip dat ADAMS et al. (1965) 'Numerical Assignment System' noemen. Zij geven van dat begrip een uiterst formele definitie. We zullen hier slechts een omschrijving geven en een voorbeeld.

Een meetsysteem bestaat uit vier klassen van elementen:

1. *objecten* of aspecten van objecten die gemeten worden;
2. *(meet)getallen;*
3. *meetfuncties* (numerical assignments) die getallen in betrekking brengen tot objecten;
4. *toegestane transformaties.*

Een eenvoudig voorbeeld: een psycholoog heeft aan een klas van 20 leerlingen een test afgenomen en daarna de scores bepaald. De leerlingen zijn in dit geval de objecten, de testscores zijn de meetgetallen en de meetfunctie is de relatie leerling-testscore.

Een *functie*, bijvoorbeeld $y = 5x$, is een voorschrift dat voor elke mogelijke waarde van x aangeeft welke de bijbehorende waarde van y is. Zo'n functie kan de vorm hebben van een rekenvoorschrift ($y = 5x$), maar ook van een tabel met een serie getallenparen of, in het geval van een meetfunctie, met een serie getal-object-combinaties.

De toegestane transformaties zijn, in ons voorbeeld, alle monotone transformaties of, als men bereid is aan te nemen dat testscores bij benadering een intervalschaalkarakter hebben, alle lineaire transformaties.

Wanneer is nu een bewering in termen van kwantitatieve grootheden ("deze functie heeft een totaalscore van 108", "die leerling heeft een testscore boven het klassegemiddelde") *zinvol*?

Een meetsysteem ontstaat wanneer men getallen aan objecten toe-

kent. Praktisch elke toekenning zoals hier bedoeld is tot op zekere hoogte willekeurig: zo'n toekenning (de meetfunctie) mag in bepaalde opzichten anders zijn zonder dat daarmee iets essentiëels verandert. (Men zegt dan, dat bepaalde transformaties van de meetgetallen toegestaan zijn.)

Hoe zwakker nu het schaaltype, des te groter is de willekeur – of zo men wil: de vrijheid – bij het toekennen van de getallen. Zo is men bij een ordinaal meetsysteem vrij in de keuze van de meetgetallen als de rangorde maar juist is.

Deze willekeur brengt met zich mee dat sommige eigenschappen van de toegekende getallen geen enkele betekenis hebben. Uitspraken gebaseerd op deze toevallige eigenschappen zijn daardoor zelf ook niet 'zinvol', in de zin dat het van toevallige omstandigheden of arbitraire beslissingen afhangt of ze waar zijn of niet. Omgekeerd worden uitspraken 'zinvol' genoemd, als deze berusten op niet-arbitraire aspecten van de meetgetallen, dat wil zeggen, op die eigenschappen, die niet door toegestane transformaties aangetast worden. SUPPES en ZINNES (1963) formuleren dit als volgt: "A numerical statement is *meaningful* if and only if its truth (or falsity) is constant under admissible scale transformations of any of its numerical assignments, that is, any of its numerical functions expressing the results of measurement". Een fictief voorbeeld moge een en ander verduidelijken.

Stel er zijn tien functies in een bedrijf en men kan deze functies betrouwbaar rangordenen. Nu geeft men de functie op het laagste niveau een rangnummer 1; de hoogst gewaardeerde functie krijgt een 10 toegewezen. Daarna worden de functies echter gegradeerd met de GM; de resulterende totaalpuntentallen (wij noemen die hier T-scores) vertonen dezelfde rangorde als de rangnummers. Het resultaat van een en ander is weergegeven in tabel 2.1. (Zie p. 24.)

De uitspraak: "functie F, gewaardeerd volgens de GM, heeft een T-score beneden het gemiddelde van de 10 bedrijfsfuncties" is juist. Het gemiddelde van de T-scores is 74 en de T-score van F is 45. De uitspraak is dus *semantisch* gesproken zinvol; zij heeft een betekenis, een toetsbaar waarheidsgehalte. Dit waarheidsgehalte is echter afhankelijk van de wijze van graderen. Kent men rangnummers toe in plaats van T-scores dan ligt F *boven* het groepsgemiddelde. De keuze tussen de twee alternatieve meetfuncties (rangnummers of T-scores) is, meettechnisch bezien, volkomen willekeurig, als men er van uitgaat dat de GM slechts een rangordening van functies mogelijk maakt. Beide meetfuncties geven evenveel informatie en in feite zelfs dezelfde infor-

TABEL 2.1

Een voorbeeld van twee meetfuncties binnen eenzelfde meetsysteem

functies	rangnummers	T-scores
A	1	20
B	2	25
C	3	30
D	4	35
E	5	40
F	6	45
G	7	131
H	8	134
I	9	139
J	10	141
gemiddelde:	5,5	74

matie. Het waarheidsgehalte van de bewering omtrent het gemiddelde hangt dus af van een willekeurige beslissing die met de gemeten objecten (functies) niets te maken heeft en daarom zegt men dat zo'n bewering *empirisch* zinledig is.

ADAMS et al. (1965) wijzen, nog duidelijker dan SUPPES en ZINNES dat doen, op het belang van de *relaties* tussen statistische bewerkingen, uitspraken en meetsystemen. Zij stellen dat men van statistische bewerkingen nooit zonder meer kan zeggen of zij 'geoorloofd' zijn of niet. Men kan slechts zeggen of ze toepasbaar zijn binnen een bepaald meetsysteem (vergelijk onze bespreking van LORD's artikel) en met betrekking tot bepaalde uitspraken. De aandacht die de uitspraken en het meetsysteem hier krijgen is voor ons probleem van belang. We worden daardoor gedwongen om ons de vraag te stellen: wat maakt het in feite voor verschil als 'verkeerde' bewerkingen op het materiaal worden toegepast? Welke uitspraken worden dan zinloos door een te variabel waarheidskarakter? En: gelden deze bezwaren ook voor ons meetsysteem? Wij zullen op deze vragen in verband met de werkclassificatie nog uitvoerig ingaan. Eerst willen we hier terug komen op een andere vorm van kritiek op STEVENS.

2.6 KRITIEK OP STEVENS VAN OPERATIONALISTISCHE ZIJDE

Tot nu toe hebben we gesproken over de kritiek die ervan uitging dat getallen tenslotte altijd getallen zijn, waar ze ook vandaan komen, en dat men hiermee dus kan doen wat men wil als maar aan bepaalde

statistische voorwaarden is voldaan. De zin van uitspraken gebaseerd op zulke bewerkingen bleek echter in veel gevallen twijfelachtig te zijn.

Van geheel andere zijde is ook kritiek gekomen op de als te stringent ervaren voorschriften van STEVENS. Gedoeld wordt hier op de minder theoretische maar meer op het praktische nut gerichte bezwaren van de kant van testpsychologen. Zo geeft COMREY (1950) in een overigens weinig helder betoog, als zijn mening dat "a set of logical and operational criteria as the ultimate standard by which a procedure of measurement is to be judged" de onderzoeker te veel beperkingen oplegt: "The true criterion by which a scientific procedure is to be judged is productivity".

In een volgende, veel duidelijker uiteenzetting stelt hij dat "the excellence of measurement methods in mental testing may be judged by the practical validity of those methods for the purposes at hand, . . ." (COMREY 1951). Dit tweede artikel is vooral duidelijker omdat COMREY nu een onderscheid maakt tussen de eisen van de praktijk ("The objectives of mental testing are held to be primarily empirical in nature") en die van de theorie-ontwikkeling. Hij ziet wel in dat voor het laatste doel isomorfie, of zoals hij het formuleert, het bestaan van een optellingsoperatie, noodzakelijk is en het ontbreken ervan dus een nadeel maar hij aanvaardt dit als voorlopig op het terrein van psychologische tests onvermijdelijk: "It seems reasonable to assert that mental testing is and will be for some time essentially an empirical science with certain rather well defined practical objectives rather than primarily a theoretical scientific enterprise" (op. cit., p. 330).

Dit onderscheid tussen praktijkeisen en theorieontwikkeling zijn we in een ietwat andere vorm bij DE GROOT (1961) ook al tegengekomen. Hij maakte daar onderscheid tussen metende en voorspellende tests en eiste van de laatste slechts dat ze zouden voorspellen. In een latere formulering komt zijn opvatting nog dichter bij die van COMREY: "Il semblerait en premier lieu, qu'historiquement parlant, le but premier des constructeurs de tests *n'était pas tant de mesurer des propriétés théoriquement fondamentales que des propriétés d'utilité pratique* telles que les performances humaines ou les capacités" (DE GROOT 1962; cursivering van de schrijver). SUPPES en ZINNES (1963) laten zich in dezelfde zin uit als zij spreken over het gebruik van tests: "The justification in the use of such instruments would then lie solely in the degree to which they are able to predict significant events, not as with most 'normal' fundamental measures, in the homomorphism between an empirical system and a numerical system".

Zeer verhelderend is de uiteenzetting van ADAMS et al. (1965) op dit punt. Juist met betrekking tot testscores, zo schrijven zij, bestaat er zo vaak onenigheid over wat men wel of niet hiermee mag doen. Dit is een direct gevolg van het feit dat er geen duidelijke isomorfie bestaat tussen deze scores en een empirisch systeem van eigenschappen. Daardoor kan ook geen preciese definitie gegeven worden van toegestane transformaties. Zij concluderen: "However, this is not to say that questions of appropriateness and significance cannot legitimately be raised about uses of statistics in connection with scales, only that our theory (and, we believe, any theory based on the idea of a permissible transformation) will not help much to resolve them".

2.7 TOEPASSING OP DE WERKCLASSIFICATIE

Laat ons na deze theoretische beschouwingen terugkeren tot het probleem dat reeds door WIEGERSMA gesignaleerd werd: als de graderingscijfers rangordegetallen zijn, mag men deze dan optellen en vermenigvuldigen? We zullen een omgekeerde benaderingswijze volgen en ons afvragen tot welk schaaltype de graderingscijfers moeten behoren, willen de T-scores minstens een ordinaal karakter hebben. We volgen hierbij de methode van SUPPES en ZINNES (1963) die door middel van enkele voorbeelden duidelijk maken wat de implicaties zijn van hun definitie van het begrip 'zinvol', hierboven reeds geciteerd. Zij kiezen daartoe bepaalde in wiskundige termen geformuleerde beweringen (bijvoorbeeld $x_a + x_b > x_c$, waar a, b en c verwijzen naar objecten en x een uitkomst van een meting – a numerical assignment – is) en laten zien wat er gebeurt met het waarheidsgehalte van deze bewering als een toegestane transformatie toegepast wordt. Blijft dit waarheidsgehalte gelijk dan noemen zij de nieuwe bewering 'equivalent' met de oude en dan is volgens hun definitie de bewering zinvol.

Een voor ons geval relevant voorbeeld is het vijfde:

$x + y = a$

of in woorden: de som van de twee metingen x en y uitgevoerd op alle elementen van een groep A is gelijk een constante. Men kan zich voorstellen dat x en y staan voor de in verhoudingsschaal uitgedrukte grootheden 'lengte' en 'gewicht'. De scalaire transformaties $x' = kx$ en $y' = my$ zijn dan toegestaan zelfs wanneer $k \neq m$ en toepassing hiervan leidt tot

$kx + my = a'$.

Er is geen enkele waarde te bedenken voor a' zodanig dat deze laatste bewering equivalent wordt aan de niet getransformeerde. De bewering is dus niet zinvol als de scalaire transformatie toegestaan is.

Dit voorbeeld kan als volgt aangepast worden aan de werkclassificatiesituatie: om de discussie te vereenvoudigen nemen we aan dat er slechts twee gezichtspunten, x en y, zijn met afweegfactor 1. Als $T(= x + y)$ een ordinaal karakter heeft is de rangorde van de objecten (nu: functies) niet willekeurig en deze rangorde mag niet door een toegestane transformatie verstoord worden.

Stel dat voor de functies a en b geldt

$T_a > T_b$,

dus

$x_a + y_a > x_b + y_b$.

Na bijvoorbeeld een scalaire transformatie wordt dit

$kx_a + my_a > kx_b + my_b$

en één enkel voorbeeld is nu voldoende om aan te tonen dat deze bewering niet equivalent is met de oorspronkelijke:

laat

$x_a = 4 \quad x_b = 1$

$y_a = 4 \quad y_b = 5$

dus

$T_a = 8$, $T_b = 6$ en $T_a > T_b$.

Laat nu $k = 1$ en $m = 4$ dan wordt

$T'_a = 4 + 16 = 20$

$T'_b = 1 + 20 = 21$

en

$T'_a < T'_b$.

Zelfs scalaire transformaties, die nog bij verhoudingsschalen toegestaan zijn, leiden hier al tot een verstoring van de rangorde in T. Alleen wanneer de graderingscijfers per gezichtspunt de eigenschappen van een absolute schaal (SUPPES en ZINNES 1963, p. 11) (of desnoods die van een 'difference scale', op. cit., p. 70) bezitten, is de uitspraak $T_a > T_b$ zinvol, maar dan is ook de T-score gemeten op een absolute schaal[4].

[4] Een absolute schaal is zelfs buiten de gedragswetenschappen een zeldzaamheid. De term impliceert dat geen enkele transformatie (of zoals wiskundigen het graag uitdrukken: alleen de identiteitstransformatie) geoorloofd is. Een voorbeeld van een absolute schaal is de toekenning van getallen aan verzamelingen van objecten in overeenstemming met de aantallen elementen per verzameling.

De Deskundigencommissie (1952 en 1959) heeft nu herhaaldelijk gesteld dat de graderingscijfers slechts een ordinaal karakter hebben. Zij heeft overigens wel gezien dat men graderingscijfers niet zonder meer kan optellen en wel omdat het hier ongelijksoortige grootheden betreft. Zij wilde deze ongelijksoortigheid echter opheffen door middel van afweegfactoren, dus door scalaire transformaties toe te passen. De ongelijksoortigheid van variabelen komt tot uitdrukking in het feit dat men ze, elk afzonderlijk, kan onderwerpen aan de binnen hun meetsysteem toegestane transformaties (waarbij dus eventueel $k \neq m$). Men kan dus door de afweegtransformatie niet van ongelijksoortige variabelen gelijksoortige maken.

Er is echter theoretisch wel een situatie denkbaar waar een procedure met afweegfactoren zinvol is. Men moet dan uitgaan van gelijksoortige variabelen die alleen in andere eenheden zijn uitgedrukt. Stel bijvoorbeeld dat men op een of andere wijze van een aantal functies kan vaststellen hoeveel zij volgens een bepaald gezichtspunt waard zijn. Als men dit per gezichtspunt doet kan het gebeuren dat men wel tot een intervalachtige waardeschaal per gezichtspunt komt maar dat de eenheden, zeg maar de 'munteenheid', per gezichtspunt verschilt. Dit verschil in eenheid zou men dan kunnen corrigeren door middel van afweegfactoren. Zo iets moet de DC wel voor de geest hebben gestaan toen zij schreef over ongelijksoortige grootheden, waarvan eerst de relatieve waarde vastgesteld moest worden voordat zij opgeteld konden worden.

Laat ons aannemen dat we de verschillende gezichtspunten inderdaad langs één en dezelfde dimensie kunnen meten – de graderingscijfers zijn dan dus gelijksoortige grootheden – en dat eventuele verschillen in 'munteenheid' door afweegfactoren gecorrigeerd zijn. We kunnen dan weer nagaan, op dezelfde wijze als hierboven, welke schaaleigenschappen de graderingscijfers minstens moeten hebben. Laat ons weer even aannemen dat er slechts twee gezichtspunten zijn, maar dat de afweegfactoren nu gelijk zijn aan w_x en w_y. We zouden dan voor de functies a en b hebben

$$w_x x_a + w_y y_a = T_a.$$
$$w_x x_b + w_y y_b = T_b.$$

Laat bovendien

$$T_a > T_b.$$

Transformatie van de variabele x moet nu gepaard gaan met eenzelfde transformatie van y om de gelijkheid van 'munteenheid' niet te verstoren. We transformeren nu de variabelen $X = w_x x$ en $Y = w_y y$

door middel van een lineaire transformatie tot $X' = kX + m$ en $Y' = kY + m$.
We hebben in termen van $X(= w_x x)$ en $Y(= w_y y)$

$$X_a + Y_a > X_b + Y_b$$

en na transformatie en groepering van termen krijgen we

$$k(X_a + Y_a) + 2m > k(X_b + Y_b) + 2m.$$

Het is duidelijk dat voor positieve k deze uitdrukking te herleiden is tot de oorspronkelijke en dat de twee uitdrukkingen dus equivalent zijn. Als de graderingscijfers minstens op een intervalschaal gemeten zijn, is de uitspraak $T_a > T_b$ zinvol, als we kunnen aannemen dat de graderingscijfers voor de verschillende gezichtspunten in feite telkens betrekking hebben op eenzelfde 'waarde' en als de afweegfactoren deze cijfers tot dezelfde 'munteenheid' terugbrengen. Uiteraard heeft onder deze voorwaarden de T-score ook het karakter van een intervalschaal.

Als we nu de DC op haar woord nemen en de graderingscijfers beschouwen als niet meer dan rangorde-getallen dan zijn alle monotoon toenemende transformaties toegestaan. Het is duidelijk dat dan een uitspraak van de vorm $T_a > T_b$ niet meer zinvol is, aangezien het waarheidsgehalte daarvan niet meer invariant is onder de toegestane transformaties. Eén voorbeeld is genoeg om dit aan te tonen:
stel

$$X_a = 1 \qquad X_b = 2$$
$$Y_a = 5 \qquad Y_b = 3,$$

zodat

$$X_a + Y_a > X_b + Y_b.$$

Definiëren wij nu een transformatie $X' = f(X)$ als volgt

X	$X' = f(X)$
1	1
2	5
3	6
5	9

(f(X) monotoon toenemend) dan leidt toepassing van deze transformatie op X en Y tot

$$\left.\begin{aligned} X_a' + Y_a' &= 10 \\ X_b' + Y_b' &= 11 \end{aligned}\right\} T_a < T_b,$$

zodat de rangorde van de functies a en b omgekeerd is. Ordinale variabelen zijn dus niet optelbaar, zelfs wanneer ze betrekking hebben

op hetzelfde begrip en dus in die zin gelijksoortig zijn. Dit resultaat is overigens niet zo verrassend. Men kan het eenvoudig zo formuleren: als afweegfactoren gebruikt worden om de 'munteenheid' van alle graderingscijfers gelijk te maken dan moet er ook een 'munteenheid' of schaaleenheid zijn! De ordinale schaal echter ontbeert die eenheid juist.

Vatten we samen in enkele punten:

a. de DC ging van de veronderstelling uit dat de graderingscijfers ongelijksoortige grootheden zijn die echter na vermenigvuldiging met afweegfactoren optelbaar worden. Ongelijksoortige grootheden blijven echter ongelijksoortig, ook na vermenigvuldiging met welke factoren dan ook. Alleen een verschil in schaaleenheid kan op deze wijze gecorrigeerd worden. De DC heeft dus met haar term "ongelijksoortig" waarschijnlijk slechts willen aangeven dat de schaaleenheden verschillen;

b. de DC ging van de veronderstelling uit dat de graderingscijfers een ordinaal karakter hebben. Dit impliceert echter dat deze cijfers, zelfs als voldaan is aan de veronderstelling van gelijksoortigheid, niet optelbaar zijn;

c. wil de methode in zijn metende functie aanvaardbaar zijn dan moeten we aannemen,
   1. dat de graderingscijfers gelijksoortige grootheden zijn die slechts in gebruikte schaaleenheid verschillen;
   2. dat de verhoudingen tussen de schaaleenheden der gezichtspunten juist geschat kunnen worden en
   3. dat de graderingscijfers minstens een intervalschaalkarakter hebben.

We hebben hierboven naar aanleiding van het standpunt van ADAMS et al. (1965) op de noodzaak gewezen om na te gaan of deze theoretische bezwaren ook voor de praktijk van de werkclassificatie gelden. Een belangrijk punt van overweging vormt hierbij de vraag: zijn de monotone transformaties die bij ordinale meetsystemen toegestaan zijn, ook in de praktijk in feite mogelijk? Zou een taakanalist, een personeelschef of een vakbondsvertegenwoordiger onverschillig staan tegenover een willekeurige monotone transformatie van de graderingscijfers? De vraag heeft niet meer dan een rhetorische betekenis! Door de binding van de GM-resultaten aan het loon hebben de cijfers

achteraf een quasi-absoluut karakter gekregen. Het is daarbij de vraag of dit wel te verantwoorden is als deze cijfers, zoals de DC stelt, niet meer dan rangordecijfers zijn. In de beginfase van de ontwikkeling van de methode moet de toekenning van cijfers in dat geval, afgezien van de rangorde, op volkomen willekeurige wijze geschied zijn. (Een Kennisgradering van bijvoorbeeld 4 betekent slechts: meer dan $3\frac{1}{2}$. Een 5 zou dezelfde betekenis gehad hebben.) Een andere, even goed te verdedigen, cijfertoekenning had echter een geheel andere rangorde op de T-schaal kunnen geven. Als de afweegfactoren los van de graderingscijfers en van de praktijksituatie vastgesteld waren, als men zich bij deze vaststelling dus uitsluitend gebaseerd had op een analyse van het relatieve gewicht der verschillende gezichtspunten, dan was de rangorde der functies volkomen willekeurig geweest. De redelijke overeenstemming die met de praktijk bereikt werd is hiermee in strijd.

Dit heeft niemand ooit verwonderd. Het werd in de inleiding reeds gesteld: de methode is ontworpen om de praktijksituatie zo goed mogelijk te reproduceren of in de in paragraaf 2.2 gebruikte terminologie: om deze als criterium zo goed mogelijk te voorspellen. De afweegfactoren zijn dus niet los van graderingscijfers en praktijksituatie vastgesteld en werden niet, zoals de DC zich had voorgesteld, gebruikt als correctie op ongelijke schaaleenheden. Zij werden gebruikt om een aanpassing aan de praktijksituatie te bewerkstelligen.

2.8 KRITIEK OP DE DESKUNDIGENCOMMISSIE

Het onderscheid dat wij hier, in navolging van DE GROOT, gemaakt hebben tussen metende en voorspellende variabelen, is door de DC niet gezien. Zij heeft enerzijds ter verdediging van haar werkwijze quasi-theoretische argumenten aangevoerd waarin begrippen als ongelijksoortige grootheden en rangordegetallen werden gehanteerd. Anderzijds heeft zij met afweegfactoren en graderingscijfers zodanig "geschoven" tot een redelijke aanpassing aan de praktijk ontstond (LIJFTOGT 1966, Hfdst. IV). Op zichzelf is er niets tegen om een instrument te ontwikkelen aan de hand van praktijkcriteria, al zou men wel iets meer systematisch te werk hebben kunnen gaan. Bezwaar wordt hier echter gemaakt tegen de ondoorzichtige verantwoording van de gevolgde procedure. Hierdoor is een zinvolle discussie over de waarde van de methode altijd bijzonder moeilijk geweest. Moest men de methode beschouwen als een criteriumvariabele, waarbij schaal-

technische overwegingen van belang zijn of moest men haar zien als een voorspeller, dat wil zeggen als een efficiënt substituut van een weliswaar meer zuivere, maar te tijdrovende (criterium-)meetprocedure? In dat geval zou definitie en meting van dit criterium het centrale probleem vormen. We hebben ons voorlopig op het eerste standpunt geplaatst en zijn tot de conclusie gekomen dat de eisen, die in dit geval aan de methode gesteld moeten worden, veel zwaarder zijn dan de DC het voorgesteld heeft; het bleek namelijk dat de graderingscijfers reeds voor de invoering van de afweegfactoren gelijksoortige grootheden zouden moeten zijn en dat zij bovendien intervalschaalkarakter zouden moeten hebben.

Men kan zich voorstellen dat de gelijksoortigheid der gezichtspunten min of meer bij fiat vastgesteld wordt. Wel moet hierbij aangetekend worden dat het dan nauwelijks meer houdbaar is om binnen één methode zowel gezichtspunten die betrekkingen hebben op kennis en kunde als ook inconveniënten te handhaven. De gelijksoortigheid van twee zozeer uiteenlopende groepen van gezichtspunten is daarvoor te twijfelachtig. Men heeft in de praktijk reeds aanwijzingen hiervoor kunnen vinden in de recente voorstellen tot wijziging van de waardering der inconveniënten (Werkclassificatie 1965). Deze voorstellen werden verdedigd door te wijzen op de sterk veranderde bereidheid tot het verrichten van zwaar en onaangenaam werk.

Tot nu toe werd eveneens aangenomen dat de graderingscijfers rangordekarakter hadden. Veel moeilijker te verdedigen is de veronderstelling als zouden zij een intervalschaalkarakter hebben. De conclusie die men voorlopig moet trekken is voor de werkclassificatie weinig gunstig: wil de T-score als gewogen som van graderingscijfer zinvol zijn, dan moeten deze graderingscijfers minstens een intervalschaalkarakter hebben en bovendien vergelijkbare grootheden zijn.

Over het schaalkarakter der gezichtspuntgraderingen is echter niets bekend. De werkclassificatie is hoogstens te beschrijven als een pseudo-afleesinstrument in de zin van SUPPES en ZINNES en de voor dit soort instrumenten noodzakelijke toetsing aan een praktijkcriterium is niet goed mogelijk. Bovendien wordt ook op andere gronden de opvatting als zou de GM een voorspeller zijn niet houdbaar geacht. De waarde van de werkclassificatie-methode in zijn huidige vorm als meetinstrument is op zijn minst uiterst twijfelachtig.

HOOFDSTUK III

Poging tot een constructieve bijdrage

Na de vooral kritische beschouwingen in hoofdstuk II, wordt in dit hoofdstuk gezocht naar oplossingen uit de impasse. Allereerst wordt daartoe nagegaan of het mogelijk is om graderingscijfers op een of andere wijze te combineren tot een totaalscore, rekening houdend met hun ordinaal karakter. Het blijkt dan dat om verschillende redenen de in de literatuur beschreven methoden van combinatie van rangordegegevens voor de werkclassificatie niet bruikbaar zijn.

Daarna wordt de mogelijkheid onderzocht om graderingscijfers om te zetten in rangnummers, dat wil zeggen meetgetallen met een absoluut schaalkarakter. Hoewel dit een oplossing lijkt te bieden voor de schaaltechnische bedenkingen tegen het optellen van de gezichtspuntwaarderingen, blijkt achteraf dat de kritiek van ADAMS et al. op LORD's artikel (zie paragraaf 2.4 hierboven) evenzeer van toepassing is op het gebruik van rangnummers, opgevat als meetgetallen met een absoluut schaalkarakter.

Wil men van de GM een bruikbaar meetinstrument maken dan zal men graderingsschalen moeten ontwikkelen met, althans bij benadering, een intervalkarakter. Een schetsmatig voorstel tot een experimentele ontwikkeling van zulke schalen wordt gegeven. Tenslotte worden verschillende vraagstukken, die op dat moment actueel zijn geworden, aan de orde gesteld: de keuze van de gezichtspunten en het probleem van de afweegfactoren. Een en ander wordt geïnterpreteerd vanuit het standpunt dat de GM tegelijk een meetinstrument is en een explicitering van waarden biedt. Op de konsekwenties hiervan wordt kort ingegaan.

3.1 HET COMBINEREN VAN RANGORDE-GEGEVENS

In het vorige hoofdstuk is aangetoond, dat de in de GM voorgeschreven procedure om graderingscijfers te combineren door optelling, niet te verdedigen valt als men uitgaat van het ordinale karakter van deze cijfers. Hier zal nu onderzocht worden of andere combinatiemethoden, die speciaal voor rangorde-gegevens ontwikkeld werden, voor de werkclassificatie bruikbaar zijn.

Het is moeilijk in de literatuur een algemene beschouwing te vinden over de combinatie van rangorde-getallen. Er is wel onderzoek gedaan naar het combineren van ordinale meetgetallen op een zeer speciaal gebied van de gedragswetenschappen, namelijk dat van de besliskunde.

Vooral pogingen om rechtvaardige methoden te vinden om individuele keuzen te combineren tot een algemeen aanvaardbare groepsbeslissing of groepsvoorkeur zijn voor ons probleem relevant. LUCE en RAIFFA (1957) besteden een heel hoofdstuk aan het vraagstuk van 'group decision making'. Het gaat er daarbij om een regel te vinden volgens welke één rangorde van alternatieve beslissingen opgesteld kan worden op grond van een aantal individuele voorkeursrangschikkingen. Zo'n regel wordt een 'social welfare function' genoemd als deze in een of andere zin 'fair' is, als hierbij zo veel mogelijk gelijkelijk rekening gehouden wordt met de individuele wensen.

Meer concreet: stel er is een groep van drie personen, die elk vier alternatieve handelingsmogelijkheden naar hun persoonlijke voorkeur hebben geordend. Zo'n situatie is als volgt in een matrix weer te geven:

		personen:		
		P_1	P_2	P_3
alternatieven:	A_1	4	3	2
	A_2	1	2	1
	A_3	3	1	4
	A_4	2	4	3

De getallen in deze matrix hebben slechts een ordinale betekenis. Ze geven per kolom de voorkeursrangorde van een persoon voor de vier gegeven alternatieven weer. Zo gaat de voorkeur van persoon P_1 uit naar het eerste alternatief (A_1), terwijl het tweede alternatief bij hem het laatst voor realisatie in aanmerking komt. Een 'social welfare function' is een procedure, die het mogelijk maakt op grond van de drie gegeven rangschikkingen een vierde op te stellen, die representatief geacht kan worden voor de individuele voorkeuren.

De situatie bij de GM vertoont hiermee formeel een duidelijke overeenkomst. In de groepsbeslissingssituatie worden alternatieven per persoon geordend (naar voorkeur); in de werkclassificatie worden functies per gezichtspunt geordend (naar functie-eisen). In de groepsbeslissing moeten de individuele wensen zoveel mogelijk in gelijke mate tot hun recht komen, bij de GM moeten de gezichtspunten (na weging) in gelijke mate bijdragen tot de totaalrangorde.

Een onderzoek naar de verschillende, in de literatuur besproken, 'welfare functions' wees uit dat een in de praktijk – en vooral in de

praktijk van de werkclassificatie – bruikbare combinatieregel voor rangschikkingen niet gemakkelijk te vinden is.

Een grote moeilijkheid daarbij blijkt het optreden van intransitiviteiten te zijn. Het is daarom wenselijk dat hier het begrip 'transitiviteit' toegelicht wordt. Een verzameling van elementen (A, B, C, . . .) voldoet aan de voorwaarde van transitiviteit als voor elk drietal van deze elementen geldt, dat wanneer ApB en BpC dan ApC, waarbij 'p' een rangorde-relatie weergeeft (ApB betekent, bijvoorbeeld, dat A wordt verkozen boven B, of dat A wint van B). Transitiviteit wordt in het algemeen als een vanzelfsprekende zaak beschouwd. Als men 'p' interpreteert als 'groter dan', 'zwaarder dan' of 'sterker dan', wordt meestal zonder meer aangenomen dat intransitiviteiten van de vorm ApB, BpC en tòch CpA niet kunnen optreden. Betekent ApB echter, dat A in een wedstrijd heeft gewonnen van B, of dat A wordt verkozen boven B, dan wordt transitiviteit ineens veel minder vanzelfsprekend; in zulke gevallen kan men alleen door empirisch onderzoek nagaan of aan de voorwaarde voor een bepaalde verzameling voldaan is of niet.

Nu blijkt dat bij toepassing van een voor de hand liggende 'social welfare function', de zogenaamde 'eenvoudige meerderheidsregel', intransitiviteiten kunnen optreden, waardoor de regel als combinatieprocedure van voorkeurs- of andere rangschikkingen onbruikbaar wordt. In zo'n geval is namelijk de samenvattende rangorde niet gedefinieerd.

De eenvoudige meerderheidsregel houdt in dat de groep A verkiest boven B (ApB) dan en dan alleen als de meerderheid van de groepsleden A boven B verkiest. Het is eenvoudig om een voorbeeld te vinden waar toepassing van deze regel tot een intransitieve relatie tussen drie alternatieven leidt. LUCE en RAIFFA (1957, p. 333) geven zo'n voorbeeld waar een samenvattende (groeps-) rangorde niet gedefinieerd is:

	P_1	P_2	P_3
A_1	3	1	2
A_2	2	3	1
A_3	1	2	3

P_1 verkiest A_1 boven A_2 en A_2 boven A_3 (A_1pA_2 en A_2pA_3). Omdat de preferenties van de personen als rangschikkingen gegeven zijn kunnen hierbinnen geen intransitiviteiten optreden; waren ze als paarsgewijze vergelijkingen gegeven dan was dit wel mogelijk geweest. We

hebben dus tevens dat voor P_1 geldt: A_1pA_3. De voorkeursrelaties voor de andere personen zijn eveneens gemakkelijk uit de matrix af te lezen. De groepsvoorkeursrelaties vindt men door toepassing van de meerderheidsregel als volgt: P_1 en P_2 (de meerderheid dus) verkiezen A_1 boven A_2 dus voor de groep geldt eveneens A_1pA_2. Zo vindt men ook dat A_2pA_3 en A_3pA_1. Deze drie binaire relaties vormen tesamen een intransitiviteit.

Door BLACK en onafhankelijk van hem door COOMBS zijn de voorwaarden bestudeerd, waaraan een verzameling moet voldoen, opdat bij toepassing van de meerderheidsregel geen intransitiviteiten optreden. Men raadplege LUCE en RAIFFA (1957, p. 354-357) en COOMBS (1964, Hfdst. 18) voor een bespreking van dit onderzoek en de conclusies daarvan. Voor ons is slechts van belang dat aan de door BLACK en COOMBS geformuleerde voorwaarden bij werkclassificatie-materiaal in het algemeen niet voldaan zal zijn.

Een andere combinatie-procedure is voorgesteld door KEMENY (KEMENY 1959 en ook KEMENY en SNELL 1962, Hfdst. II). Hij heeft een afstandsfunctie gedefinieerd die aan elk paar rangschikkingen een afstand toekent. Hoe meer de rangschikkingen overeenkomen, des te kleiner is de afstand ertussen. De afstand tussen twee rangschikkingen die volledig overeenstemmen is nul. De groepsbeslissing behoort, volgens KEMENY, die rangorde te zijn die het meest lijkt op alle individuele rangschikkingen, of in termen van de afstandsfunctie: de best samenvattende rangorde is die rangorde, waarvan de som van de afstanden (of ook: de som van de kwadraten van de afstanden) tot de individuele rangschikkingen zo klein mogelijk is[1].

Een groot nadeel van deze twee oplossingen van het combinatieprobleem is, dat de samenvattende rangorde niet altijd uniek is: er is altijd minstens één oplossing (het intransitiviteitsprobleem doet zich niet voor), maar vaak zijn er meer en dit geldt zowel voor de kleinst mogelijke som van de afstanden als voor de kleinst mogelijke som van

[1] De definitie van de afstandsfunctie vertoont een opvallende overeenkomst met de grootheid S in KENDALL's formule voor de rangcorrelatiecoëfficiënt, tau, ook een maat voor de overeenkomst tussen twee rangschikkingen (KENDALL 1962, p. 5). De relatie tussen de afstandsfunctie, d, en S is gegeven door $S = -d + \frac{1}{2}n(n-1)$, waar n het aantal gerangschikte objecten is. Deze relatie geldt alleen dan wanneer in geen van de rangschikkingen objecten voorkomen, die dezelfde plaats in de rangorde bezetten (geen 'ties'). In zo'n geval zal dan ook de rangorde met de kleinst mogelijke gemiddelde afstand tot de individuele rangschikkingen tevens de rangorde zijn waarvan de gemiddelde tau met alle rangschikkingen maximaal is.

de kwadraten van de afstanden. Bovendien is het vinden van een samenvattende rangorde bij meer dan drie alternatieven praktisch niet meer uitvoerbaar, aangezien het aantal mogelijkheden onoverzienbaar groot wordt en er geen systematische methode bestaat om tot een oplossing te komen.

LUCE en RAIFFA (1957, p. 358 e.v.) vermelden nog twee modificaties van de meerderheidsregel, voorgesteld door COPELAND[2] om het optreden van intransitiviteiten te voorkomen. In het eerste geval wordt de samenvattende rangorde gekozen in overeenstemming met een index u(x), die aangeeft hoeveel alternatieven het verliezen van een alternatief x bij toepassing van de meerderheidsregel min het aantal alternatieven dat het wint van x. Een voorbeeld moge deze definitie verduidelijken. Stel we hebben de volgende situatie:

	P_1	P_2	P_3
A_1	1	3	3
A_2	2	2	1
A_3	3	1	2

De indices voor de drie alternatieven zijn respectievelijk $u(A_1) = 2 - 0 = 2$ (A_2 en A_3 verliezen het van A_1, géén alternatief wint het van A_1), $u(A_2) = 0 - 2 = -2$ en $u(A_3) = 1 - 1 = 0$. De samenvattende rangorde, in overeenstemming met deze indices, is dus

	T
A_1	3
A_2	1
A_3	2

De tweede, door COPELAND voorgestelde, wijziging houdt in de bepaling van een index v(x), die is gedefiniëerd als het totaal aantal malen dat het alternatief x boven een ander alternatief gesteld wordt (in alle rangschikkingen) min het aantal malen dat het lager gesteld wordt. In het bovengegeven voorbeeld vinden we $v(A_1) = 4 - 2 = 2$, $v(A_2) = 2 - 4 = -2$ en $v(A_3) = 3 - 3 = 0$. De samenvattende rang-

[2] Het bleek niet mogelijk de hand te leggen op de, door LUCE en RAIFFA in dit verband genoemde, gestencilde publikatie van COPELAND.

orde is in dit geval dezelfde als die volgens de index u(x), maar dit is in het algemeen niet het geval, hetgeen uit een hieronder te behandelen voorbeeld duidelijk zal blijken.

De index v(x) staat in een eenvoudige relatie tot rangnummersommen. Het rangnummer van x in een bepaalde rangorde is immers gelijk aan het aantal malen dat x in die rangorde boven een ander alternatief geplaatst wordt plus 1. Als er in totaal n alternatieven zijn, is het aantal malen dat x in een bepaalde rangorde lager gesteld wordt n — R(x), waar R(x) het rangnummer van x is. De index v(x) is dus per definitie [R(x) — 1] minus [n — R(x)] en dit verschil gesommeerd over alle k rangschikkingen, of

$$v(x) = 2\Sigma R(x) - k(1 - n),$$

waar $\Sigma R(x)$ de rangnummersom van x is.

Deze twee door COPELAND voorgestelde methoden zijn in de praktijk gemakkelijk hanteerbaar en zullen daarom nader vergeleken worden. Dit kan het eenvoudigst aan de hand van een voorbeeld geschieden. In tabel 3.1 zijn de rangschikkingen van vijf functies op 11 gezichtspunten gegeven. Een 5 betekent daar dat een functie op dat gezichtspunt het hoogst gewaardeerd wordt, een 1 dat deze functie lager gegradeerd wordt dan de andere vier functies. Voorlopig wordt het probleem van de afweging even genegeerd.

TABEL 3.1

functies	gezichtspunten											som rangnrs.	u(x)
	1	2	3	4	5	6	7	8	9	10	11		
A	1	1	1	1	1	5	5	5	5	5	5	35	4
B	5	5	5	5	5	4	4	4	4	4	4	49	2
C	4	4	4	4	4	3	3	3	3	3	3	38	0
D	3	3	3	3	3	2	2	2	2	2	2	27	—2
E	2	2	2	2	2	1	1	1	1	1	1	16	—4

Dit wat extreme voorbeeld laat zien dat de som der rangnummers – of COPELAND's v(x) – tot een rangorde leidt waar B bovenaan staat, gevolgd door C en waar A op de derde plaats komt, terwijl COPELAND's eerste methode – met de index u(x) – leidt tot een volgorde waar A bovenaan staat.

Welke rangorde zou nu in het kader van de werkclassificatie het meest aanvaardbaar zijn? Laat ons eens even aannemen, dat de eerste

vijf gezichtspunten in de sfeer van het kunnen liggen en de laatste zes in die van het willen (inconveniënten). In dat geval kunnen de gegevens van tabel 3.1 als volgt samengevat worden: er is wat betreft het kunnen een rangorde waar B bovenaan staat en A onderaan (B, C, D, E, A); er is volgens de inconveniënten een rangorde waar A bovenaan staat, onmiddellijk gevolgd door B en waar E onderaan staat (A, B, C, D, E). Omdat de inconveniënten in dit fictieve voorbeeld zwaarder wegen (in de meerderheid zijn, 6 tegen 5) dan het kunnen, is te verwachten dat A, die bij de inconveniënten als eerste naar voren komt, door zijn lage plaats op het gebied van het kunnen niet al te zeer geschaad wordt. De rangorde volgens de rangnummersommen – of volgens COPELAND's v(x) – is hiermee in overeenstemming: A staat nog op de derde plaats, vlak onder C. De rangorde volgens u(x) is echter wat onverwacht: A die bij vijf van de elf gezichtspunten geheel onderaan staat, verschijnt als eerste op de totaalranglijst, die geheel in overeenstemming is met de inconveniënten. De andere vijf gezichtspunten hebben dus geen enkele invloed gehad op de einduitkomst. Hoe is dit te verklaren? Het doorslaggevende bij de index u(x) is de uitkomst van de meerderheidsregel: Of A het van de overige functies wint met grote verschillen (11 tegen 0 gezichtspunten zoals B het van alle andere functies behalve van A wint) of met slechts kleine verschillen (6 tegen 5) heeft geen konsekwenties. Het gaat er slechts om hoe váák A het wint, niet om de score. Zouden de graderingscijfers, in dit geval de rangnummers, foutloos zijn en zou men absoluut niets kunnen aannemen over de afstanden tussen de verschillende functies, dan is u(x) de beste index. Het is zeker, dat A bij 6 gezichtspunten hoger is dan B, ook al verschillen ze slechts één rangnummer (5 en 4), de gegevens zijn immers bij aanname foutloos. Het verschil tussen rangnummers 5 en 4, waardoor A op 6 gezichtspunten hoger staat dan B, en het verschil tussen 1 en 5, waardoor A op 5 gezichtspunten lager staat, geeft dezelfde soort informatie die uitsluitend ordinaal van aard is: $A > B$ of $A < B$. Het is dus niet juist om uit het grotere verschil tussen 1 en 5, vergeleken met dat tussen 5 en 4, konsekwenties te trekken voor het eindresultaat, de samenvattende rangorde. Tellen we de rangnummers op dan doen we in feite juist dit.

Zijn de rangnummers echter niet foutloos en kan het dus zijn dat A 'per ongeluk' boven B staat, terwijl eigenlijk A het rangnummer 4, of zelfs 3 en B het rangnummer 1 had moeten krijgen en is men bereid aan te nemen dat de kans, dat de rangnummers door toevallige omstandigheden een verkeerde rangorde aangeven, kleiner wordt naar

mate de rangnummers meer verschillen dan wordt de v(x)-index aantrekkelijker. In de praktijk waar men altijd met feilbare gegevens moet werken en waar men, noodgedwongen, dus wel – soms niet geheel verantwoorde – aannamen moet maken, is het optellen van rangnummers (het gebruik van de v(x)-index) de meest voor de hand liggende van de twee methoden. Het wordt hier dan ook voorlopig verdedigbaar geacht om graderingscijfers uit te drukken als rangnummers of als relatieve rangnummers en deze op te tellen over gezichtspunten.

Is dit nu niet in strijd met wat in voorgaande bladzijden betoogd is, namelijk dat het empirisch zinloos is om ordinale meetgetallen op te tellen, of nog sterker, dat het alleen dan zinvol is om ongelijksoortige grootheden op te tellen als ze langs een absolute schaal gemeten zijn, dat wil zeggen dimensievrij zijn, zoals bijvoorbeeld wanneer het om aantallen gaat? (Zie voetnoot bladzijde 27.)

Deze tegenwerping klinkt overtuigend, want rangnummers lijken toch wel bij uitstek een ordinaal karakter te hebben. Rangnummers zijn echter in strikte zin géén ordinale grootheden: na elke andere dan de identiteits-transformatie houden ze op rangnummers te zijn. De enige transformatie die is toegestaan is $R'(x) = (n + 1) - R(x)$, waar n het aantal objecten en dus het hoogst voorkomende rangnummer is. Door deze transformatie toe te staan wordt slechts aangegeven, dat het een kwestie van afspraak is of men het hoogste rangnummer aan het hoogst gewaardeerde object geeft of aan het laagst gewaardeerde. Als men voor één van deze mogelijkheden kiest en deze dan ook konsekwent toepast, heeft men dus in feite met een absolute schaal te maken. Dit resultaat is niet zo verwonderlijk: het is immers niet moeilijk in te zien dat de som van rangnummers een zinvol gegeven is; het geeft namelijk aan hoeveel objecten, gerekend over alle gezichtspunten, lager gewaardeerd werden.[3]

Wellicht ten overvloede zij er hier op gewezen, dat we ons, ten gevolge van deze interpretatie van de index v(x), nu niet langer bezig houden met het probleem van de combinatie van rangorde-gegevens. Rangnummers zijn immers geen ordinale, maar absolute grootheden. Twee vragen doen zich daarbij voor. Ten eerste kan men zich afvragen wèlke betekenis aan de rangnummers gehecht kan worden, naast de manifeste van "aantallen functies die, gerekend over alle gezichtspunten, lager gewaardeerd worden dan de functie in kwestie". Dit zal in de volgende

[3] Zie bijvoorbeeld KENDALL (1962, p. 1), die er eveneens op wijst, dat rangnummers 'cardinal numbers' zijn, die ontstaan door tellen.

paragraaf ter sprake komen. Ten tweede kan het wegingsprobleem niet langer genegeerd worden. Uit het voorbeeld, hierboven gegeven, bleek dat het van doorslaggevende betekenis was, dat er zes inconveniënten-gezichtspunten tegenover vijf gezichtspunten op het gebied van het kunnen stonden. Wil men met rangnummersommen werken, dan zal men een belangrijk gezichtspunt meerdere keren met zijn rangschikking van de functies moeten laten 'meestemmen' om het zo zwaarder te laten wegen. Dit houdt in dat men, zoals dat in de GM gebruikelijk is, de graderingscijfers met een afweegfactor vermenigvuldigt. Bij het bepalen van de afweegfactoren zal dan rekening gehouden moeten worden met mogelijke dupliceringen of groeperingen van gezichtspunten: wil men dat kennis of meer in het algemeen, vaardigheden, zwaarder wegen dan inconveniënten, dan zal de som van de afweegfactoren voor de eerste groep gezichtspunten groter moeten zijn dan die voor de tweede. In deze som zit zowel de *grootte* van de afzonderlijke afweegfactoren als het *aantal* gezichtspunten.

3.2 HET SCHAALKARAKTER VAN DE RANGNUMMERS OPNIEUW BEZIEN

Definiëren we empirisch zinvol op dezelfde wijze als SUPPES en ZINNES dan lijkt het schaalprobleem nu opgelost: de aan een functie toegekende (gewogen) rangnummersom geeft aan hoeveel functies lager gewaardeerd worden volgens willekeurig welke gezichtspunten. Deze som is een ondubbelzinnig gegeven. We weten wat we er aan hebben, want rangnummers zijn geen rangorde-gegevens maar absolute meetgetallen, die niet getransformeerd mogen worden.

Een en ander hangt direct samen met de isomorfie tussen rangnummersommen en aantallen functies (die lager gewaardeerd worden). Het probleem is hiermee echter slechts schijnbaar opgelost: in feite is het alleen maar verschoven. De vraag wordt nu, wat voor betekenis men kan hechten aan deze aantallen in het kader van de beloning. Als deze aantallen voor de beloning irrelevante gegevens blijken te zijn, maken wij dezelfde fout die ook LORD's statisticus maakte (LORD 1953). De keuze van het meetsysteem is dan niet adekwaat.

Het lijkt een bijzonder gelukkige situatie dat het mogelijk is om door middel van rangnummers tot absolute meetgetallen te komen. Deze meetgetallen hebben echter niet direct betrekking op de eisen die een bepaalde functie aan de werknemer stelt. Zij zijn mede afhankelijk van de samenstelling van de functie-populatie. Het rangnummer kan

slechts gebruikt worden als een *indirecte* maat voor deze functie-eisen, omdat over de relatie tussen deze maat en de zwaarte van de eisen niet meer bekend is dan dat ze monotoon toenemend is. Meer kan men hierover niet zeggen omdat dit verband van populatie tot populatie en van gezichtspunt tot gezichtspunt kan variëren. Als men dus de rangnummers wil beschouwen en gebruiken als indices voor functie-eisen dan is de isomorfie tussen rangnummer en aantal functies niet meer interessant. Het gaat dan veeleer om de isomorfie tussen rangnummer en functie-eis en in het daarbij behorende meetsysteem kan men niet meer dan een ordinaal karakter aan de rangnummers toekennen.

Meer concreet: stel in een bedrijf wordt een functie A als zwaarder beoordeeld op het gezichtspunt Kennis dan een functie B. Tussen A en B vallen, wat betreft Kenniseisen, tien functies. Het rangnummer van A voor Kennis is in dat geval elf hoger dan dat van B. Worden nu in dit bedrijf tien nieuwe functies gecreëerd, die ook wat Kennis betreft tussen A en B vallen, dan wordt het verschil in rangnummer tussen A en B 21 punten groot. De verschillen in rangnummers tussen alle functies beneden B blijven echter ongewijzigd. Aangezien aan de functie-eisen zelf niets veranderd is, betekent dit dat de relatie tussen rangnummers en functie-eisen wel veranderd is. Alleen de rangorde binnen de rangnummers is gelijk gebleven aan die binnen de functie-eisen. De conclusie is weer: bezien we rangnummers binnen het relevante meetsysteem, waar functie-eisen de te meten aspecten van de objecten zijn (vergelijk paragraaf 2.5 hierboven), dan hebben zij slechts een ordinaal karakter.

De beschouwingen in de paragrafen 3.1 en 3.2 kunnen als volgt samengevat worden: een voor de werkclassificatie bruikbare combinatie-methode van rangorde-gegevens kon niet gevonden worden. De tweede methode van COPELAND leek eerst aanvaardbaar, omdat rangnummers in een bepaald meetsysteem als meetgetallen met een absoluut schaalkarakter kunnen worden beschouwd. Bij nader inzien is dit meetsysteem echter niet relevant. In een wèl voor ons probleem relevant meetsysteem hebben ook rangnummers een niet meer dan ordinaal karakter en is dus optelling niet zinvol.

3.3 DE ONTWIKKELING VAN INTERVALSCHALEN PER GEZICHTSPUNT

We zullen in deze paragraaf de konsekwenties moeten trekken uit de conclusie van de meettheoretische analyse van paragraaf 2.7. We weten

dat de gezichtspuntgraderingen gelijksoortige variabelen zouden moeten zijn, die een intervalschaalkarakter hebben. In de eerste plaats zal er daartoe een bezinning moeten plaats vinden op de gelijksoortigheid: kan men de eisen die een functie stelt op het gebied van kennis en kunde gelijksoortig achten met die op het gebied van het willen dulden van ongemak of gevaar? De waarde van een functie in loontechnisch opzicht wordt toch op geheel andere wijze belicht als het gaat om kenniseisen dan wanneer het gaat om inconveniënten?

In de tweede plaats zal althans een poging gedaan moeten worden om een schaalkarakter te krijgen waarbij schaalafstanden enigszins vergelijkbaar worden. De poging door middel van het toekennen van rangnummers faalde omdat het absolute schaalkarakter slechts gold in een meetsysteem dat voor werkclassificatiedoeleinden niet bruikbaar is. Een andere mogelijkheid is om een eenvoudige schaaltechniek toe te passen om tot min of meer gelijke intervallen te komen. Men kan bijvoorbeeld een aantal beoordelaars functies paarsgewijs laten vergelijken en hen daarbij dwingen altijd een oordeel 'hoger' of 'lager' te geven. Daarna kan men per definitie vastleggen dat de schaalafstand tussen twee functies A en B een halve (of een hele) punt bedraagt als 80% of 90% van de beoordelaars functie A hoger (of lager) waardeert dan functie B (vergelijk TORGERSON 1958, p. 137). Op deze wijze kan men een aantal voorbeeldfuncties vinden die *per gezichtspunt* de gehele schaal bestrijken (voor Kennis bijvoorbeeld 17 functies per bedrijfstak, voor contact misschien maar vier). Een bijkomend voordeel van deze methode zou wel eens kunnen zijn dat men gedwongen is functies per gezichtspunt te beoordelen en te vergelijken met telkens andere sleutelfuncties. Dit verzwakt een eventueel optredend halo-effect en bovendien dwingt het tot een analytische benadering van de taakanalyse. Het is mogelijk dat dit leidt tot meer betrouwbare resultaten. Of dit werkelijk zo is, is nog steeds een "matter of opinion" (VITELES 1941) en experimenteel onderzoek hierover is dringend gewenst.

Bij de ontwikkeling van een plan tot schaalconstructie *per gezichtspunt* zal er naar gestreefd moeten worden om zoveel mogelijk een onafhankelijkheid van het praktijk-oordeel te verkrijgen. De gezichtspuntschalen moeten meetinstrumenten worden die zo min mogelijk door irrelevante factoren beïnvloed worden, dus noch door waardeoordelen noch door bestaande loonverhoudingen. Men zal daartoe beoordelaars moeten kiezen die niet op de hoogte zijn van de relatief kleine niveauverschillen (in beloning of GM-score) tussen de functies die zij paarsgewijs met elkaar moeten vergelijken. Men zou bijvoor-

beeld een aantal gevorderde studenten in de (bedrijfs)psychologie of van een Technische Hogeschool kunnen gebruiken. Deze beoordelaars zouden dan een grondige training kunnen krijgen in het beoordelen van functies zoals die in de praktijk uitgeoefend worden. Een moeilijk punt daarbij zal zijn het contact dat de beoordelaar nu eenmaal onvermijdelijk met de functionaris en zijn chef(s) zal hebben. Men zal daarbij attent moeten zijn op de mogelijkheid van informatielekken betreffende niveau en beloning.

Voordat men met deze training begint is het noodzakelijk dat de gezichtspunten zoals die door de DC omschreven zijn kritisch worden bekeken. In de volgende paragrafen wordt daar nader op in gegaan.

Wellicht ten overvloede wordt hier nog even samengevat wat de verschillen zijn tussen de toepassing van de methode der paarsgewijze vergelijkingen zoals voorgesteld door DE GROOT (1953) en zoals hier beschreven. DE GROOT gebruikt deze methode om tot een operationalisatie van waardeoordelen (het rechtsbewustzijn) te komen en stelt daarom ook voor om functies in hun geheel naar waarde te laten vergelijken. Het resultaat zal dan een *rangorde* zijn van functies naar hun waarde. De bedoeling is om deze rangorde als criterium voor de GM te gebruiken.

Hier wordt de methode gebruikt om zo scherp mogelijk gedefinieerde aspecten van functies, los van enige waardering daarvan, in een intervalschaal te kunnen meten. Er moge overigens op gewezen worden dat de intervalschaal die op deze wijze bereikt kan worden schaalafstanden bevat die *per definitie* gelijk zijn. Of de waarderingen die over de functies uitgesproken worden ook 'werkelijk' een intervalschaalstructuur bezitten, of er dus werkelijk sprake is van een isomorfie in de zin van SUPPES en ZINNES (1963) is hiermee nog niet aangetoond. ("It is important to note, however, that the interval or ratio properties are obtained by definition and not by experimentation". TORGERSON 1958, p. 151)

3.4 DE KEUZE VAN GEZICHTSPUNTEN

De keuze van gezichtspunten is niet los te zien van het bepalen van afweegfactoren, al was het alleen maar omdat gezichtspunten slechts een selectie vormen uit een veel groter aantal mogelijk relevante aspecten. Alle andere mogelijke maar niet gekozen functie-eisen hebben in feite een afweegfactor nul gekregen.

Er is echter een voor de praktijk en de ontwikkeling van de methode veel belangrijker interrelatie tussen gezichtspuntkeuze en weging: men kan een gezichtspunt namelijk veel zwaarder wegen dan uit de afweegfactor blijkt door het tweemaal, maar dan onder verschillende namen, op te nemen. Men zou bijvoorbeeld een gezichtspunt 'Scholing en ervaring' naast 'Kennis' kunnen introduceren en in de definitie van deze gezichtspunten iets andere bewoordingen gebruiken. Ook hier is er een direct en nauw verband tussen het definiëren van gezichtspunten en de weging daarvan. Het spreekt vanzelf dat in geval van opzettelijke duplicering van hetzelfde gezichtspunt een zeer hoog verband zal bestaan tussen de graderingscijfers op deze 'tweeling-gezichtspunten'. Het is verleidelijk om omgekeerd te redeneren en bij het constateren van een hoge correlatie tussen twee gezichtspunten te veronderstellen dat hier in feite sprake is van gehele of gedeeltelijke duplicering; deze behoeft dan natuurlijk niet altijd opzettelijk te zijn.

Deze omgekeerde redenering vormt het uitgangspunt van de factoranalyse. Men tracht correlaties tussen variabelen te verklaren door middel van 'gemeenschappelijke factoren'. Deze factoren zijn *hypothetische* variabelen die verondersteld worden de scores op verschillende *gemeten* variabelen (zoals tests) mede te bepalen. In die zin vormen zij *gemeenschappelijke* oorzaken van testuitkomsten, vandaar de naam. Zo kan men bijvoorbeeld de correlaties tussen een aantal tests, die alle opgaven bevatten waarbij men met cijfers moet manipuleren, verklaren door de aanname van een rekenvaardigheidsfactor. Men spreekt in zo'n geval graag van een fundamentele vaardigheid (primary ability) waarin personen kunnen verschillen en die tot op zekere hoogte testscores, schoolprestaties en andere gedragingen kunnen beïnvloeden. Het is echter goed denkbaar dat niet de inhoud, de aard van het denkproces of oplossingsproces dat bij bepaalde testvragen optreedt de correlaties tussen tests veroorzaakt, maar bijvoorbeeld het feit dat in onze maatschappij met onze opvoedings- en scholingsmethoden, bepaalde zaken altijd min of meer gelijktijdig en in vergelijkbare situaties (thuis, op school) bijgebracht worden. Als kinderen om welke redenen dan ook meer of minder 'bereikbaar' zijn voor zo'n leerbeïnvloeding zal dit door de gelijktijdigheid en gelijkgesitueerdheid konsekwenties hebben op inhoudelijk sterk verschillende gebieden. TRYON schreef in dit verband reeds in 1935, handelend over hoog gecorreleerde testvragen: "Such positive correlations may arise, however, from a number of 'external agencies' and not from common conceptual content" (TRYON 1935) en hij noemt als zulke

'uiterlijke invloeden': "Systematic differences . . . in the quality of the educational and cultural medium in which the different concepts are evolved". Zo zouden echter ook correlaties tussen gezichtspunten, die elkaar niet dupliceren, kunnen ontstaan door min of meer toevallige, in de zin van: vanuit ons gezichtspunt niet relevante, factoren in de bedrijfssituatie. Organisatorische ontwikkelingen zouden bijvoorbeeld zulke irrelevante factoren (external agencies) kunnen zijn[4].

De factorstructuur die bij toepassing van een factor-analytische methode gevonden wordt, wordt nog te weinig beschouwd als één uit vele mogelijke verklaringen van correlatieve samenhangen. Factoranalyse is een te gemakkelijke techniek[5] en de verleiding is daarom groot om deze als een tovermiddel te misbruiken om orde te brengen in een chaotische veelheid van gegevens. Ze wordt hierdoor vaker als een 'trucje' gebruikt (vergelijkbaar met de goochelaar die een konijn uit een hoed tovert) dan als een wetenschappelijke methode. We zagen hierboven dat TRYON al in 1935 de pretentie van de factor-analyse heeft bestreden, dat hierdoor fundamentele processen en dimensies opgespoord zouden worden.

Sindsdien hebben vele critici met meer of minder technische argumenten op de factor-analytische stellingen stormgelopen. Een belangrijk argument is daarbij veelal geweest het gebrek aan uniciteit, een vorm van kritiek verwant aan onze opmerking dat een factorstructuur slechts één mogelijke verklaring is. Het rotatieproces is altijd enigszins subjectief geweest en pogingen om deze willekeur op te heffen door de ontwikkeling van analytische criteria voor een optimale rotatie hebben door hun veelheid het probleem slechts verschoven: nu moet de onderzoeker, op vrij willekeurige wijze, kiezen tussen varimax, quartimax, oblimax, oblimin en hoe al deze patent-geneesmiddelen nog meer mogen heten (zie bijvoorbeeld HARMAN 1960).

De algemeen gevoelde noodzaak om gevonden factoren te roteren is echter niet de enige oorzaak van het ontbreken van een unieke oplossing. Als men gemeenschappelijke

[4] Ook toevallige omstandigheden als vermoeidheid, motivatie, affectieve toestanden e.d. kunnen, als zij interindividueel verschillen maar intra-individueel een gelijkgerichte invloed uitoefenen op de scores van alle variabelen, de factorstructuur beïnvloeden. KALVERAM (1965) vond toename in de ladingen op de eerste centroïde, niet significant worden van een of meer der andere factoren en bemoeilijking van de interpretatie van de gevonden factoren. Bij de werkclassificatie kan men bijvoorbeeld de relatie taakanalist – arbeider of de toevallige stemming van elk van deze twee als zulke toevallige omstandigheden beschouwen.

[5] Dit geldt zeker nu elektronische rekenmachines zo gemakkelijk toegankelijk zijn geworden en programma's voor factor-analyse in de meeste programma-bibliotheken aanwezig zijn. Men kan dan ook "veilig stellen dat hooguit 10% hiervan (van factor-analytische bewerkingen van gegevens) relevant en niet overbodig is" (STOUTHARD 1965).

en unieke factoren onderscheidt en dus niet te maken heeft met wat men gemeenlijk componenten-analyse noemt (DE LEVE 1956) zijn er nog genoeg problemen, zoals bijvoorbeeld het bepalen van het aantal factoren. Hierover zegt WRIGLEY (1958): "More than twenty methods have been suggested (to determine the number of factors). Evidence is accumulating that the form of the rotated structure is affected by the number of factors included in the rotation. Addition of further dimensions to the common-factor space seems to lead to splitting of larger factors into smaller ones!" En elders (WRIGLEY 1957): "The fact that the factorial structure may very possibly change with increase in the sample of persons, or in the selection of tests, should warn us that any hope that factors can be regarded as real causal entities is unlikely to be realized. Factors must be regarded as operational fictions". Met de Engelse school zou men hieraan kunnen toevoegen "of als classificaties van gelijksoortige variabelen" (VERNON 1965).

Ondanks deze felle kritiek blijven stugge verdedigers argumenten leveren ten gunste van de realiteit van de factoren. Een van de belangrijkste figuren in dit kamp is CATTELL. In een betrekkelijk recent artikel (CATTELL en SULLIVAN 1962) treft men zijn geloofsbelijdenis aan: "We believe factor analysis is the most powerful method available to-day to reveal the number and nature of influences at work in producing variation in an array of complex variables". Het zou niet fair zijn om hierbij te verzwijgen dat hij althans pogingen doet om zijn geloof met empirische argumenten te staven. Ongelukkigerwijs put hij deze argumenten steeds uit zogenaamde demonstratie-analyses, factor-analyses uitgevoerd op fysische grootheden, waarvan de structuur van te voren bekend is (zie ook CATTELL en DICKMAN 1962). THURSTONE (1947) was de eerste die een dergelijke analyse uitgevoerd heeft en sindsdien is zijn 'dozen-probleem' het klassieke voorbeeld van een demonstratie-analyse geworden. Over de bewijskracht van zulke analyses schrijft OVERALL (1964) het volgende: "Demonstration analyses are easily contrived for almost any purpose. Previous analyses have been loaded to yield results corresponding to preconceived primary dimensions. It is just as easy to load examples in such a way that results do not correspond to what we conceive to be primary characteristics of objects. Logically, the demonstration analysis has its proper place as a counter example. A single exception will refute the general argument, but no small number of demonstration analyses can prove its truth".

Nog een zwakte van de factor-analytische methode wil ik hier noemen, vooral omdat deze door slechts weinigen onderkend wordt. Zelfs als het uniciteitsprobleem opgelost zou zijn, zodat elke factor-analyse slechts één mogelijk resultaat zou hebben, in de vorm van een matrix van factorladingen, dan nog is men niet waar men zijn wil. Dan weet men of denkt men te weten wat het verband is tussen de geobserveerde variabelen en een aantal factoren, al of niet beschouwd als oorzakelijk eenheden.

Wil men met deze kennis wetenschappelijk of praktisch iets doen dan zal men echter in concrete gevallen in staat moeten zijn om deze factoren te meten. In de factor-analytische terminologie: men zal factorscores moeten kunnen bepalen. Deze scores kunnen echter alleen maar geschat worden. Dit is, wat GUTTMAN (1955) noemt, de onbepaaldheid van de factor-scores.

GUTTMAN bewijst dat, gegeven een bepaalde factormatrix, zeer verschillende factor-scorematrices in het algemeen mogelijk zijn. Hoezeer deze mogelijke factoren kunnen verschillen hangt af van de multipele correlatie van de geobserveerde variabelen met een factor. Als een factor F een multipele correlatie gelijk 1,0 heeft met de variabelen (als dus deze factor in de variabelen-ruimte ligt) dan is slechts één oplossing moge-

lijk. Is deze multipele correlatie kleiner dan 1,0 dan moeten de factorscores geschat worden. De correlatie tussen twee mogelijke, maar maximaal verschillende versies van F (dezelfde factor dus!) is gegeven door

$$r = 2R^2 - 1,$$

waar R de bedoelde multipele correlatie van F met de geobserveerde variabelen is. Elke versie van F heeft daarbij dezelfde (multipele) correlatie, R, met de variabelen.

Als de multipele correlatie met een factor bijvoorbeeld 0,90 is, zijn twee realisaties van deze factor mogelijk, die onderling slechts 0,62 correleren. Wanneer een factor echter tot uiting komt in een groep variabelen met betrekkelijk lage communaliteiten zal de multipele correlatie lager zijn en hij hoeft maar beneden de 0,71 te dalen om twee versies van dezelfde factor mogelijk te maken die in het geheel niet meer gecorreleerd zijn! Vele (wiskundige) statistici geven daarom de voorkeur aan componenten-analyse, d.i. factor-analyse op de correlatie-matrix met enen in de diagonaal (zie bijvoorbeeld SCHAAFSMA 1961).

Men kan al deze technische bezwaren tegen de factor-analyse samenvatten door wederom op het uitgangspunt te wijzen: als factoren bestonden en werkzaam zouden zijn zoals de theorie dat beschrijft, zouden daaruit bepaalde correlatiestructuren resulteren. Vindt men nu deze structuren dan wordt op grond daarvan een factorstructuur aangenomen. Daarbij is zo'n factorstructuur altijd één uit vele mogelijke (zie de hierboven weergegeven kritiek) en de factor-analytische verklaring zelf alleen maar een mogelijke, geen noodzakelijke. Verdere toetsing van de nog zeer voorlopige werkhypothese, 'deze factoren verklaren de gevonden samenhangen op de beschreven wijze', blijft meestal uit.

We hebben dit soort theorievorming al meer gezien: als de psychoanalytische hypothesen en theorieën waar zouden zijn, zou men daaruit gemakkelijk de in therapeutische (en andere) situaties geobserveerde (verbale) gedragingen begrijpelijk kunnen maken. Aangezien toetsing van deze theorieën echter nog weinig plaats heeft gevonden en ten dele principieel onmogelijk is krijgen zij vaak het karakter van een geloof, een dogma (vergelijk bijvoorbeeld DE GROOT 1961, Hfdst. 2).

Dit dogmatische karakter van de factor-analytische zowel als van de psycho-analytische theorieën is door CRONBACH (1954) op geestige wijze in beeld gebracht. Hij sprak over het uitvoeren van een factoranalyse door een psycholoog als "a mystical experience in which various deities appear before his eyes and name themselves to him". Het benoemen van factoren is altijd een moeilijke zaak geweest! Wat betreft het uniciteitsprobleem: "These deities, or *Factors* as they are

called in the native language, are very numerous. Since "Factor Analysis" is a highly personal experience, each Psychometrikan has his own assemblage of *Factors*, different from that of his neighbors". Het moeilijk toetsbare karakter van de psycho-analytische theorieën (meer in het algemeen: klinische theorieën) werd als volgt beschreven: "Clinicia has developed a language of great poetic beauty and obscurity. In this obscure language, a Clinician can speak for hours on uncertain matters without fear of contradiction".

Voor psychologen lijkt het voor de hand te liggen om een factor-analyse toe te passen ter bepaling van het aantal en de aard van de gezichtspunten in een werkclassificatie-methode. De resultaten lijken er dan meestal op te wijzen dat een sterke overlap bestaat tussen de tot dan gebruikte gezichtspunten. Zo vonden LAWSHE en SATTER (1944) dat ze met 2 tot 4 (afhankelijk van het bedrijf) factoren uit konden komen. WIEGERSMA (1958) kwam tot de conclusie dat vijf factoren evenveel informatie gaven als de oorspronkelijke elf door hem in de analyse betrokken gezichtspunten. Zolang men het zo formuleert: vijf factoren geven in de onderzochte populatie (zie paragraaf 3.5) bijna evenveel informatie als elf gezichtspunten, kan men hier vrede mee hebben en kunnen de resultaten van een factor-analyse in een beperkte, niet te heterogene populatie gebruikt worden om het klassificeren efficiënter te maken (zie bijvoorbeeld LAWSHE 1945). Beschouwt men het als een theoretische fundering van de keuze van gezichtspunten dan overschat men de waarde van de factor-analytische methode.

De kritiek die tot hiertoe op de factor-analyse gegeven werd was vooral gericht op het moeilijk toetsbare karakter van de hypothesen en op het ontbreken van een unieke oplossing zelfs binnen de factor-analytische school. Er is echter nog een aspect van de factor-analyse dat ook bij het verwerpen van onze rangnummer-methode een rol speelde: de afhankelijkheid van de resultaten, in dit geval correlaties (en dus factoren) van de populatiesamenstelling. Een voorbeeld uit de economische sfeer moge deze afhankelijkheid illustreren. Stel dat de prijs van personenauto's bepaald zou worden onder andere door Grootte en Motorvermogen. Het zal bij een correlatieberekening blijken dat het verband tussen deze twee 'gezichtspunten' bij normale personenauto's zeer hoog is. Men zou geneigd kunnen zijn hieruit de conclusie te trekken dat Grootte en Motorvermogen in feite terug te voeren zijn op eenzelfde factor en dat deze factor dubbel betaald wordt. Neemt men echter in de steekproef een groot aantal sport-

wagens op dan wordt de populatie heterogener en zal blijken dat de samenhang slechts gold voor de homogene populatie.

Deze invloed van de populatiekeuze op correlaties en daardoor op de structuur van de gevonden factoren is een bekend verschijnsel. De conclusies uit factor-analytische onderzoekingen getrokken kunnen niet zonder meer over populaties van meetobjecten (zoals personen of functies) gegeneraliseerd worden[6]. Dit is overigens een beperking die voor alle statistische methoden geldt. Bij het presenteren van nieuwe factoren krijgt deze beperking echter niet altijd voldoende aandacht. Dit is vooral daarom ernstig omdat factoren in de psychologie nog steeds door sommigen gezien worden als afbeeldingen van werkelijke eigenschappen van de gemeten objecten (zie bijvoorbeeld OVERALL 1964).

Wordt het feit van de populatieafhankelijkheid niet altijd voldoende benadrukt, aan de andere kant wordt dit zelfde feit door aanhangers van de factor-analyse nogal eens misbruikt. Doordat de factor-structuur zo zeer van de populatiesamenstelling afhangt, kan men in fictieve voorbeelden elke gewenste factorstructuur te voorschijn roepen door variatie van de populatie. De zo 'gevonden' maar in feite *geïnduceerde* structuur wordt dan als argument voor een bepaalde opvatting gebruikt. We hebben dit hierboven reeds gesignaleerd en OVERALL's kritiek aangehaald op het gebruik van zulke voorbeelden, niet als tegenvoorbeeld maar als een positief argument voor een algemene bewering.

Het voorbeeld van de prijsvorming bij personenauto's zou men als een, fictief, tegenvoorbeeld kunnen beschouwen, dat laat zien dat correlaties niet noodzakelijkerwijs op gemeenschappelijke factoren behoeven te wijzen.

3.5 INHOUDELIJKE ANALYSE VAN GEZICHTSPUNTEN

Overigens blijft met dit al de vraag onbeantwoord naar een juiste keuze van de gezichtspunten, die aan de ontwikkeling van gezichtspuntschalen met intervalkarakter vooraf moet gaan. Het is dan wel niet noodzakelijk dat twee gezichtspunten die hoog gecorreleerd zijn, zoals

[6] Met andere woorden de factorstructuur is niet invariant over populaties van objecten. Men verwarre dit niet met de factor-invariantie waarover KAISER (1958) schrijft in zijn artikel over de varimax. KAISER doelt namelijk op invariantie over veranderingen in de testbatterij.

bijvoorbeeld Kennis en Zelfstandigheid (de correlaties liggen meestal in de grootte-orde van 0,90) "als vrijwel identiek moeten worden beschouwd" (WIEGERSMA 1958, p. 204), uitgesloten is het ook niet. Het is zelfs zeer waarschijnlijk dat in dit speciale geval inderdaad sprake is van een vérgaande duplicering. De Deskundigencommissie (1959, p. 34) zegt namelijk over het gezichtspunt Kennis: "Onder kennis wordt zowel theoretische kennis als kennis van de dagelijkse praktijk van het werk begrepen". Even verder luidt het dan: "Het toepassen van kennis, waaronder vallen de behandeling van problemen en het kiezen uit verschillende mogelijkheden, wordt derhalve niet tot kennis gerekend, doch komt tot uitdrukking in andere gezichtspunten, met name in het gezichtspunt Zelfstandigheid". Nu is het niet gemakkelijk om onderscheid te maken tussen kennis en het toepassen daarvan. Het is moeilijk voorstelbaar dat voor een functie kennis vereist wordt, maar dat deze niet toegepast behoeft te worden. Bij de beschrijving van het gezichtspunt Zelfstandigheid wordt vooral de nadruk gelegd op de mate waarin de functionaris vrij is om te kiezen uit verschillende mogelijkheden. Hier lijkt een geheel nieuw aspect naar voren te komen. Onder de eigenschappen die door middel van dit gezichtspunt tot uitdrukking zouden komen vallen het meest op: begrip, inzicht, oordeel, voorstellingsvermogen, vindingrijkheid, feeling (op. cit., p. 38). Dit zijn begrippen die vrijwel iedere psycholoog als kenmerkend voor intelligentie zou beschouwen, zeker wanneer men de laatste twee nog zou schrappen. Het onderscheid tussen kennis en intelligentie is echter zeer onduidelijk. De DC noemt in het zelfde rijtje eigenschappen ook: besluitvaardigheid en initiatief. Dit is weer meer in overeenstemming met het accent dat de vrijheid van beslissen in dit gezichtspunt krijgt.

De DC heeft onder het gezichtspunt Zelfstandigheid aspecten willen vatten die in de beschrijving van het kennisaspect niet aan de orde komen. Zij is daarin niet goed geslaagd. Het moeten beslissen in onzekere situaties, waar voorschriften of andere gegevens met betrekking tot de juiste keuze van de te volgen gedragslijn ontbreken, doet een beroep op wat men zelfstandigheid zou kunnen noemen. Dit eist niet in de eerste plaats intelligentie of kennis, maar een bereidheid en een vermogen om verantwoordelijkheid te aanvaarden in riskante situaties. Alleen wanneer men het gezichtspunt Zelfstandigheid vanuit dit uitgangspunt tracht te omschrijven kan men verwarring met Kenniseisen voorkomen.

Het nut van het berekenen van correlaties tussen gezichtspuntgraderingen ligt niet in de mogelijkheid om bijvoorbeeld via een factor-

analyse een aantal fundamentele en onderling onafhankelijke functie-eisen te isoleren. Men kan echter correlaties, vooral wanneer deze uitzonderlijk hoog zijn, wel benutten als waarschuwingssignalen, waardoor men soms geattendeerd kan worden op begripsmatig slecht onderscheiden gezichtspunten. De uiteindelijke beslissing of twee gezichtspunten voldoende overeenstemmen om ze als één en hetzelfde gezichtspunt te mogen beschouwen mag echter nooit uitsluitend op correlaties gebaseerd zijn. Een analyse zoals hierboven gegeven van de gezichtspunten Kennis en Zelfstandigheid is in zo'n geval een noodzakelijke aanvulling.

Voor de verdere gang van het betoog wordt nu aangenomen dat een aantal gezichtspunten is gekozen na kritische statistische maar vooral ook inhoudelijke analyse. Zeer waarschijnlijk zal daarna het gezichtspunt Zelfstandigheid zoals dat nu in de GM gehanteerd wordt opgenomen zijn in het gezichtspunt Kennis en wellicht zal daarnaast dan een gezichtspunt Zelfstandigheid of Verantwoordelijkheid omschreven zijn. Of dit nieuwe gezichtspunt nog van het huidige GM-gezichtspunt Afbreukrisico te onderscheiden zal zijn moet nog blijken. Voor elk van deze nieuw gedefinieerde of gehandhaafde gezichtspunten kunnen dan schalen ontwikkeld worden waarop een aantal sleutelfuncties de verschillende schaalpunten representeren. Als dit alles geschied is komt het probleem van de weging weer aan de orde.

3.6 HET PROBLEEM VAN DE AFWEEGFACTOREN

De Deskundigencommissie (1952, p. 16 en 1959, p. 25) heeft tot tweemaal toe gesteld dat de afweegfactoren voorlopig zijn en dat deze aangelegenheid nog in onderzoek is. Ook hier heeft de DC door op twee gedachten te hinken een ondoorzichtige situatie geschapen. Enerzijds heeft ze de afweegfactoren beschouwd als een maat voor de "relatieve waarde" der gezichtspunten (bijvoorbeeld: "deze relatieve waarde wordt tot uitdrukking gebracht door zogenaamde afweegfactoren", Deskundigencommissie 1959, p. 25).

Anderzijds werden ze gebaseerd op "zo goed mogelijk benaderde praktijkverhoudingen" (Stichting van den Arbeid 1952). Deze opvatting is meer in overeenstemming met de sterk op de bestaande loonverhoudingen gerichte opzet van de methode. Men heeft waarschijnlijk gehoopt uit de aanpassing van de GM-uitkomsten aan de bestaande loonstructuur af te kunnen leiden of deze afweegfactoren inderdaad de

relatieve waarde der gezichtspunten, zoals deze zich in de praktijk ontwikkeld heeft, juist tot uitdrukking brengen.

Men heeft zich daarbij kennelijk niet gerealiseerd dat een maximale overeenkomst (uitgedrukt als een correlatie-coëfficiënt) tussen GM-gegevens (T-scores) en loonverhoudingen in de meeste gevallen bereikt zal worden bij toepassing van afweegfactoren die wel heel moeilijk als uitdrukking van relatieve waarde opgevat kunnen worden. Een extreem voorbeeld hiervan levert ITZ (1960) die door middel van multipele regressie tot een negatief gewicht (!) voor zwaarte van het werk komt. Er is een zekere tegenstelling tussen het gebruik van afweegfactoren om aanpassing te bereiken aan een of ander criterium (bestaande loonverhoudingen) en de opvatting dat deze wegingsfactoren de relatieve waarde der gezichtspunten tot uitdrukking brengen. Deze tegenstelling kan het duidelijkst aan het licht komen wanneer men langs systematische weg (bijvoorbeeld door multipele regressie) de aanpassing tracht te bereiken, zoals ITZ gedaan heeft. Gaat men minder systematisch, meer volgens een methode van 'trial and error' te werk, zoals de DC gedaan heeft (zie LIJFTOGT 1966) dan valt deze tegenstelling minder op omdat men dan door het gebrek aan een dwingende systematiek ruimte overhoudt voor een compromis-oplossing.

Wanneer hier gesproken wordt over 'aanpassing aan de praktijkverhoudingen' wordt daarmee bedoeld de ontwikkeling van een valide voorspeller. De voornaamste functie van zo'n voorspeller is een zuiver pragmatische. Er is nog een geheel andere aanpassing denkbaar. Wanneer een wetenschappelijke theorie een belangrijk aspect van de werkelijkheid op adekwate wijze in kaart brengt, "de verschijnselen zo goed mogelijk dekt" (DE GROOT 1961, p. 43) kan men ook van aanpassing spreken. Het is deze vorm van aanpassing waar DA SILVA (1951) op gedoeld moet hebben toen hij stelde dat de uitkomsten van de GM moeten "kloppen met de – met verstand bekeken – practijk". DE GROOT (1953) heeft dit idee verder uitgewerkt en zegt terecht: "Het hypothetische karakter en de *noodzaak tot toetsing aan goed gekozen criteria komen hiermee in een ander licht te staan. . .*" (cursivering van de schrijver). Inderdaad, de criteria hebben hier niet meer de functie die criteria bij voorspellers hebben. Het gaat hier om toetsing van een theorie en niet om validering van een voorspeller.

Het is niet zo verwonderlijk dat de DC dit schijnbaar subtiele, maar in wezen fundamentele, onderscheid in het gebruik van het begrip 'aanpassing aan de praktijk' niet heeft gemaakt. Het is wel te be-

treuren, gezien de begripsverwarring die daardoor ontstaan is en die nog steeds doorwerkt in theoretische discussies rondom de methode. Het komt steeds weer neer op het al eerder gemaakte onderscheid tussen theorievorming enerzijds en het ontwikkelen van voorspellers met het oog op duidelijk omschreven praktische doelstellingen anderzijds (vergelijk ook COMREY 1951, DE GROOT 1962). Enigszins analoog daarmee loopt het onderscheid tussen meten en voorspellen (DE GROOT 1961) en het onderscheid dat MEEHL (1954) maakt tussen, wat hij noemt "structural (or analytic)" en "discriminative (or validating) use of statistics". MEEHL noemt als prototype van het eerst genoemde factoranalyse; als een goed voorbeeld van het tweede zou men de reeds vermelde multipele regressietechniek kunnen noemen. Vergelijking van de interpretatiemogelijkheden bij deze twee laatst genoemde technieken brengt duidelijk het verschil naar voren: een factorlading wordt beschouwd als een aanwijzing voor een mogelijke causale verklaring; tevens geeft de grootte van de lading een idee van de relatieve invloed van de als oorzaak gedachte factor. Een factoranalyse is dan ook meestal een methode om tot theorievorming te komen. Een regressie-coëfficiënt in een multipele regressievergelijking laat zo'n theoretische interpretatie niet toe.

Het gebruik van regressiecoëfficiënten als afweegfactoren is niet aanvaardbaar. De GM is ook, of misschien wel in de eerste plaats, een politiek communicatiemiddel en de waarden van de afweegfactoren moeten daarom aanspreken als redelijk. Men kan in zo'n geval geen negatieve gewichten verdedigen met correlatie-technische argumenten. De huidige situatie is echter ook onbevredigend omdat het niemand duidelijk is wat de afweegfactoren betekenen, wat nu precies hun functie is en in hoeverre zij deze functie goed vervullen. Enerzijds zijn de afweegfactoren gebruikt om aanpassing aan de praktijk in pragmatisch-voorspellende zin te bereiken, anderzijds wilde de DC het belang van elk gezichtspunt in zijn afweegfactor tot uitdrukking brengen. Deze twee functies van wegingsfactoren sluiten elkaar echter uit. Het is vooral ook hierom dat wij, uitgaande van de noodzaak om meten en waarderen in de GM gescheiden te houden, de opvatting als zou de GM een voorspeller (moeten) zijn, bestreden hebben.

Als nu een afweegfactor het belang van, dat is de waardering voor, een gezichtspunt tot uitdrukking brengt, dan moeten gezichtspunten met grote afweegfactoren 'meer bijdragen tot', 'meer invloed hebben op' het totaalpuntental, T, dan minder belangrijke gezichtspunten.

Om te kunnen controleren of dit inderdaad het geval is zal een definitie van 'bijdrage' of 'invloed' beschikbaar moeten zijn. Zo'n definitie is echter verre van gemakkelijk te geven.

Elders hebben wij een overzicht gegeven van gangbare definities van het bijdragebegrip (HAZEWINKEL 1964). Het bleek dat deze definities sterk uiteen lopen en tot zeer verschillende conclusies kunnen leiden omtrent de feitelijke invloed van de gezichtspunten. Men kan ze indelen in twee hoofdklassen: definities waarbij op een of andere wijze rekening is gehouden met de correlaties die tussen de gezichtspunten bestaan en een definitie die alleen gebaseerd is op de spreiding van het betrokken gezichtspunt. Volgens deze laatste definitie is de bijdrage gelijk aan de standaarddeviatie van het gewogen graderingscijfer.

Het is echter ongewenst om zonder meer de traditionele redenering te volgen, volgens welke het belang of de invloed van een variabele altijd toeneemt met de spreiding. Als men 'invloed' ziet in de statistische zin van correlatie met het eindresultaat is deze opvatting ongetwijfeld juist (men vergelijke HAZEWINKEL 1964). Als men echter invloed definieert door: het aantal T-punten dat een functie meer krijgt wanneer deze een categorie hoger geklasseerd wordt op een gezichtspunt, dan is alleen de afweegfactor van belang. Om deze definitie zin te geven is het echter nodig dat een punt verschil op verschillende gezichtspunten op een of andere wijze vergelijkbaar is, bijvoorbeeld door de bovenbeschreven methode van de 'even vaak waargenomen verschillen'.

Een zodanige definitie van invloed van een gezichtspuntgradering is meer in overeenstemming met ons standpunt dat de afweegfactoren slechts de functie hebben om de 'munteenheden', waarin de overigens gelijksoortige gezichtspuntgraderingen zijn uitgedrukt, gelijk te maken. Een statistisch invloeds- of bijdragebegrip is hier volkomen irrelevant en een statistische benadering van het probleem van de bepaling der afweegfactoren is dan ook niet aangepast aan aard en functie van de methode. De vraag wordt dan uiteraard welke benadering wèl adekwaat zou zijn om tot geschikte afweegfactoren te komen. Onder geschikt moeten wij verstaan afweegfactoren die de verschillende graderingscijfers vergelijkbaar maken (in dezelfde munteenheid omzetten) en die toch ook resultaten geven die enigszins in overeenstemming zijn met de bestaande loonstructuur.

We hebben gezien dat een regressietechniek niet bruikbaar is. Men zou kunnen denken dat de theoretisch ideale methode zou zijn om – uitgaande van de opvatting dat de GM meet en niet voorspelt – on-

geacht de resultaten de relatieve waarde van de gezichtspunten vast te stellen en deze vervolgens weer te geven door middel van een stelsel afweegfactoren.

Er zijn twee bezwaren tegen deze procedure:

1. Het vaststellen van de relatieve waarden zoals bovenbedoeld kan niet in abstracto geschieden. Men realiseert zich pas welke waarde men aan een gezichtspunt toekent door een analyse van het waarderingsproces met behulp van concrete functies. Dit houdt echter reeds in dat men niet geheel 'ongeacht de resultaten' te werk kan gaan.
2. Een systeem van afweegfactoren heeft een zekere overeenkomst met een axiomastelstel, dat algemene regels vaststelt volgens welke we functies moeten beoordelen. Blijkt achteraf bij toepassing van zo'n systeem dat de resultaten niet aanvaardbaar zijn dan moeten de axioma's, in dit geval de afweegfactoren, in revisie. Men vergelijke wat LUCE en RAIFFA (1957, p. 122) schrijven over axioma's die uitgangspunten vormen voor 'welfare functions' (vergelijk paragraaf 3.1): "If, however, certain of them (arbitration schemes compatible with the axioms) still yield arbitrated results which we deem 'unfair', then one must search for further stipulations as to the meaning of 'fairness' to add to our axioms and thereby eliminate those we consider undesirable".
   In het geval van de afweegfactoren kan een dergelijk onaanvaardbaar resultaat leiden tot het formuleren van nieuwe gezichtspunten, het schrappen van oude en/of het bijwerken van sommige van de afweegfactoren. Dus ook na een, voorlopige, vaststelling van de afweegfactoren moet men deze toch weer toetsen aan de resultaten.

3.7 THEORIEVORMING EN EXPLICITERING

Bovenstaande punten zijn niet in strijd met de opvatting naar voren gebracht door DA SILVA en later door DE GROOT (1953) dat de GM opgevat kan worden als een theorie die, als elke theorie, getoetst moet worden aan de praktijk. Maar de GM wordt ook praktisch gehanteerd en wel als *norm*. Zij is daarin vergelijkbaar met bijvoorbeeld een examentoets. DE GROOT (1964) zegt over zo'n toets en meer in het bijzonder over de norm in de vorm van een minimumscore (die men moet halen om te 'slagen'), dat zij behoort te berusten op de inhoud van het gemetene. Dat wil vooral zeggen dat niet meer relevant is

hoeveel leerlingen bij een bepaalde normdefinitie zouden slagen. Het gaat om de inhoud van de geleverde prestatie en niet om hoeveel leerlingen deze prestatie kunnen leveren. De norm behoort ontdaan te worden van zijn "fatale statistische basis" (op. cit., p. 439). Men zou letterlijk hetzelfde kunnen zeggen over graderingscijfers en afweegfactoren: het gaat om wat deze functie waard is en niet om het (toevallige) aantal functies dat lager gewaardeerd werd en het gaat om wat dit gezichtspunt waard is en niet om de (toevallige) spreiding van functies op dit gezichtspunt.

MILLER (1964) geeft een bijzonder sprekende illustratie van de noodzaak om bij het stellen van (minimum-)eisen absolute, dus niet-statistische normen te hanteren. Hij schrijft: "The State of Illinois in its wisdom has determined for the safety of its citizens that no one can be granted the license to drive a motor car unless he can identify without error each of the seven roadside signs that alert him to impending danger. *No matter what the mean score or range of scores among candidates, an individual will remain unlicensed until he can identify all seven*". (cursivering van de schrijver) Overigens betekent dit niet dat men de statistische resultaten mag negeren. MILLER heeft bijvoorbeeld bij examens bij wijze van experiment absolute normen ingevoerd en hij vond dat tussen 85% en 100% van de kandidaten zakten. Hij vroeg zich toen terecht af wat hier fout gegaan was. "Whatever the answer, the most important fact is that the questions are being asked rather than obscured by a method of scoring and reporting that would have allowed distribution rather than absolute achievement to determine what was acceptable and what unacceptable" (op. cit., p. 295).

Statistische gegevens worden dus wel gebruikt, zoals wij ook de hoge correlatie tussen kennis en zelfstandigheid gebruikt hebben, als een waarschuwingssignaal dat er misschien iets mis is. Ze mogen echter niet als enige basis voor een normstelling aangewend worden.

Zelfs deze stelling is in zijn absoluutheid nog wel aanvechtbaar. Als een bepaald niveau op een gezichtspunt vrijwel nooit toegekend wordt spreekt men terecht van "zeldzaam moeilijk". Dit zeldzame, een statistisch begrip, geeft aan het niveau van dit gezichtspunt een extra gewicht. Een analoge redenering kan men opstellen met betrekking tot extreme gevallen van spreiding: als een gezichtspunt vrijwel niet differentieert tussen functies is dat gezichtspunt *om praktische redenen* niet zinvol.

Impliciet spelen deze overwegingen bij het stellen van (quasi-) absolute normen ongetwijfeld een corrigerende rol. Hier wordt slechts

gewaarschuwd voor een normstelling die uitsluitend en alleen op statistische gegevens gebaseerd is.

Het waarderen in de breedste zin van het woord, het evalueren en interpreteren van meetuitkomsten door deze te relateren aan een populatie van uitkomsten is in de gedragswetenschappen gebruikelijk en vaak ook zinvol en nodig. Zodra men echter uitsluitend op deze wijze normen gaat stellen komt men tot absurde resultaten.

Als men de GM beschouwt als een normstellend instrument moet men dus een uitsluitend statistische benadering verlaten. Maar er is nog een gevolg van deze opvatting omtrent de GM: DE GROOT (1964) noemt als een voordeel van zijn 'kernitemmethode' dat hierdoor vage onderwijsdoelstellingen geëxpliciteerd worden. Men beschouwe de GM niet als een theorie maar als een explicitering van een systeem van waarden, in de zin waarin LUCE en RAIFFA's axioma's dat ook zijn. Men zou kunnen tegenwerpen dat deze axioma's toch in feite de basis vormen van een theorie. Inderdaad is dit het geval. De theorievorming is echter een uitwerking van de konsekwenties van deze axioma's en in het geval van de 'social welfare function' het opsporen van mogelijke interne tegenstrijdigheden. Daarmee is echter de explicitering zelf niet tot theorie geworden.

Er zijn belangrijke overeenkomsten tussen explicitering en theorievorming. Hierdoor wordt het soms moeilijk om deze twee processen duidelijk te onderscheiden. Bij explicitering is er, evenals bij theorievorming, een voortdurende wisselwerking tussen formulering (stelling, axioma, norm-operationalisatie) en empirische werkelijkheid. Bij de theorievorming worden de konsekwenties van de hypothesen in de vorm van voorspellingen getoetst (vergelijk DE GROOT's "Cyclus van het empirisch-wetenschappelijke onderzoeken", DE GROOT 1961). Bij de *explicitering* van waarden en normen worden de implicaties van de formuleringen getoetst aan eigen oordeel en aan dat van anderen. Bij duidelijke discrepanties vindt herziening van theorie of expliciete formulering plaats.

De verschillen tussen deze twee processen zijn voor de GM echter veel belangrijker. Een theorie is descriptief[7]. Theorievorming vooronderstelt een reeds aanwezige werkelijkheid die min of meer adekwaat in kaart gebracht kan worden. Explicitering brengt iets nieuws. Uit

[7] We zien hier even af van de mogelijkheid dat een theorie ook (voorwaardelijk) normatief kan zijn (LUCE en RAIFFA 1957). Ook bij de vorming van zo'n theorie valt het bepalen van doelstellingen en waarde van bepaalde gebeurtenissen (outcomes) buiten de theorie.

vage, impliciete, niet of nauwelijks geverbaliseerde doelstellingen, waarden of normen wordt iets nieuws ontwikkeld dat scherp omlijnd, duidelijk geformuleerd is. Discussie over een theorie is, idealiter, altijd een rationele[8]: is de theorie logisch verenigbaar met de feiten? Welke feiten kunnen en moeten nog verzameld worden? Discussie over een explicitering heeft in laatste instantie een irrationeel karakter. Door de explicitering wordt de norm niet *toetsbaar* zoals een hypothese toetsbaar wordt, ze wordt slechts *discutabel*. Er ontstaat niet kennis, maar "kansen op vruchtbare (zelfkritiek en) kritiek van anderen zijn verbeterd" (DE GROOT 1964, p. 439). Voor een instrument als de GM: de kansen op een vruchtbare discussie zijn verbeterd.

Er is wetenschappelijke discussie mogelijk over de vorm van operationalisatie (kernitemmethode, overwegingen van schaaltechnische aard bij de GM) en over mogelijke tegenstrijdigheden binnen een geëxpliciteerd waardesysteem (ARROW's 'impossibility theorem', LUCE en RAIFFA 1957, p. 339). Er is geen wetenschappelijke discussie mogelijk over de inhoud van een explicitering. Men komt hier op het terrein van de (gezamenlijke) beleidsbepaling bijvoorbeeld in de vorm van loononderhandelingen.

De opvatting, als zou het vaststellen van de afweegfactoren een expliciteringsproces en geen theorievorming zijn, is geen vrijblijvende. Dit houdt immers in dat hier sprake is van beleidsvorming en beleid moet binnen bepaalde grenzen flexibel zijn. Verstarring van de loonstructuur is uit den boze. De afweegfactoren moeten punt van discussie blijven en wel zo frekwent als met het oog op een stabiele loonontwikkeling gewenst is. Een tweede implicatie van het beleidskarakter van de afweegfactoren is dat niet een deskundigencommissie maar een politieke commissie (zoals bijvoorbeeld een looncommissie van de Stichting van de Arbeid) de verantwoordelijkheid voor deze explicitering in de vorm van afweegfactoren op zich behoort te nemen. De Deskundigencommissie voor Werkclassificatie heeft deze verantwoordelijkheid inderdaad formeel niet op zich genomen, maar de reden hiervan is niet altijd duidelijk gesteld. Uit wat de Stichting van den Arbeid (1952) hierover stelt kan men namelijk gemakkelijk opmaken dat het voornaamste motief was dat de afweegfactoren nog niet af

[8] Bij de operationalisatie van vage begrippen als angst, rigiditeit, e.d. is ook sprake van iets nieuws dat gecreëerd wordt, ontwikkeld vanuit een vaag idee. Het verschil ligt echter in de methode van toetsing: een theorie toetst men aan feiten die verzameld worden tijdens een niet door de theorie beïnvloed observatie-proces; een explicitering 'toetst' en verfijnt men in een discussie over meningen en opvattingen.

waren, nog 'in ontwerp'. Men kan echter ook tot de conclusie komen, dat deze *in principe* niet door loontechnici bepaald dienen te worden. Zij zijn overigens wèl door de DC (loontechnici dus) bepaald en niet door loonpolitici en naderhand zijn zij niet meer gewijzigd. Men vergelijke wat LIJFTOGT (1966, p. 42-43) over deze zaak schrijft onder het veelzeggende paragraaf-hoofd *Onzekerheid in beleid.*

3.8 KEUZE EN WEGING VAN GEZICHTSPUNTEN

De conclusie die zich naar aanleiding van het in de voorgaande paragrafen behandelde opdringt is wetenschappelijk weinig bevredigend. Er bestaat geen systematische en wetenschappelijk verantwoorde methode om gezichtspunten te kiezen en afweegfactoren vast te stellen. Men zal het eens moeten worden omtrent de bruikbaarheid van een aantal gezichtspunten, hun onderlinge overlap en de mate waarin zij gezamenlijk alle relevante aspecten van functie-eisen dekken. Er is echter geen wetenschappelijke methodiek die hierbij veel steun kan geven. De waarde van factor-analyse en multipele regressie-technieken voor dit probleem is sterk overschat.

De functie van een analyse als hier gemaakt is vooral om een zekere verheldering te geven, om aan te tonen welke veronderstellingen, vaak impliciet, gemaakt werden of zouden moeten worden en om te laten zien in hoeverre deze veronderstellingen gerechtvaardigd zijn. Hierdoor wordt duidelijk welke aspecten van de methode voor wetenschappelijke analyse toegankelijk zijn en welke aspecten veel meer een beleidskarakter dragen waarover de wetenschap het laatste woord niet kan en mag spreken.

Tot slot nog enkele opmerkingen over het vaststellen van afweegfactoren die beschouwd kunnen worden als een goede afspiegeling van algemeen geaccepteerde waardeverhoudingen tussen de gezichtspunten. Er is hier nogal kritiek geleverd op de weinig systematische wijze waarop door de DC de afweegfactoren zijn vastgesteld (men vergelijke LIJFTOGT 1966, Hfdst. IV). Het is, gezien de moeilijkheid van het probleem, niet helemaal fair om de DC te verwijten dat ze niet meer systematisch te werk is gegaan. Ook hier kan geen voorstel tot een kant en klare systematiek gegeven worden. Als belangrijkste bezwaren tegen de werkwijze van de DC blijven echter:

1. dat zij niet wist wat zij wilde: aanpassing aan de praktijk in de voor-

spellende zin of aanpassing aan de praktijk in de zin van een theorie? (Men bedenke dat een complex beschrijvingsmodel ook een soort theorie is.)
2. dat zij haar beslissingen baseerde op veel te weinig functies (22! Zie LIJFTOGT 1966, p. 46) en
3. dat zij ondanks dit alles de tenslotte gekozen gezichtspunten in feite op vrij dogmatische wijze heeft gepresenteerd als de juiste.

Van enige aarzeling of relativering van de waarde van juist deze keuze is in de publikaties van de DC weinig te bemerken. Toch zou een iets voorzichtiger presentatie de discussie rond de keuze van de gezichtspunten en daarmee rond het wegings- en dus waarderingsaspect van de methode hebben vergemakkelijkt, hetgeen met name met betrekking tot de toepassingsmogelijkheden van de GM voor niet-handarbeidersfuncties geen overbodige luxe zou zijn geweest.

Wij zouden hier allereerst als eis willen stellen dat een explicitering van waarden getoetst wordt aan een groot aantal functies. Bovendien willen we hier pleiten voor een sterke relativering van de waarde van elke explicitering en vooral ook voor een explicitering die hoogstens betrekking heeft op de functies binnen één bedrijfstak. Het lijkt nauwelijks aanvaardbaar om zeer uiteenlopende functies met eenzelfde waardesysteem te benaderen, al was het alleen maar omdat schijnbaar dezelfde gezichtspunten bij sterk verschillende functies een andere kleur krijgen. Het wordt hierom ook op zijn minst onwaarschijnlijk geacht dat éénzelfde methode tegelijk functies van zeer verschillend niveau kan waarderen.

HOOFDSTUK IV

Het experimentele onderzoek

Het betoog in de voorafgaande hoofdstukken kan men beschouwen als een theoretische analyse, die soms een terreinverkenning en een ontdekkingstocht werd: na de op theoretische gronden gebaseerde conclusie dat de GM als meetinstrument niet voldoet, werd gezocht naar wegen en middelen om uit de impasse te geraken.

Theoretische argumenten missen voor velen de overtuigingskracht van feiten. Mede hierom en tevens om aan te tonen, dat de experimentele methode van de psychologie ook op het ondoorzichtige en gecompliceerde gebied van de werkclassificatie met succes toegepast kan worden, werd een experiment ontworpen. Hierbij werd de hypothese getoetst dat het graderen van functies volgens de GM mede beïnvloed wordt door externe factoren, dat zijn factoren die niet in een direct verband staan met de functie-eisen en die daardoor soms met deze in strijd kunnen zijn.

Om tot een operationalisatie van het begrip 'externe factoren' te komen werd gefingeerde informatie omtrent het globale niveau van de functies aan de functiebeschrijvingen toegevoegd. De hypothese dat deze informatie het graderen zou beïnvloeden, zelfs als deze afweek van de informatie vervat in de functiebeschrijvingen, werd bij één groep functies en twee categorieën van proefpersonen bevestigd. Bij een andere groep functies bleek toetsing niet mogelijk, omdat de gefingeerde informatie niet voldoende afweek van de interpretatie die de proefpersonen, ook zonder deze informatie te kennen, aan de functiebeschrijvingen gaven.

Op meer overtuigende wijze dan alleen door middel van een theoretische analyse mogelijk is, werd hiermee aangetoond dat de GM als meetinstrument op fundamentele punten te kort schiet.

4.1 INLEIDING

Er is veel gepraat en geschreven over werkclassificatie in het algemeen en over allerlei werkclassificatie-systemen, zoals de GM. Men is bij al dit praten en schrijven uitgegaan van een abstractie of op zijn best van een hoeveelheid uitkomsten van toepassing van een methode. Wat werkclassificatie nu precies is, dat wil zeggen wat er gebeurt als de methode toegepast wordt, is veel minder besproken. Wel is er een groot aantal artikelen en publikaties die beschrijvingen geven van de uiterlijke gang van zaken. 'Praktijk van de werkclassificatie' (Werclassificatie 1962) is een typisch voorbeeld van zo'n beschrijving. 'Werkclassi-

ficatie' (Nederlands Verbond van Vakverenigingen) gaat iets verder, maar heeft met de andere hier bedoelde publikaties gemeen dat het een normatieve beschrijving geeft, een voorschrift van hoe men te werk moet gaan. Wat in feite gebeurt als een taakanalist de methode toepast, welke factoren dit proces beïnvloeden, zijn vragen die nooit beantwoord werden, laat staan dat er een experimenteel onderzoek naar gedaan werd. RITTER (1958, paragraaf 4.2), bijvoorbeeld, heeft een analyse gemaakt van het graderen en hij komt daarbij tot vier mogelijke methoden, maar zijn analyse was 'van achter de schrijftafel'.

Empirisch onderzoek met betrekking tot de GM is tot nu toe vrijwel geheel beperkt gebleven tot het bepalen van intercorrelaties tussen graderingscijfers en het toepassen van factor-analyses en multipele regressie-technieken (bijvoorbeeld SANDEE en RUITER 1957, WIEGERSMA 1958, ITZ 1960). Ook in de Verenigde Staten van Amerika is het meeste werk gedaan op dit gebied: LAWSHE (o.a. 1944 met SATTER, en 1945) heeft een aantal publikaties gewijd aan factor-analyses van gezichtspunten. De betrouwbaarheidscoëfficiënten van verschillende verkorte versies van bestaande methoden werden berekend en vergeleken met de onverkorte versies door o.a. CHESLER (1948, 1949) en weer door LAWSHE (met FARBRO 1949) en vergelijkingen zijn gemaakt tussen graderingen door werknemers en werkgevers (LAWSHE en FARBRO 1949). Steeds werd hier gekeken naar de resultaten van de graderingen; hoe deze tot stand kwamen en welke factoren deze beinvloedden, vragen die nauw verband houden met de waarde en de betekenis van de methode, werden nauwelijks of niet gesteld.

Er zijn ons slechts twee onderzoekingen bekend die meer in de richting van een experimenteel onderzoek gaan. Zo werden door SATTER (1949) twee methoden vergeleken, een met paarsgewijze vergelijkingen en een andere waar gebruik gemaakt werd van een soort graderingsschema's, waar echter meer gedetailleerde en concrete beschrijvingen gegeven werden van de verschillende niveaus der gezichtspunten dan bij de GM het geval is. Men kan hier dus spreken van een experimenteel onderzoek naar de invloed van graderingsmethoden. Voor zover ons bekend is het slechts MYERS (1958) geweest die geprobeerd heeft langs experimentele weg verklaringen te vinden voor de in empirische onderzoekingen gevonden resultaten. Zijn vraagstelling was enigszins verwant aan de onze. Hij werd ook getroffen door wat hij noemt de "forcing of job ratings", wat wij steeds aangeduid hebben als "het toewerken naar een bepaald gewenst eindresultaat". Zijn experimentele opzet verschilt echter sterk van de onze.

door de proefpersonen enige informatie te verschaffen omtrent het niveau van de functies die ze moesten graderen. Om te zorgen dat de analyse van de functies en daardoor de functiebeschrijvingen door deze informatie niet beïnvloed kon worden, kregen de proefpersonen tot taak functies te graderen op grond van functiebeschrijvingen die door andere taakanalisten waren opgesteld. De informatie omtrent het niveau van de functies was van dien aard dat er geen enkel verband bestond met het niveau dat in de praktijk, door de opstellers van de functiebeschrijvingen, aan de functies was toegekend. Soms kwam deze informatie overeen met de praktijkcijfers, soms waren er ook vrij aanzienlijke verschillen tussen experimentele informatie en in de praktijk gegeven T-score.

4.3 HET VOORONDERZOEK

Toen bovengenoemde hypothese eenmaal geformuleerd was, werd in een vooronderzoek nagegaan of een experimentele toetsing mogelijk was. Een taakanalist van het Raadgevend Bureau Berenschot werd bereid gevonden als proefpersoon op te treden. Hem werden 20 functiebeschrijvingen voorgelegd met het verzoek deze per gezichtspunt te graderen volgens de GM en bij het neerschrijven van de graderingscijfers gebruik te maken van een formulier zoals in Bijlage 1 is weergegeven. De proefpersoon werd gevraagd om te graderen alleen op grond van de beschrijvingen zonder gebruik te maken van voorbeeldfuncties van welke aard dan ook. Op elke functiebeschrijving was een klasse-aanduiding aangebracht. De instructie daarbij luidde als volgt: "Wel wil ik U enig idee geven van het niveau van de functie in kwestie. Tenslotte heeft U ook in de praktijk vrijwel altijd al enig idee van het niveau van een functie voor U met de analyse begint. U zult daarom bij elke functiebeschrijving een klasse-indeling vinden. Hier is een overzichtje van deze klassen met hun betekenissen".
Hij ontving dan het volgende tabelletje:

Klasse	T-score
V	75-89
IV	60-74
III	45-59
II	30-44
I	15-29

met de mededeling: "Dit niveau is uiteraard niet bindend voor U. Wij willen *Uw* oordeel over deze functies".

Deze niveau-informatie was zo gekozen dat deze bij tien functies overeenkwam met de in de praktijk, door de opstellers van de functiebeschrijvingen, gegeven T-score. Bij de overige tien functies was de correlatie tussen de informatie (verder 'T-informatie' genoemd) en het praktijk-cijfer nul.

Verondersteld werd dat de taakanalist de aanpassing van zijn T-score aan de gegeven informatie zou realiseren vooral door aanpassing van de graderingscijfers voor de zwaarst gewogen gezichtspunten (Kennis en Zelfstandigheid) aan deze T-informatie. Op grond van deze redenering werd verwacht dat de invloed van T-informatie het duidelijkst waarneembaar zou zijn op deze gezichtspunten. Gezien de in het algemeen zeer hoge correlatie tussen de cijfers voor Kennis en die voor Zelfstandigheid werd als afhankelijke variabele uitsluitend het Kennis-cijfer gebruikt.

Uit de te toetsen hypothese werd met betrekking tot het vooronderzoek de voorspelling afgeleid dat er een hoge correlatie zou bestaan tussen de T-informatie en het door de proefpersoon in het experiment gegeven graderingscijfer voor Kennis, zowel bij de tien functies waar de T-informatie juist was als bij de tien waar deze gefingeerd was. Dit hoge verband met de T-informatie zou bij de laatste tien functies moeten leiden tot een lagere correlatie van de Kennisgradering van de proefpersoon met het praktijkcijfer. Het resultaat was bemoedigend:

TABEL 4.1

Correlaties tussen T-informatie en praktijkgraderingen enerzijds en graderingen van de proefpersoon anderzijds (berekend over telkens 10 functies)

	juiste T-informatie Kennisgradering proefpersoon	gefingeerde T-informatie Kennisgradering proefpersoon
T-informatie	0,93	0,88
Kennisgradering praktijk	0,94	0,40

De correlatie met de T-informatie was hoog, ongeacht de juistheid van deze informatie. De rangcorrelatie (volgens SPEARMAN) was 0,93 bij de juiste informatie en 0,88 bij de gefingeerde. De correlatie tussen proef-

persoon-beoordeling en praktijk-beoordeling was bij de juiste T-informatie hoog (0,94), maar werd door de nadelige invloed van de gefingeerde informatie tot 0,40 gereduceerd.

De proefpersoon liet zich bij zijn graderen dus mede leiden door de niveau-informatie, ook waar deze in strijd was met de informatie die de functiebeschrijvingen gaven. Het gevolg was een interpretatie van deze beschrijvingen die in een slechte overeenkomst met de praktijkcijfers resulteerde. De uitkomsten van dit vooronderzoek waren dus van dien aard, dat voortzetting van het experiment gerechtvaardigd leek.

Er kwamen echter ook twee zwakke punten van de experimentele opzet aan het licht. De proefpersoon klaagde over de summiere functiebeschrijvingen en het is dus goed mogelijk dat hij wel op de T-informatie (juist of niet juist) af *moest* gaan bij gebrek aan betrouwbare informatie vanuit de beschrijvingen. Over de kwaliteit van de beschrijvingen was niets bekend omdat ook bij de controle-functies T-informatie was verstrekt. Een ander zwak punt had betrekking op de mogelijkheid van een statistische analyse. In een experimentele situatie waar bepaalde factoren door de experimentator als onafhankelijke variabelen ingevoerd worden om de invloed daarvan op een afhankelijke variabele (in ons geval de graderingen van de proefpersonen) na te gaan, is regressieanalyse of variantie-analyse verre te verkiezen boven een correlatieberekening (zie bijvoorbeeld HAYS 1963, p. 536-538). De aantallen functies per (gefingeerde) T-informatie-klasse waren echter voor zo'n analyse te gering en varieerden bovendien. Figuur 4.1 geeft een beeld, voor wat betreft de functies met gefingeerde T-informatie, van de verdeling der functies over de verschillende waarden van de onafhankelijke variabelen, Kennis (in de praktijk gegeven graderingen) en T-informatie.

Kennis (praktijkcijfers)	I	II	III	IV	V
4–4½	0	1	0	1	0
3–3½	0	1	0	0	1
2–2½	1	0	0	0	1
1–1½	0	1	0	2	0
0– ½	0	0	1	0	0

Gefingeerde T-informatie

Fig. 4.1 *Verdeling der functies over T-informatie en Kenniscijfers*

Voor een goede opzet is een gelijk aantal functies in elk van de 25 cellen vereist.

4.4 BESCHRIJVING VAN HET EXPERIMENT

De experimentele opzet

Om de zojuist gesignaleerde nadelen van de experimentele opzet in het vooronderzoek te vermijden werd de opzet van het experiment op twee punten grondig gewijzigd.

Ten eerste werd een controle-conditie ingevoerd. Voor elke proefpersoon of groep van proefpersonen die een aantal functies mèt (gefingeerde) T-informatie te graderen kreeg, werden vergelijkbare proefpersonen gezocht, die dezelfde functies te graderen kregen. Deze controle-proefpersonen kregen dezelfde instructies als de experimentele proefpersonen, met dien verstande dat zij géén T-informatie (noch gefingeerd, noch juist[1] ontvingen. Alle onderdelen van de instructie die op deze T-informatie betrekking hadden bleven daarbij uiteraard ook achterwege.

De functie van deze controle-conditie was tweeledig. Het is mogelijk dat de proefpersonen, evenals de proefpersoon in het vooronderzoek, zouden klagen over de kwaliteit van de functiebeschrijvingen. Zelfs wanneer dit voorkomen kan worden door betere functiebeschrijvingen voor het experiment te gebruiken, kan nog de betekenis van een eventuele invloed van de T-informatie gemakkelijk gerelativeerd worden door te wijzen op de kunstmatigheid van een situatie waar men alleen op grond van beschrijvingen moet graderen. In zo'n situatie, zo zou men kunnen aanvoeren, is er zelfs met goede functiebeschrijvingen onvoldoende informatie en moet men dus wel afgaan op informatie buiten de legitieme. Blijkt echter dat in de controle-conditie de overeenstemming met de praktijk hoog is, dan vervalt dit bezwaar.

Een tweede functie van de controle-conditie is een – in de experimentele psychologie – meer traditionele: men is pas zeker dat een bepaalde experimentele ingreep (in dit geval het verschaffen van de gefingeerde niveau-informatie) *causaal* samenhangt met de afhankelijke variabelen, als in een controle-conditie deze samenhang afwezig of

[1] Wellicht ten overvloede zij er hier op gewezen, dat in het definitieve experiment in tegenstelling tot het vooronderzoek uitsluitend met *gefingeerde* T-informatie is gewerkt. Juiste T-informatie is hier nergens gegeven.

significant minder sterk is. We zullen op deze functie van de controle-conditie nog nader ingaan, speciaal bij de bespreking van de covariantie-analyse.

De tweede wijziging in de experimentele opzet heeft betrekking op de keuze van de functies, zodanig dat zij gelijkmatig over de dimensie Kennis (praktijkcijfers) en over de T-informatie-klassen verdeeld zijn. Wij zagen hierboven (fig. 4.1), dat in het vooronderzoek de verdeling over de 25 cellen (Kennis–T-informatie–combinaties) ongelijk was.

Gelijkmatige verdeling van de functies over de 25 cellen van figuur 4.1 vraagt echter een veelvoud van 25 functies. In verband met de bedoeling om bij toepassing van een variantie-analyse ook voor interactie te toetsen (zie paragraaf 4.6) zijn minstens twee functies per cel nodig en zouden minstens 50 functies gegradeerd moeten worden. Om de proefpersonen niet te veel te belasten is gekozen voor een reductie van het aantal cellen tot 16 (vier maal vier).

De niveau-informatie werd bij deze gewijzigde opzet als volgt toegelicht:

Klasse-indeling van functies naar hun totaalscore:

Klasse	Totaalscore
A	70–84
B	55–69
C	40–54
D	25–39

De aanduidingen op de functiebeschrijvingen, die in het vooronderzoek de vorm van Romeinse cijfers hadden, kregen hier de vorm van de letters A, B, C of D. De verdeling der functies over de klassen werd als aangegeven in figuur 4.2.

Kennis (praktijkcijfers)	D	C	B	A
4–4½	2	2	2	2
3–3½	2	2	2	2
2–2½	2	2	2	2
1–1½	2	2	2	2

T-informatie

Fig. 4.2 *Verdeling der functies over T-informatie en Kenniscijfers*

In iedere cel kwamen twee functies; iedere proefpersoon moest dus 32 functies volledig graderen. De instructie was weer om de functies te graderen zonder gebruik te maken van voorbeeldfuncties. Het was niet mogelijk om te controleren of deze instructies opgevolgd werden. De proefpersonen hadden namelijk twee tot drie volle werkdagen nodig om alle functies te graderen en sommigen onder hen konden hieraan slechts incidenteel werken, zodat meer dan een maand verstreek voordat het materiaal terug kwam. Het was in deze omstandigheden uiteraard onmogelijk om bij het graderen aanwezig te zijn.

Evenals in het vooronderzoek, waren wij in de eerste plaats geïnteresseerd in de Kennisgraderingen van de proefpersonen. Deze graderingen moesten namelijk als afhankelijke variabelen fungeren. De onafhankelijke variabelen, die verondersteld werden de Kenniscijfers van de proefpersonen te beïnvloeden, waren de T-informatie en de informatie omtrent het gezichtspunt Kennis, zoals die in de functiebeschrijvingen beschikbaar was. De praktijkcijfers voor Kennis werden gebruikt als een voor de hand liggende maat voor deze Kennis-informatie.

De toekenning van de T-informatie (dat is de plaatsing van functies in één van de vier kolommen van fig. 4.2) geschiedde binnen bepaalde grenzen aselect. De grenzen werden gevormd door de door ons geschatte aanvaardbaarheid voor de proefpersonen van de T-informatie. Al te opvallende discrepanties met het werkelijke niveau van de functies werden vermeden. Blijkt bijvoorbeeld uit de naam al dat een functie van een zeer laag niveau is (straatreiniger, corveeër, handlanger) dan werd geen hoge T-informatie gegeven en was een functie te duidelijk van een hoog niveau (elektricien) dan werd een lage T-informatie vermeden.

De functiebeschrijvingen werden gekozen uit de voorbeelden van gradering van de Vereniging van Nederlandse Gemeenten (Commissie Werkclassificatie Gemeenten). Deze beschrijvingen werden door de proefpersonen bijna zonder uitzondering als goed beoordeeld en de functies zijn zeer verschillend van inhoud hetgeen de generaliseerbaarheid der uitkomsten ten goede komt. Zo kan men daarin een beschrijving vinden van de functie 'wachtsman hoofdrioolgemaal', maar ook van functies als 'zweminstructeur', 'timmerman', 'grofbankwerker' en 'zalfbereider' om maar enkele zeer uiteenlopende voorbeelden te noemen.

De proefpersonen

Het was voor de relevantie van de resultaten noodzakelijk dat ervaren taakanalisten als proefpersoon hun medewerking zouden verlenen. Zij mochten de functies echter ook weer niet te goed kennen, want dan zouden zij in veel gevallen de graderingen en in de meeste gevallen toch minstens het niveau reeds vrij nauwkeurig kennen. Aan de andere kant moesten de functies ook niet zo vreemd zijn dat de beschrijvingen ten enenmale onvoldoende zouden zijn om een redelijke indruk te geven van de functie-inhoud. Een en ander heeft ertoe geleid dat het vinden van proefpersonen een uitermate moeizame zaak is geweest.

Tenslotte hebben we twee groepen proefpersonen gevonden die enigermate aan de eisen voldeden. De eerste groep bestond uit taakanalisten van de onderafdeling Taakanalyse van de afdeling Formatiezaken van het Ministerie van Binnenlandse Zaken. De andere groep werd met opzet gezocht in een geheel andere sfeer, namelijk in werknemerskringen.

4.5 DE DRIE FASEN VAN HET EXPERIMENT

Het lag in de bedoeling om het experiment in twee fasen uit te voeren, die slechts zouden verschillen in de categorie waartoe de proefpersonen behoorden. Het experiment werd in fase I uitgevoerd met medewerking van vier taakanalisten van het Ministerie van Binnenlandse Zaken. Twee daarvan, de experimentele proefpersonen van fase I, kregen functiebeschrijvingen met T-niveau-aanduidingen (A, B, C of D). De twee controle-proefpersonen kregen geen T-informatie. Na elk deelexperiment met één proefpersoon werd met hem gesproken over zijn ervaringen, zijn indruk van de beschrijvingen, zijn werkwijze en eventuele moeilijkheden, die hij bij het graderen ondervonden had. Uit deze vier gesprekken bleek o.a. dat deze proefpersonen sterk globaal te werk gingen[2]. Zij bepaalden na lezing van de functiebeschrijving (waaraan in de experimentele conditie de gefingeerde T-informatie was toegevoegd) bij benadering een T-niveau van de functie en stelden pas daarna de graderingscijfers per gezichtspunt vast. Het totaal puntental

[2] Uit de gesprekken bleek ook dat de proefpersonen de T-informatie bewust hebben genegeerd. In paragraaf 4.8 zal blijken dat dit een invloed van deze informatie niet in de weg heeft gestaan.

werd "verdeeld over de gezichtspunten" zoals sommigen het uitdrukten. Geheel in overeenstemming hiermee had men minder vertrouwen in de afzonderlijke gezichtspuntgraderingen dan in de totaalscore.

Als echter de graderingen per gezichtspunt en vooral de graderingen die door hoge afweegfactoren niet los van een gewenst T-niveau bepaald kunnen worden (in casu het graderingscijfer voor Kennis), na de T-score vastgesteld worden, zal een eventuele invloed van de T-informatie hierop minder direct en daardoor misschien minder duidelijk zijn dan de invloed van T-informatie op de primair geschatte T-score. Het materiaal voor fase I was zo gekozen dat bij een variantie-analyse de in de publikaties van de Commissie Werkclassificatie Gemeenten gegeven kennisgraderingen (de praktijkcijfers voor Kennis) en de T-informatie onafhankelijke variabelen zouden zijn en de kenniscijfers van de proefpersoon de afhankelijke variabele. Invloed van T-informatie direct op T kon hierbij niet exact nagegaan worden.

De onderzoeker moest dus op het moment, dat de resultaten van fase I bekend waren, kiezen uit twee mogelijke voortzettingen van het experiment. Hij kon het vermoedelijke nadeel van de indirecte invloed van T-informatie op de Kenniscijfers aanvaarden en het experiment in fase II ongewijzigd herhalen met vakbondsproefpersonen of hij moest de opzet wijzigen. Het tweede alternatief zou met zich meebrengen dat een naar verhouding groot aantal functies (het bleken er achteraf 17 van de 32 te zijn) in de tweede fase nieuw zouden zijn. Dit alternatief houdt namelijk in dat de opzet, zoals schematisch weergegeven in fig. 4.2, nu in die zin gewijzigd wordt, dat de onafhankelijke variabele 'praktijkcijfer voor kennis' vervangen wordt door 'praktijk-T-score'. Om toch een gelijkmatige verdeling der functies over de cellen te behouden zouden enkele nieuwe functies ingevoerd moeten worden. De vergelijkbaarheid van de resultaten zou daaronder lijden. Zouden echter de resultaten in de tweede fase analoog zijn aan die in de eerste dan zou hiermee wel zeer aannemelijk gemaakt zijn dat de invloed van de T-informatie een vrij algemeen karakter heeft; het zou dan immers onafhankelijk van functiegroep en aard van de proefpersonen-groep opgetreden zijn.

Zou achteraf blijken dat de resultaten in de tweede fase verschillen van – bijvoorbeeld minder duidelijk zijn dan – die in de eerste fase, dan is ook dit een waardevol gegeven. Het gaat er in dit stadium van het onderzoek naar de GM niet om een meer omvattende theorie te confirmeren. Veeleer is het doel hier het leveren van, wat DE GROOT (1961) noemt, een *existentiebewijs*: wij wensen de invloed van de ge-

fingeerde informatie omtrent het niveau van de functies bij bepaalde proefpersonen en bepaalde functiegroepen aan te tonen. Er wordt daarmee nog niet gesteld dat dit bij andere proefpersonen en/of andere functies ook aantoonbaar zal zijn (men vergelijke wat DE GROOT hierover zegt, speciaal op. cit., p. 163 e.v.). Het is in dit licht bezien belangrijk om enige variatie in de experimentele opzet aan te brengen. Hierdoor kan men zich een eerste indruk vormen van de reikwijdte van de waargenomen verschijnselen en op grond van dit gegeven kan men desgewenst eventueel plannen maken voor een meer gericht experimenteel onderzoek.

In fase II van het experiment werd gewerkt met vier proefpersonen uit de vakcentrales. Twee daarvan, de experimentele proefpersonen, kregen T-informatie, de twee controle-proefpersonen kregen deze informatie niet. De opzet werd nu dus in die zin gewijzigd dat niet per Kennisklasse, maar per T-klasse acht functies gekozen werden. (De T-klassen waren 7 punten breed; 29-35, 44-50, 59-65 en 74-80.) In deze fase werd de invloed van T-informatie op de Totaalscore van de proefpersonen in plaats van op de Kennisscore nagegaan.

Achteraf bleek deze wijziging in de opzet niet nodig geweest te zijn: de proefpersonen in fase II gaven een beschrijving van hun werkwijze die vrijwel het tegenovergestelde was van die van de proefpersonen van Binnenlandse Zaken! Zij werkten juist bewust analytisch. Gezichtspunt na gezichtspunt werd afzonderlijk gegradeerd, de T-score werd daarna door weging en optelling bepaald en dit resultaat werd vrijwel altijd verder ongewijzigd geaccepteerd. Een enkele maal, zo zei men, kan het voorkomen dat de T-score sterk blijkt af te wijken van de globale indruk (die ook deze proefpersonen natuurlijk na lezing van de functiebeschrijving, maar vóór het graderen der gezichtspunten, van de functie hebben). In zo'n geval neemt men de afzonderlijke gezichtspuntgraderingen nog eens kritisch door om te zien of er iets niet goed zit. Het komt voor dat men dan toch de eerst gevonden T-score handhaaft en de discrepantie tussen globale indruk en langs analytische weg gevonden resultaat aanvaardt.

Er is hier dus, in vergelijking met de proefpersonen in fase I, een belangrijke accentverschuiving naar de gezichtspunt-graderingen en weg van de globale indruk. Voor de vakbondsproefpersonen was dus juist de opzet van fase I (invloed van T-informatie op het Kenniscijfer) optimaal geweest.

Bij de presentatie van de resultaten zal blijken dat er een groot verschil bestaat tussen fase I en II. De invloed van T-informatie was

in fase I duidelijk aanwezig, in fase II daarentegen afwezig of in ieder geval niet aantoonbaar. Was dit verschil vooral te wijten aan het verschil in functie-keuze of waren de vakbondsproefpersonen door hun meer analytische werkwijze minder beïnvloedbaar door buiten de functiebeschrijvingen liggende factoren? Om op deze vragen antwoord te kunnen geven werden in fase III van het experiment de functies van fase I aangeboden aan twee nieuwe vakbondsproefpersonen. Een van hen kreeg daarbij T-informatie, de ander diende als controle-proefpersoon.

De drie fasen van het experiment zijn in onderstaand overzicht nog eens samengevat. De keuze van de 32 functiebeschrijvingen was afhankelijk van de aard van de (onafhankelijke en afhankelijke) variabelen (Kenniscijfers in praktijk en experiment, of T-scores in praktijk en experiment). De functies in de fasen I en III zijn dus dezelfde; die in fase II zijn ten dele verschillend.

Overzicht van de drie fasen van het experiment

Fase	Proefpersonen	Variabelen Onafhankelijke	Afhankelijke
I	Binnenlandse Zaken: 2 experimentele; 2 controle.	Kennisgradering praktijk T-informatie T-score praktijk	Kennisgradering proefpersoon
II	Vakcentrales: 2 experimentele; 2 controle.	T-informatie	T-score proefpersoon
III	Vakcentrales: 1 experimentele; 1 controle.	dezelfde als in fase I	dezelfde als in fase I

4.6 OVERZICHT VAN DE GEBRUIKTE BEWERKINGSMETHODEN MET SPECIFIEKE VOORSPELLINGEN VAN DE RESULTATEN DAARVAN

De algemene hypothese, waarvan de toetsing het motief vormde voor vooronderzoek en experiment, was dezelfde bij het vooronderzoek en bij de drie fasen van het experiment. In paragraaf 4.2 is reeds een formulering hiervan gegeven. De kern van de hypothese is, dat taak-

analisten bij het graderen van functies naar een van te voren gewenst eindresultaat toewerken en dat dit resultaat gewenst wordt op grond van 'externe factoren', dat zijn factoren die geen directe relatie hebben met de functie-inhoud als zodanig. In zoverre deze externe factoren wel (indirect) positief met de functie-eisen samenhangen, is hun invloed niet meer interessant. Voor zover de invloed van deze factoren afwijkt van de functie-inhoud, kan men ze als storend en voor de GM (opgevat als meetinstrument) als nadelig beschouwen. Het is dit aspect van de externe factoren dat in het experiment door middel van de niveau-informatie is nagebootst.

De voorspellingen met betrekking tot de specifieke onderzoek-resultaten variëren wèl van vooronderzoek tot experiment en van fase tot fase. In het vooronderzoek werden op grond van deze algemene hypothese correlaties voorspeld tussen T-informatie en Kennisgraderingen van de proefpersonen en werd een verschil van de correlatie tussen in het experiment gegeven graderingen en in de praktijk gegeven graderingen voorspeld: in de controle-conditie (juiste T-informatie) hoger dan in de experimentele conditie (gefingeerde T-informatie).

De voorspellingen voor fasen I en III hadden betrekking op de relatie tussen de praktijkcijfers voor Kennis en de T-informatie enerzijds en de Kennisgraderingen van de proefpersonen anderzijds. De voorspellingen voor fase II betroffen de relatie tussen de in de praktijk gegeven T-score en de T-informatie enerzijds en de T-score van de proefpersonen anderzijds.

Men kan de gegevens op verschillende manieren bewerken en zo bovengenoemde relaties op verschillende manieren weergeven en nader onderzoeken. Met betrekking tot elk van deze bewerkingsmethoden kan men specifieke voorspellingen doen. In de volgende paragrafen zullen de resultaten achtereenvolgens op drie manieren gepresenteerd worden.[3]

1. *Weergave van de resultaten in grafieken*

Als een eerste verkenning van de relatie tussen de T-informatie en de afhankelijke variabele (Kennisgraderingen in fasen I en III, T-scores in fase II) zal in paragraaf 4.8 dit verband grafisch weergegeven worden.

De voorspelling op grond van de uitgangshypothese was, dat het gemiddelde graderingscijfer voor Kennis van de experimentele proefper-

[3] Voor de basisgegevens verwijzen wij naar de bijlagen 2 tot en met 4, blz. 115-117.

sonen van fasen I en III zou toenemen met het niveau van de ontvangen T-informatie. Verwacht werd dat dit verband bij de controle-proefpersonen, met de door hun dus niet ontvangen T-informatie, zwakker[4] of afwezig zou zijn. Voor fase II golden dezelfde voorspellingen, met dien verstande dat deze daar niet de Kennisgraderingen maar de T-scores van de proefpersonen betroffen.

In grafische termen uitgedrukt: verwacht werd dat de curven van de experimentele proefpersonen omhoog zouden lopen, terwijl die van de controle-proefpersonen horizontaal zouden zijn of minder steil[4] zouden oplopen dan die van de experimentele proefpersonen.

2. Beschrijving van de resultaten door middel van een variantie-analyse

Als de voorspellingen hierboven onder 1. genoemd geheel of gedeeltelijk bevestigd worden, doen zich verschillende vragen voor.

De eerste is die naar de statistische significantie van de gevonden samenhangen. We zullen in paragraaf 4.12 zien, dat onder bepaalde omstandigheden (geen enkel verband tussen T-informatie en de afhankelijke variabele in de controle-conditie) een variantie-analyse een voorzichtige conclusie met betrekking tot deze significantie mogelijk maakt.

Een tweede vraag is die naar de sterkte van de samenhang tussen T-informatie en afhankelijke variabele in vergelijking tot die tussen de praktijk-Kenniscijfers (of T-score) – als index voor de impliciete informatie omtrent de zwaarte van de functie-eisen, zoals gegeven in de functiebeschrijvingen – en de graderingen van de proefpersonen.

Wat ons hier interesseert, is: als externe factoren het graderingsproces beïnvloeden, hoe sterk is deze invloed dan vergeleken met de invloed van de legitieme (interne) factoren, zoals aanwezig in de functiebeschrijvingen? Een tweezijdige variantie-analyse kan een antwoord op deze vraag geven.

[4] Het lijkt vreemd om in de controle-conditie zelfs maar een zwak verband te verwachten tussen de *daar niet gegeven* T-informatie en de graderingen van de proefpersonen. De toekenning van de T-informatie-niveaus (A, B, C en D) aan de functies is echter niet geheel aselect gedaan. De experimentator heeft zich mede door plausibiliteitsoverwegingen laten leiden (zie paragr. 4.4 onder 'De experimentele opzet'). Het is in principe niet uitgesloten dat hij daarbij heeft geanticipeerd op wat ook voor de proefpersonen 'plausibel' was en dat daardoor indirect een verband zou ontstaan tussen de niet gegeven T-informatie en de graderingen. Het zal blijken dat dit in fase II inderdaad het geval is geweest.

Een derde vraag is die naar de mogelijke afhankelijkheid van de invloed van externe factoren van de mate van discrepantie tussen deze externe informatie en de in de functiebeschrijvingen gegeven informatie. Wellicht zou de invloed geringer zijn wanneer de T-informatie duidelijk in strijd is met de functiebeschrijvingen, dan wanneer het verschil tussen T-informatie en praktijkcijfer slechts klein is. Wanneer de invloed niet constant is, zou dit bij een tweezijdige variantie-analyse tot een zogenaamd interactie-effect moeten leiden.

De voorspellingen met betrekking tot deze drie vragen luidden:

i. Het verband tussen de T-informatie en de afhankelijke variabele zal bij toepassing van een variantie-analyse significant blijken te zijn.
ii. Wat betreft de sterkte-verhouding tussen de externe en de interne invloeden verwachten wij, dat de externe invloed relatief sterk zal zijn. Om deze voorspelling iets minder vaag te maken: in termen van 'verklaarde variantie' verwachten wij dat de door de T-informatie verklaarde variantie bijvoorbeeld meer dan een derde bedraagt van die door de andere onafhankelijke variabele (praktijk-Kenniscijfer of praktijk-T-score) verklaard wordt. Het lijkt ons niet waarschijnlijk dat de invloed van T-informatie die van de andere onafhankelijke variabele zal overtreffen.
iii. Wat betreft het interactie-verschijnsel hebben wij geen onvoorwaardelijke voorspelling durven doen. Het lijkt ons echter het meest waarschijnlijk dat, als interactie optreedt in overeenstemming met onze hypothese, te grote discrepantie tussen T-informatie en functiebeschrijvingen zal leiden tot negeren van de T-informatie.

3. Toetsing van de significantie van de invloed van de T-informatie op het graderingsproces door middel van een covariantie-analyse

Onder 2. hierboven werd erop gewezen dat met de onderhavige experimentele gegevens significantie-toetsing door middel van variantie-analyse slechts onder bepaalde omstandigheden, en dan nog maar te nauwer nood, mogelijk is. De reden voor deze voorzichtigheid is, dat een variantie-analyse slechts conclusies toelaat omtrent significante (eventueel zelfs causale) samenhangen, als de indeling van de objecten in de waarnemingscellen (in ons geval de cellen van fig. 4.1 en 4.2) volkomen aselect heeft plaats gevonden.

We hebben er in paragraaf 4.4 echter reeds op gewezen dat de toekenning van de T-informatie slechts binnen bepaalde grenzen aselect

geschiedde. Al te weinig plausibele T-informatie-toekenningen zijn door ons zoveel mogelijk vermeden. Een covariantie-analyse maakt het in ons geval mogelijk om voor zo'n 'selectieve vertekening' te corrigeren, dank zij het feit dat wij ook over gegevens van controle-proefpersonen beschikken. In paragraaf 4.10 zal dit nader toegelicht worden.

De voorspelling met betrekking tot de uitkomsten van een covariantie-analyse luidde: Het verband tussen de T-informatie en de afhankelijke variabele zal, ook na correctie voor selectieve vertekening, significant zijn. De voorspellingen zojuist genoemd onder 2. en 3. zullen in de paragrafen 4.12 en 4.13 getoetst worden.

4.7 WAS DE EXPERIMENTELE SITUATIE TE KUNSTMATIG OM TE KUNNEN GRADEREN?

Bij de beschrijving van de experimentele opzet (paragraaf 4.4) werd erop gewezen, dat eventuele positieve resultaten van het experiment gemakkelijk gerelativeerd konden worden door te wijzen op de kunstmatigheid van de experimentele situatie. In deze paragraaf zullen gegevens gepresenteerd worden, waaruit blijkt dat – kunstmatig als de situatie ook was – nauwkeurig graderen zeer goed mogelijk was.

Men kan de nauwkeurigheid van het graderen op twee manieren nagaan: men kan de graderingen van de controle-proefpersonen vergelijken met de praktijkcijfers en men kan ze onderling vergelijken. Tabel 4.2 geeft de correlaties van de Kennisgraderingen en de T-scores van de proefpersonen met die in de praktijk gegeven. Men ziet dat de Kennisgraderingen en de T-scores van de eerste controle-proefpersoon van fase I (C-I-1) en van de (enige) controle-proefpersoon van fase III (C-III) zeer goed overeenkomen met de praktijkcijfers (correlaties van 0,79 en hoger).

TABEL 4.2

Correlaties van Kennisgraderingen en T-score van de proefpersonen met praktijkcijfers in de controle-conditie (berekend over 32 functies)

	fase I	fase II		fase III
proefpersonen:	C-I-1	C-II-1	C-II-2	C-III
Kennis	0,84	0,66	0,62	0,79
T-score	0,90	0,70	0,70	0,79

De correlaties van de tweede controle-proefpersoon van fase I (C-I-2) zijn niet gegeven. Het is achteraf gebleken, dat deze proefpersoon in feite, door een niet geheel waterdichte experimentele situatie, in een verzwakt experimentele conditie heeft gegradeerd. Hierop wordt, bij de bespreking van de grafische weergave van de resultaten (paragraaf 4.8), nog nader ingegaan.

De overeenkomst met de praktijk was in de fase II minder goed. Gelukkig waren in die fase de gegevens van twee controle-proefpersonen beschikbaar, zodat daar de onderlinge overeenkomst nog nagegaan kon worden. Het bleek dat de correlaties tussen de graderingen van deze proefpersonen zeer hoog waren: Kennis, $r = 0{,}90$; T-score, $r = 0{,}94$. De relatief slechte overeenkomst met de praktijk moet hier veeleer toegeschreven worden aan verschil in opvatting tussen deze categorie proefpersonen en de opstellers van de praktijkcijfers met betrekking tot deze functies. Een zo hoge correlatie tussen twee, onafhankelijk van elkaar graderende, proefpersonen zou immers niet mogelijk geweest zijn als graderen in de kunstmatige experimentele situatie niet goed mogelijk geweest was. De conclusie is dus: de kunstmatigheid van de experimentele situatie is geen voldoende reden om de betekenis van eventueel positieve resultaten te relativeren.

4.8 WEERGAVE VAN DE RESULTATEN IN GRAFIEKEN

In de figuren 4.3 en 4.4 is grafisch het verband weergegeven tussen de T-informatie en de gemiddelde waarden van de afhankelijke variabele (Kennisgradering van de proefpersonen in fasen I en III (fig. 4.3); T-score van de proefpersonen in fase II (fig. 4.4)).

In figuur 4.3 val tonmiddellijk op, dat de experimentele curven steil oplopen en dat de controle-curven vrijwel horizontaal verlopen. Dit is dus geheel in overeenstemming met de verwachtingen, afgeleid uit de hypothese van de invloed van de T-informatie. De curve van de tweede controle-proefpersoon van fase I (C-I-2) blijkt een tussenpositie in te nemen: minder vlak dan de twee andere controle-curven (C-I-1 en C-III), maar niet zo steil als de experimentele curven (E-I-1, E-I-2 en E-III). De verklaring daarvan is eenvoudig en interessant. Proefpersoon C-I-2 heeft, evenals C-I-1, een controle-instructie ontvangen. De functiebeschrijvingen waren echter ook bij hem voorzien van de T-informatie-letters. Hiervan werd bij de instructie slechts even terloops gezegd dat het een codering voor de proefleider betrof die met de

Relatie tussen de variabele 'T-informatie' en de in het experiment gegeven graderingen.

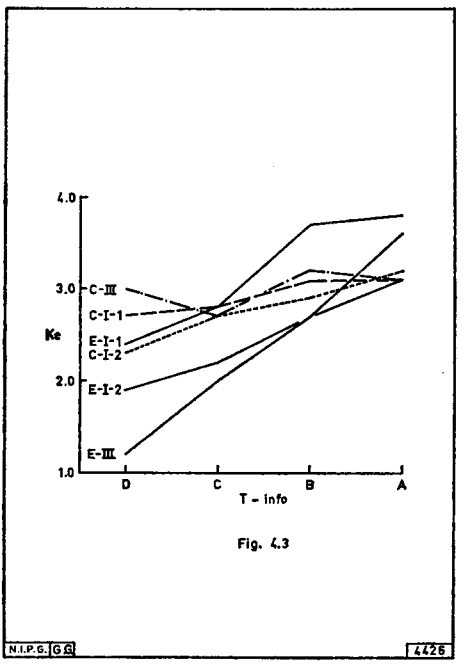

Fig. 4.3

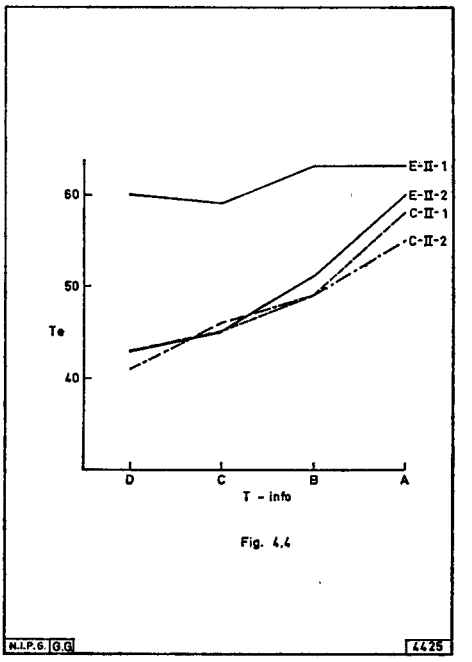

Fig. 4.4

De afkorting 'T-info' staat voor de gefingeerde T-informatie. De afkorting 'Ke' in fig. 4.3 staat voor 'in het experiment door de proefpersonen gegeven Kennisgraderingen'. De afkorting 'Te' staat voor de in het experiment toegekende T-score. De getrokken curven hebben betrekking op de gemiddelden van de experimentele proefpersonen, de onderbroken curven zijn die van de controle-proefpersonen.

functie als zodanig niets te maken had en dus verder genegeerd kon worden. Terwijl dit door C-I-1 zonder meer aanvaard en onthouden werd, heeft C-I-2 halverwege de week, waarin hij aan het graderen was, de proefleider opgebeld om te vragen naar de betekenis van de letters. Wat was het geval? Hij had gemerkt dat er nog al wat lage functies met een D waren en toen hij daarop enkele A-functies opzocht bleken daar nogal wat hoge bij te zijn. Zijn veronderstelling was dat er misschien een verband was tussen de letters en het niveau.

Om te voorkomen dat ook andere controle-proefpersonen dergelijke onbedoelde conclusies zouden trekken kregen de controle-proefpersonen in de volgende fasen van het experiment functiebeschrijvingen *zonder* letters.

Achteraf werd door C-I-2, evenals trouwens door de experimentele

proefpersonen E-I-1 en E-I-2 van fase I, ten stelligste beweerd dat hij de letters genegeerd had. C-I-2 beriep zich hierbij op de, telefonisch nogmaals herhaalde, instructie; de (experimentele) proefpersoon E-I-1 vroeg zich af of de T-informatie een grapje was: hij kon er niet in geloven. E-I-2 gaf wel toe dat hij soms door grote verschillen tussen zijn eigen oordeel en de T-informatie onzeker werd, maar, zo zei hij, hij was toch steeds helemaal zijn eigen weg gegaan. E-III (fase III) had ook verschillen geconstateerd, maar deze rustig genegeerd, zo zei hij. Het is in het licht van deze stellige uitspraken des te merkwaardiger dat zo'n duidelijk verband bestaat tussen de T-informatie en de graderingen.

Het beeld in fig. 4.4 (fase II) is geheel anders. Daar doet zich niet het duidelijke verschil voor tussen controle-curven en experimentele curven. Drie curven zijn alle even steil en vallen zelfs vrijwel samen. Dit zijn echter twee controle-curven (C-II-1 en C-II-2) en een curve van een experimentele proefpersoon (E-II-2). De enige proefpersoon, die zich gedraagt zoals we van een controle-proefpersoon verwacht hadden (een vlakke curve) is de experimentele proefpersoon E-II-1! Dit laatste is achteraf wel begrijpelijk: het bleek namelijk dat deze proefpersoon deel had uitgemaakt van een commissie, die destijds bij de totstandkoming van de hier gebruikte (praktijk-)graderingen meegewerkt had. Ondanks het feit dat sindsdien geruime tijd verlopen was, was hij toch nog in staat om een overeenkomst met de praktijkcijfers te krijgen van $r = 0{,}93$. Het valt niet te verwachten dat in zo'n geval de T-informatie nog veel invloed kan uitoefenen.

Blijft nog het verschijnsel van de steile controle-curven. Hier hebben we kennelijk te maken met het resultaat van wat wij hier boven 'selectieve vertekening' hebben genoemd. Onze plausibiliteitsoverwegingen bij de toekenning van de T-informatie (vergelijk paragraaf 4.4 en 4.6) waren kennelijk in fase II ten dele in overeenstemming met de opvattingen van de controle-proefpersonen, die van deze T-informatie niets af konden weten (zij hadden géén letters op hun functiebeschrijvingen).

Hoe valt dit verschil tussen fasen I en III enerzijds en fase II anderzijds nu te verklaren? Het is zeer onwaarschijnlijk dat het aan de soort proefpersonen ligt, aangezien zowel in fase III als in fase II taakanalisten uit de vakcentrales aan het experiment hebben deelgenomen. Het is veel waarschijnlijker dat het vrij grote aantal functies dat alleen in fase II gebruikt was, tot dit zo verschillende resultaat heeft geleid.

De voorspellingen met betrekking tot de samenhang tussen de gemiddelden van de afhankelijke variabele en de T-informatie, zoals geformuleerd in paragraaf 4.6 onder 1. zijn dus voor fasen I en III beves-

tigd. De resultaten van fase II waren niet in strijd met de uitgangshypothesen (de experimentele curven waren niet beide vlak), maar vormden ook geen bevestiging hiervan, doordat er geen verschil was tussen controle- en experimentele conditie.

4.9 HET STATISTISCHE SIGNIFICANTIE-BEGRIP

De volgende fase in de bewerking en presentatie van de onderzoekresultaten is het toetsen van de significantie van de in de vorige paragraaf geconstateerde samenhangen. Dit zal in eerste instantie geschieden door middel van een variantie-analyse, die echter vooral gebruikt zal worden als een descriptieve techniek om de sterkte van de samenhang van T-informatie met de afhankelijke variabele te kunnen vergelijken met die tussen de praktijkcijfers en de afhankelijke variabele. De definitieve statistische toetsing van de hypothese, dat de T-informatie het graderingsproces beïnvloedt, zal door middel van de covariantie-analyse plaats vinden.

Als inleiding tot deze presentatie van de resultaten van variantie- en covariantie-analyse zal hier eerst een korte beschouwing gegeven worden over het significantie-begrip, gevolgd door een uiteenzetting betreffende de toetsing door middel van een covariantie-analyse.

De curven weergegeven in figuur 4.3 en 4.4 hebben betrekking op gemiddelden. In de experimentele conditie van fasen I en III neemt de *gemiddelde* kennisgradering toe. Men kan uit deze figuren niet aflezen hoe sterk de graderingscijfers per functie binnen een en dezelfde T-informatie-categorie rondom deze gemiddelde waarden gespreid zijn, noch ook of deze spreiding *vergeleken met de verschillen tussen de gemiddelden* groot of klein genoemd mag worden. Toch zijn deze gegevens voor de statistische significantie van de gevonden verbanden van het grootste belang. In het volgende zal op deze statistische aspecten van de resultaten nader ingegaan worden.

De vraag naar de significantie van gevonden resultaten is een vraag naar de generaliseerbaarheid of algemeengeldigheid van deze resultaten. Men beschouwt daarbij de uitkomsten van het experiment als een steekproef uit een verzameling van uitkomsten (in de statistiek meestal 'populatie' genoemd), die bij herhaling van het experiment gevonden zouden kunnen worden. Belangrijk is hierbij wat men verstaat onder een 'herhaling van het experiment'. Dit is de vraag van de populatiedefinitie. Het experiment kan natuurlijk strictu sensu nooit herhaald

worden. Een herhaling is nooit in alle details gelijk aan het oorspronkelijke experiment. Er zijn dus bepaalde variaties die toegestaan zijn om nog van 'hetzelfde experiment' te kunnen spreken. Ze worden als irrelevant beschouwd. Zo kan men een experiment herhalen op een andere dag met dezelfde proefpersonen of men kan het experiment gelijktijdig met andere proefpersonen uitvoeren. Die andere proefpersonen zullen dan aan bepaalde voorwaarden moeten voldoen om het experiment nog als een herhaling te kunnen beschouwen. Hoe ruimer deze voorwaarden des te groter de kans op 'toevallige' variatie, maar ook des te groter de generaliseerbaarheid als de uitkomsten significant blijken.

Meer concreet: als wij voor ons experiment een zeer homogene steekproef van functies hadden gekozen en een invloed van de T-informatie hadden geconstateerd dan zouden we nog niet hebben geweten of dit resultaat ook voor andere, niet tot die groep behorende functies zou gelden. Hadden we daarentegen functies gekozen die bijvoorbeeld qua niveau meer verschilden dan de nu gebruikte (bijvoorbeeld ook functies met T-scores boven 100) dan konden we ook iets zeggen over de invloed van de T-informatie bij hoger geclassificeerde functies. Hetzelfde geldt voor onze keuze van proefpersonen. Taakanalisten uit werkgeverskringen en van raadgevende efficiency-bureaux zijn niet bij het onderzoek betrokken en de resultaten mogen dan ook niet zonder meer tot die groepen gegeneraliseerd worden.

In ons experiment is er sprake van twee soorten steekproeven: we hebben steekproeven van proefpersonen en steekproeven van functies. De eerstgenoemde steekproeven zijn bijzonder klein, per populatie maximaal slechts vier personen. Het was praktisch niet mogelijk om dit aantal belangrijk uit te breiden. Met statistische technieken is hier dan ook weinig uit te richten. Toch valt er wel iets te zeggen over de algemeengeldigheid van de uitkomsten. Het zal blijken dat het niet erg waarschijnlijk is dat de hier gevonden resultaten uitsluitend het gevolg zijn van een toevallige keuze van proefpersonen en dat deze bij andere proefpersonen dus niet zouden optreden. We komen daarop in paragraaf 4.14 terug.

De functiesteekproeven waren veel groter: $N = 32$. Hier kan het toepassen van statistische technieken wel enige zin hebben. De opzet van het experiment is zodanig geweest dat een variantie-analyse mogelijk is, en wel een tweezijdige, waarbij tegelijkertijd de invloed van de functiebeschrijvingen en die van de T-informatie op de graderingen nagegaan kon worden. Bovendien werden per cel (men zie figuur 4.2)

twee functies genomen zodat de mogelijkheid bestond om voor interactie te toetsen (men zie voor bijzonderheden over deze techniek bijvoorbeeld HAYS 1963). Dit laatste werd belangrijk geacht omdat het niet onmogelijk is dat de invloed van T-informatie anders is wanneer deze meer van de functiebeschrijving afwijkt. Zo'n variatie in de invloed van de T-informatie zou als interactie in een variantie-analyse tot uiting moeten komen.

Als eenmaal een significante relatie geconstateerd is tussen T-informatie en de graderingen, moet nagegaan worden of deze relatie beschouwd kan worden als een causale. Vergelijking met de controle-conditie waar geen T-informatie gegeven is, is daarvoor noodzakelijk. Deze vergelijking zou kunnen geschieden door eenzelfde variantie-analyse uit te voeren in de controle-conditie. Vindt men daar géén significante relatie dan wordt het waarschijnlijk dat, in het verband tussen T-informatie en de graderingen, T-informatie inderdaad de oorzaak is geweest. De betekenis van het verschil tussen controle- en experimentele conditie is echter niet vast te stellen. Als het verband tussen T-informatie en graderingen in de experimentele conditie net significant is (bijvoorbeeld op het 5%-niveau) en in de controle-conditie net niet significant (P bijvoorbeeld tussen 5% en 10%) is dat verschil dan zelf weer significant?

4.10 COVARIANTIE-ANALYSE

Er bestaat een directere wijze om van de gegevens in de controle-conditie gebruik te maken, namelijk door middel van een covariantie-analyse, een variant van de variantie-analyse. Bij een variantie-analyse wordt nagegaan of groepen waarnemingen, die op bepaalde aspecten (de onafhankelijke of experimentele variabelen) verschillen, *daardoor* ook op de afhankelijke variabele verschillen. In ons geval: groepen functiebeschrijvingen verschillen in het niveau van de hen toegekende T-informatie (de experimentele variabele) en wij willen nagaan of ze daardoor ook in Kennisgraderingen of T-score (van de proefpersonen) verschillen. Vinden wij dat de gemiddelde door proefpersonen gegeven graderingen van de verschillende T-informatie-groepen (A, B, C en D) significant verschillen dan lijkt het voor de hand te liggen om die verschillen toe te schrijven aan de invloed van de T-informatie. Deze conclusie mogen we echter alleen trekken als voldaan is aan een belangrijke voorwaarde: de functiebeschrijvingen moeten volkomen

aselect, willekeurig, in T-informatie-groepen ingedeeld zijn. Als er enige selectie heeft plaats gevonden kan het zijn dat de T-informatie-groepen behalve op T-informatie, ook nog op andere aspecten niet geheel vergelijkbaar zijn en dan is achteraf niet meer na te gaan of de verschillen in de graderingen moeten worden toegeschreven aan de T-informatie of aan die andere aspecten.

Een voorbeeld dat veel gebruikt wordt om het principe van een covariantie-analyse mee te illustreren is dat van het didactische experiment. Stel er zijn drie onderwijsmethoden, A, B, en C. Om na te gaan of deze methoden verschil maken in wat de leerlingen in een bepaalde tijd kunnen leren worden drie groepen leerlingen gevormd. Aan elke groep wordt onderwijs gegeven in hetzelfde onderwerp maar met verschillende methoden. Als de leerlingen volkomen lukraak in de drie groepen ingedeeld zijn kunnen systematische verschillen in bijvoorbeeld intelligentie niet optreden. (Met toevallige verschillen wordt in een variantie-analyse rekening gehouden.) Verschillen tussen de groepen in wat geleerd wordt moeten dus wel gevolg zijn van de didactische methoden.

Als echter de leerlingen reeds in klassen gegroepeerd waren en men wil de klassen intact houden dan is het zeer wel mogelijk dat sommige klassen betere leerlingen hebben dan andere. De verschillen die tussen de resultaten van de drie groepen gevonden worden, behoeven dan niet (uitsluitend) gevolg te zijn van de gevolgde didactiek. De covariantie-analyse is nu een methode om de niet volkomen aselect gekozen groepen achteraf vergelijkbaar te maken door de resultaten te corrigeren op grond van wat men aan (niet experimentele) verschillen heeft geconstateerd. Men schat welke verschillen in resultaten men kan verwachten op grond van bijvoorbeeld de scores op een intelligentie-test. Als de groepen nu nog méér verschillen dan verwacht werd, moeten deze verschillen gevolg zijn van de experimentele (in dit geval didactische) ingreep. Een covariantie-analyse is dus eigenlijk een variantie-analyse op gegevens die gecorrigeerd zijn voor systematische, maar in het experiment niet gewenste, verschillen tussen de experimentele groepen.

Ook in ons experiment was het mogelijk dat de functiebeschrijvingen, die in de verschillende T-informatie-groepen ingedeeld werden, niet geheel vergelijkbaar waren doordat wij bij onze indeling plausibiliteitsoverwegingen hebben laten meespelen. We hebben gezorgd dat de T-informatie-groepen wat betreft de praktijkgraderingen (Kennis en T-score) niet verschilden, maar dat was niet genoeg. Ten gevolge van onze selectie zouden ze nog kunnen verschillen in de waardering die zij

van de proefpersonen krijgen. Dit is nagegaan door deze functies zonder T-informatie aan proefpersonen voor te leggen. De functie van de graderingen van de controle-proefpersoon werd daardoor dezelfde als de functie van de intelligentietest in het didactische experiment: correctiemogelijkheid leveren voor verschillen die door de experimentator niet bedoeld waren maar die door het ontbreken van volledige aselectiviteit toch opgetreden zijn.

4.11 VOORONDERSTELLINGEN BIJ TOEPASSING VAN VARIANTIE-ANALYSE EN COVARIANTIE-ANALYSE

Mag men nu een statistische bewerkingsmethode als een variantie-analyse of een covariantie-analyse op dit materiaal toepassen? Is wel voldaan aan de voorwaarden die men in zo'n geval aan het materiaal moet stellen? We hebben in hoofdstuk II gezien dat één voorwaarde waaraan voldaan moet zijn, willen statistische uitspraken empirisch zinvol zijn, het bestaan van een bepaald schaaltype is. Is aan de statistische voorwaarden, gedicteerd door de wiskundige waarschijnlijkheidstheorie, voldaan dan mag men de statistische methoden wel toepassen (vergelijk LORD 1953) maar de interpretatie van de resultaten ("the return trip from the numbers to properties of objects", HAYS 1963, p. 76) is dan nog niet altijd mogelijk. Wanneer wij echter willen aantonen dat factoren buiten de functiebeschrijvingen (T-informatie) graderingen kunnen beïnvloeden interesseert het ons *in dit kader* niet of de graderingen zinvol zijn en zo ja, wat hun betekenis is. Wij stellen ons, voor wat betreft dit experimentele onderzoek, op het standpunt van de gebruiker van de methode, die de graderingscijfers hanteert alsof ze intervalschaalkarakter hebben. Vanuit dit standpunt trachten wij dan aan te tonen dat een T-informatie-effect bestaat waardoor de methode, *zelfs als zij schaaltechnisch gezond zou zijn*, onmogelijk erg valide kan zijn.

De voorwaarden die bij toepassing van een variantie-analyse op grond van het statistische model aan het cijfermateriaal moeten worden gesteld hebben betrekking op de verdeling binnen de experimentele subpopulaties (normaal en met gelijke variantie) en op de onafhankelijkheid van fouten (zie bijvoorbeeld HAYS 1963, p. 364). Bij een covariantie-analyse komen daar nog voorwaarden bij die betrekking hebben op de regressie van de afhankelijke variabele op de controlevariabele (zie bijvoorbeeld DE JONGE 1964, p. 590-591).

Het is bij het aantal waarnemingen waarover wij per experimentele subpopulatie beschikken niet mogelijk om voor normaliteit te toetsen. Als de afwijking van de normaliteit in deze subpopulaties dezelfde is, is het negatieve effect op de waarde van de F-toets echter betrekkelijk gering (HAYS 1963, p. 379). Er is geen enkele reden om bij ons materiaal *verschillende* verdelingen aan te nemen, wel om een niet normale verdeling te verwachten. De gelijkheid van foutenvariantie is evenmin erg belangrijk als het aantal waarnemingen per steekproef maar gelijk is (HAYS, loc. cit.). In ons geval hebben we bij de enkelvoudige analyse steeds acht functies per T-informatie-categorie en bij de tweevoudige analyse steeds twee functies per cel. Over eventuele variantie-verschillen behoeven we ons dus ook geen zorgen te maken.

De voorwaarde waaraan in ieder geval voldaan moet worden is die van de onderlinge onafhankelijkheid van toevalsfouten (meetfouten), dat wil zeggen dat de toevalsfouten over alle waarnemingsparen ongecorreleerd moeten zijn (zie HAYS 1963, p. 364). Het is echter altijd bijzonder moeilijk om te controleren of aan deze voorwaarde voldaan is. HAYS waarschuwt speciaal voor die gevallen waar meerdere observaties (metingen) gedaan worden aan dezelfde proefpersonen. In leerexperimenten is dit bijvoorbeeld van belang. De eerste prestatie kan dan vaak op systematische wijze de volgende prestaties beïnvloeden. In ons geval is hoogstens te verwachten dat een proefpersoon die een functie toevallig wat hoger waardeert dan hij gewoon is (een positieve toevalsfout) ook de andere functies hoger zal graderen. In die zin heeft de eerste fout invloed op de latere graderingen. Omdat deze invloed zich echter bij alle graderingen van deze proefpersoon zal voordoen en er per analyse telkens slechts sprake is van één en dezelfde proefpersoon, komt dit neer op een geringe verschuiving van de door hem gebruikte graderingsschaal. Op de resultaten van de analyse heeft dit geen enkele invloed. Om deze redenen zullen we er hier vanuit gaan dat de resultaten van de F-toetsen met enige reserve aanvaard mogen worden en dat we dus zeer significante uitkomsten als significant mogen beschouwen.

4.12 STATISTISCHE BEWERKING VAN DE RESULTATEN VAN FASEN I EN III

Wij hebben gemeend enige resultaten van de tweevoudige variantie-analyse hier te moeten weergeven, ondanks het nadeel dat de significantie van de resultaten in de experimentele conditie moeilijk te vergelijken is met die in de controle-conditie (vergelijk paragraaf 4.9). Een vergelijking van de invloed van T-informatie met die van de praktijkcijfers voor Kennis (als maat voor de informatie omtrent het gezichtspunt Kennis dat in de functiebeschrijvingen vervat is) is echter bij de covariantie-analyse niet mogelijk. Daar wordt namelijk gecorrigeerd

voor verschillen in kennisgraderingen door de controle-proefpersoon en deze graderingen correleren zeer hoog met de overeenkomstige praktijkcijfers. De invloed van het praktijkcijfer voor Kennis wordt bij een tweevoudige covariantie-analyse dan ook vrijwel geheel geëlimineerd.

Evenals bij een correlatiecoëfficiënt kan men ook bij een variantie-analyse onderscheid maken tussen de *mate van samenhang* en *de graad van significantie* daarvan. In het algemeen zal een sterkere samenhang ook eerder een bepaald significantie-niveau bereiken, maar bij zeer grote steekproeven kan zelfs de zwakste samenhang significant worden en bij zeer kleine steekproeven kunnen zeer duidelijke samenhangen niet significant blijken te zijn. Terwijl men bij correlatiecoëfficiënten vooral let op de sterkte van de samenhang en dan meestal ook nog wel controleert of deze significant is, wordt de variantie-analyse vrijwel uitsluitend gebruikt als een toetsingsmethode. Vrijwel nooit vindt men een aanduiding van de mate van samenhang; men schijnt alleen geïnteresseerd in de mate waarin het resultaat significant is, in de P-waarde (de overschrijdingskans) dus. Terecht zegt HAYS (1963, p. 326) hierover dat een significante samenhang zo zwak kan zijn dat ze triviaal wordt. Het is daarom nodig om ook altijd de sterkte van de samenhang in de beschouwing te betrekken.

Hij geeft als een bruikbare maat een grootheid die hij aanduidt met ω^2 en die evenals het kwadraat van de correlatiecoëfficiënt beschouwd kan worden als een maat voor het gedeelte van de variantie van een (meestal afhankelijke) variabele dat 'verklaard' wordt door de (meestal onafhankelijke) variabele. HAYS spreekt zowel van "proportion of variance accounted for" als van "relative reduction in uncertainty". Het laatste trekt het begrip meer in de sfeer van het kunnen voorspellen en vermijdt daardoor eventuele causale connotaties. Om die reden is het te verkiezen boven de term 'verklaarde variantie'. Wij zullen verder steeds het symbool w^2 gebruiken om de schatting van ω^2 volgens één van HAYS' schattingsformules aan te duiden (HAYS 1963, p. 407).

Met behulp van de index w^2 kan men de variantie-analyse gebruiken als een descriptieve methode. "It is useful to think of the analysis of variance as a device for 'sorting' the information in an experiment into nonoverlapping and meaningful portions" (HAYS 1963, p. 409). In tabel 4.3 wordt zo'n descriptie gegeven voor de zes proefpersonen van fasen I en III (die dezelfde functiebeschrijvingen hebben gegradeerd).

We zien hieruit dat in de experimentele conditie het verband tussen T-informatie en Kennisgradering (proefpersonen) in twee gevallen ongeveer gelijk is aan de helft van het verband van het praktijk-Ken-

TABEL 4.3
De samenhang tussen praktijkcijfers voor Kennis en T-informatie enerzijds en de Kennisgraderingen van de proefpersonen anderzijds, uitgedrukt in de index w^2

	experimentele conditie			'half-experimentele' conditie	controle-conditie	
proefpersonen:	E-I-1	E-I-2	E-III	C-I-2	C-I-1	C-III
Kennis (praktijk)	48%	43%	21%	69%	68%	65%
T-informatie	23%	29%	38%	7%	1%	0,2%
interactie	0%	10%	12%	2%	3%	9%

niscijfer met deze Kennisgradering en in één geval (E-III) zelfs groter dan dat tussen praktijk- en proefpersonengradering op het gezichtspunt Kennis.

Iets minder voorzichtig geformuleerd: bij twee proefpersonen is de invloed van de fictieve T-informatie op de kennisgraderingen ongeveer de helft van die, uitgaande van de (interne) informatie vervat in de functiebeschrijvingen. Bij één proefpersoon is die invloed zelfs groter dan de invloed van de interne informatie. Deze proefpersoon laat zich dus bij het graderen meer leiden door 'externe' informatie dan door de legitieme.

Zoals hierboven reeds vermeld zijn significantie-schattingen in dit geval, gezien het ontbreken van aselectiviteit, niet erg betrouwbaar. De selectie is echter in de fasen I en III niet zo ernstig: de resultaten in de controle-conditie geven namelijk nauwelijks enig verband tussen T-informatie en de Kennisgraderingen (bij C-I-1 hangt Kennis slechts voor 1% samen met T-informatie, bij C-III is dit zelfs maar 0,2%). Daarom wordt hier, onder voorbehoud, toch enige significantie-informatie verschaft. (zie tabel 4.4)

De voor onze vraagstelling belangrijkste gegevens zijn de P-waarden voor de T-informatie-effecten in de experimentele conditie. Deze zijn zo klein dat aan de significantie hiervan nauwelijks meer getwijfeld behoeft te worden ondanks mogelijke storende invloeden als het ontbreken van een volledige aselectiviteit. Deze conclusies worden bij de covariantie-analyse dan ook bevestigd.

De interactie is niet of nauwelijks significant. Alleen bij proefpersoon E-I-2 is P kleiner dan 5%. Op grond van deze gegevens lijkt het niet erg waarschijnlijk dat de invloed van T-informatie afhankelijk is van specifieke combinaties van T-informatie en praktijk-Kenniscijfer.

TABEL 4.4

Significantieschattingen van de relaties beschreven in tabel 4.3 in P-waarden (overschrijdingskansen)

	experimentele conditie			'half-experimentele' conditie	controle-conditie	
proefpersonen:	E-I-1	E-I-2	E-III	C-I-2	C-I-1	C-III
Kennis (praktijk)	<0,001	<0,001	<0,005	<0,001	<0,001	<0,001
T-informatie	<0,001	<0,001	<0,001	<0,05	NS	NS
interactie	NS	<0,05	±0,05	NS	NS	NS

Na deze eerste exploratie en het voorlopig verwerpen van de interactie-hypothese, zijn we gereed om de resultaten van de covariantieanalyse te bespreken. Zij zijn weergegeven in tabel 4.5 in de vorm van P-waarden. Deze overschrijdingskansen hebben betrekking op de invloed van T-informatie op de Kennisgraderingen in de experimentele conditie na correctie voor selectieve vertekening. Anders gezegd: ze hebben betrekking op de mate waarin het verband tussen T-informatie en de Kennisgraderingen van de proefpersonen verschilt van (uitgaat boven) dat wat men op grond van de gegevens in de controle-conditie zou verwachten.

TABEL 4.5

Significantieschattingen van de invloed van T-informatie op de Kennisgraderingen, uitgedrukt in P-waarden

proefpersonen:	E-I-1*	E-I-2*	C-I-2*	E-III**
P-waarden:	<0,05	<0,01	NS	<0,001

* gecorrigeerd voor Kennis-scores van proefpersoon C-I-1
** gecorrigeerd voor Kennis-scores van proefpersoon C-III

Zoals hierboven reeds gesteld werd zou het beter geweest zijn als eenzelfde proefpersoon zowel in een controle-conditie als in een experimentele conditie dezelfde functies gegradeerd had. Dit is uiteraard niet mogelijk. Vergelijkbare functies zijn, ten eerste, moeilijk te vinden. Ten tweede zou zo'n procedure te veel van de tijd van de proefpersonen gevergd hebben. De controle-proefpersonen in deze opzet zijn derhalve zodanig gekozen dat men redelijkerwijs mag aannemen dat ze verge-

lijkbaar zijn met de experimentele proefpersonen. Vandaar dat in fase III voor E-III de proefpersoon C-III als controle diende en in fase I voor de proefpersonen E-I-1, E-I-2, en C-I-2 de proefpersoon C-I-1.

Het blijkt dat ook na correctie de T-informatie op significante wijze met de Kennisgraderingen samenhangt; alleen de eerst nauwelijks significante relatie bij C-I-2 blijkt nu niet meer significant te zijn. Hieruit blijkt de zin van de toepassing van de covariantie-analyse. Interessant is dat niets erop wijst dat de vakbondsproefpersoon, die een sterk analytische graderingsmethode volgt, minder gevoelig zou zijn voor T-informatie dan de proefpersonen van Binnenlandse Zaken. Het is eerder zo dat bij E-III het verband met T-informatie nog sterker (tabel 4.3) en significanter (tabel 4.5) is dan bij de proefpersonen van fase I.

Het is niet onmogelijk dat dit een gevolg is van de grotere vertrouwdheid van de taakanalisten van de rijksoverheid met de functies en functiebeschrijvingen van de gemeente. De hogere correlatie tussen C-I-1 en de praktijkgraderingen (0,84 voor kennis en 0,90 voor T) vergeleken met die tussen de vakbondsproefpersonen en de praktijkcijfers (variërend van 0,62 tot 0,79; zie tabel 4.2) wijzen in dezelfde richting. Hoe beter men (in de controle-conditie) de functiebeschrijvingen kan interpreteren des te minder gevoelig zal men (in de experimentele conditie) zijn voor T-informatie. Deze 'verklaring' – het zij hier met nadruk gesteld – is niet meer dan speculatief. Een andere mogelijke verklaring voor het meer uitgesproken T-informatie-effect in fase III is, dat hier de experimentele opzet (invloed van T-informatie op Kennisgraderingen) het best aangepast was aan de werkwijze van de proefpersonen, die immers vooral analytisch was.

Over het algemeen genomen zijn de voorspellingen geformuleerd in paragraaf 4.6 onder 2., i. en ii. voor wat betreft de fasen I en III hiermee bevestigd. Alleen de verwachting, dat het verband tussen de T-informatie en de graderingen van de proefpersonen niet sterker zou zijn dan dat tussen T-informatie en de relevante praktijkgradering, bleek nog te voorzichtig: in fase III was dit (bij één proefpersoon) immers wèl het geval.

Ook de voorspelling van paragraaf 4.6 onder 3., met betrekking tot de resultaten van een covariantie-analyse, is voor de fasen I en III uitgekomen.

4.13 STATISTISCHE BEWERKING VAN DE RESULTATEN VAN FASE II

In tabel 4.6 zijn – evenals in tabel 4.3 voor de fasen I en III is gebeurd – de descriptieve resultaten van een tweevoudige variantie-analyse weergegeven.

TABEL 4.6
De samenhang tussen praktijk-Totaalscores en T-informatie enerzijds en de T-scores van de proefpersonen anderzijds, uitgedrukt in de index w^2

	experimentele conditie		controle-conditie	
Proefpersonen:	E-II-1	E-II-2	C-II-1	C-II-2
T-score (praktijk)	81%	45%	45%	47%
T-informatie	0%	19%	14%	14%
interactie	0%	0%	0%	4%

Het percentage door T-informatie 'verklaarde' variantie is, zoals te verwachten was, bij E-II-1 (het commissielid) nul; bij de andere experimentele proefpersoon was dit percentage iets hoger dan bij de twee controle-proefpersonen. De toch nog relatief sterke samenhang tussen T-informatie en T-score in de controle-conditie (geschat significantie-niveau kleiner dan 2½%!) bevestigt onze eerste interpretatie van de gegevens: er heeft een selectieve vertekening plaats gevonden, waardoor significantie-berekeningen op grond van een variantie-analyse zinloos zijn. Of het geringe verschil tussen experimentele en controle-conditie (19% tegenover 14%) significant is, kan weer onderzocht worden door middel van een covariantie-analyse. Als correctie-variabele is hier gebruikt het gemiddelde van de T-scores van de twee controle-proefpersonen. De invloed van T-informatie op de T-score van proefpersoon E-II-2 bleek na correctie niet meer significant te zijn.

De voorspellingen van paragraaf 4.6 onder 2. hadden betrekking op de uitkomsten van een variantie-analyse. Voorspeld werd een positief en significant verband tussen T-informatie en de afhankelijke variabele. Zo'n resultaat is alleen maar overtuigend als het verband in de controle-conditie zwakker of liever nog: geheel afwezig is. In de fasen I en III bleek dit het geval te zijn. In fase II was het verband in de controle-conditie echter nauwelijks zwakker. Veel meer relevant voor deze situatie is de voorspelling met betrekking tot de uitkomsten van de covariantie-analyse, gedaan in paragraaf 4.6 onder 3. Deze voorspelling is, voor wat betreft fase II, niet uitgekomen.

4.14 SAMENVATTING EN CONCLUSIE

Experimenteel onderzoek met betrekking tot systemen van werkclassificatie is zeldzaam. Een van de weinige publikaties op dit gebied is van de hand van MYERS. Hij had belangstelling voor een verschijnsel dat hij 'forcing of job ratings' noemde. Dit toewerken naar een gewenst eindresultaat is voor een (meet-) methode met enige wetenschappelijke pretentie wel een zeer zwak punt. Ook wij kregen de indruk, bij een aantal oriënterende gesprekken met werkclassificatiedeskundigen, dat de taakanalist in de praktijk naar een uitkomst toewerkt die past in (zijn beeld van) de bestaande loonstructuur.

Het in dit hoofdstuk beschreven experiment was in het bijzonder gericht op het toetsen van de hypothese, dat externe factoren, dat zijn factoren die geen directe relatie hebben tot de functie-inhoud als zodanig, de uitkomsten van een werkclassificatie kunnen beïnvloeden. Voorts hebben wij getracht enig inzicht te krijgen in de wijze waarop de proefpersonen de invloed van de informatie, die wij hen als een nabootsing van zo'n externe factor hebben aangeboden, hebben ervaren.

Meer in het algemeen was dit onderzoek bedoeld om aan te tonen dat ook op het gebied van de werkclassificatie, dat – zoals wij in de hoofdstukken II en III hebben gezien – moeilijk in kaart te brengen is, experimenteel onderzoek mogelijk is. Wij hebben antwoorden gezocht op methodologische vragen als:

- kan men langs experimentele weg vat krijgen op het verschijnsel dat MYERS aanduidt met "forcing of job ratings" en dat hier omschreven is als "toewerken naar een gewenst eindresultaat"?
- wat is de reikwijdte van dit fenomeen?
- welke moeilijkheden ontmoet men bij het opzetten en uitvoeren van zo'n experiment?

De boven omschreven hypothese is door het experiment bevestigd: externe factoren, zoals door ons geoperationaliseerd in een gefingeerde T-niveau-aanduiding, beïnvloeden onder bepaalde omstandigheden de graderingen van ervaren taakanalisten op statistisch significante wijze, als men deze vergelijkt met graderingen van vergelijkbare taakanalisten die zonder deze factoren dezelfde functies graderen.

Strikt genomen zijn deze conclusies slechts ondubbelzinnig aangetoond voor enkele proefpersonen. Maar dit waren *ervaren taakanalisten* en aangetoond is in ieder geval dat het effect bij zulke proefpersonen kan voorkomen èn zeer groot kan zijn. De significantieschattingen had-

den steeds betrekking op steekproeven van functies en niet op steekproeven van proefpersonen en generalisatie naar andere proefpersonen is dus op grond van de statistische gegevens niet verantwoord. De grote overeenkomst in de structuur van de resultaten bij zulke uiteenlopende proefpersonen als die in fasen I en III maakt het echter op zijn minst onwaarschijnlijk dat het fenomeen van de invloed van T-informatie specifiek is voor deze, willekeurige, proefpersonen.

De operationalisering van het begrip 'externe factoren' is in fase II niet erg gelukkig geweest. Wij hebben de T-niveau-aanduidingen in het algemeen zo gekozen, dat ze ons in meerdere of mindere mate 'plausibel', althans aanvaardbaar toeleek op grond van een oppervlakkige lezing van de functiebeschrijving. In fase II van het experiment bleek dit zover te gaan dat de informatie althans voor een deel overeenkwam met de opvattingen van de controle-proefpersonen. Wij spraken daar van selectieve vertekening om aan te duiden dat wij bij de keuze van functies voor de verschillende T-informatie-categorieën niet aselect te werk zijn gegaan. Wij hebben hier kennelijk voor een deel rekening gehouden met informatie die in de functiebeschrijvingen aanwezig was.

Men kan nog een tweede, meer principieel bezwaar tegen de hier gekozen operationalisatie inbrengen. Als externe factoren worden hier beschouwd al die gegevens waarover een taakanalist beschikt (zoals positie in de loonstructuur en in de hiërarchie, status van de functie e.d.) die niet noodzakelijkerwijs in overeenstemming behoeven te zijn met de functie-eisen als gedefinieerd door de GM. Nu kan men stellen dat de proefpersonen in dit experiment de T-informatie hebben opgevat als wel relevant in het kader van de GM, dus niet als externe factor. De experimentator zei immers in zijn instructie dat hij een indruk wilde geven van het niveau van de functie en hij noemde in dat verband punten werkclassificatie (A betekent een T-score van 70 tot 85 punten). De proefpersoon was dus gerechtigd, zo zou men kunnen aanvoeren, om de T-informatie als informatie te beschouwen die wél betrekking had op de functie-inhoud als zodanig. Ze was immers – zover de proefpersoon kon nagaan – zeer waarschijnlijk afkomstig van een vorige analyse van deze inhoud volgens de GM.

Deze redenering lijkt sterk, maar houdt bij nader inzien geen stand. De kern van het betoog rust op de interpretatie door de proefpersoon van de T-informatie. Nu weten wij uit nabesprekingen met de proefpersonen dat de T-informatie eenvoudig niet au sérieux genomen werd (zie paragraaf 4.8). De toevoeging aan de instructie, "Dit niveau is

HOOFDSTUK V

Slotbeschouwingen

5.1 SAMENVATTING

"De techniek der werkclassificatie heeft de mogelijkheid geschapen tot een fijnere en meer verantwoorde indeling der functies te komen. Met behulp van deze techniek kan op veel nauwkeuriger en objectiever wijze dan vroeger mogelijk was, een rangorde der functies met betrekking tot het loon worden vastgesteld." Dit citaat uit 'Werkclassificatie als hulpmiddel bij de loonpolitiek', een brochure van de Stichting van den Arbeid, geeft duidelijk de verwachtingen weer die met betrekking tot de GM hebben geleefd en die nog steeds in brede kring de houding tegenover deze methode karakteriseren (zie bijvoorbeeld Bijlage II in LIJFTOGT 1966). *Nauwkeurigheid* en *objectiviteit* zijn hierin de kernbegrippen, die reeds doen vermoeden dat men maar al te zeer geneigd is om de GM te beschouwen als een methode die langs wetenschappelijke weg ontwikkeld is. Dit vermoeden wordt bij verder lezen van deze brochure bevestigd als gesproken wordt van "een stuk wetenschappelijke arbeid, afgestemd op de practijk" (op. cit., p. 16). Het is de bedoeling van dit onderzoek geweest om na te gaan of deze pretenties van nauwkeurigheid, objectiviteit en wetenschappelijke basis gerechtvaardigd zijn.

Dit onderzoek valt in twee delen uiteen: een (meet-)theoretische analyse en een experimenteel onderzoek. Door middel van de theoretische analyse zijn wij tot de conclusie gekomen dat de methode, zoals deze nu gebruikt wordt, onbevredigend is. Zij is alleen te verdedigen als men zich op het standpunt stelt dat zij slechts een stabiliserende functie behoeft te hebben (men vergelijke wat hierover in de inleiding gezegd werd). Stelt men zich echter op het standpunt dat men met behulp van deze methode in staat zou moeten zijn om aan de hand van feitelijke gegevens een waardering van functies in loontechnisch opzicht op te stellen, dan is het noodzakelijk dat deze twee aspecten – het feitelijke en het waarderende – methodisch duidelijk onderscheiden

worden. Dit standpunt werd hier uitdrukkelijk verdedigd. De GM is een onderhandelingsinstrument, waarmee een waardering van feitelijke gegevens tot uitdrukking kan worden gebracht. Zolang deze waardering niet duidelijk van de gegevens onderscheiden wordt, is de betekenis van de resultaten van een werkclassificatie-onderzoek ondoorzichtig, hetgeen een zinvolle discussie vrijwel onmogelijk maakt.

In de methode zoals deze ontwikkeld en in praktijk gebracht is, zijn nu inderdaad deze twee aspecten – het waarderende en het feitelijk waargenomene (gemetene) betreffende – onontwarbaar vermengd. Wil men de methode zien als voorspeller van – dat wil in de praktijk zeggen: als efficient substituut voor – een criterium, dan hebben de afweegfactoren slechts de (statistische) functie om een zo goed mogelijke overeenkomst met het criterium te bewerkstelligen. Zij kunnen dan niet tevens explicitering van waarden zijn, zodat het waarde-aspect niet afzonderlijk ter discussie staat. Alleen wanneer men uitdrukkelijk een meetinstrument of een batterij van meetinstrumenten tracht te ontwikkelen, rekening houdende met de schaaltechnische voorwaarden waaraan dan voldaan moet zijn, is het mogelijk om daarnaast de afweegfactoren als explicitering van waarden te hanteren.

Nadat in onze beschouwingen de metende functie losgemaakt was van de waarderende kon de combinatie van de meetuitkomsten als zuiver meettheoretisch probleem aangevat worden. Het bleek dat de gezichtspuntgraderingen minstens een intervalkarakter moesten hebben wilden zij in het gunstigste geval – de graderingscijfers zijn gelijksoortige grootheden, die slechts in schaaleenheid verschillen – na vermenigvuldiging met afweegfactoren optelbaar zijn. Een voorlopig plan werd opgesteld tot ontwikkeling van graderingscijfers met intervalschaalkarakter.

Pogingen om ook het probleem van de afweegfactoren en het daarmee verwante probleem van de keuze der gezichtspunten langs wetenschappelijke weg op te lossen (door middel van factor-analyse en regressietechnieken) bleken op principiële gronden tot mislukking gedoemd. De opvatting werd verdedigd dat keuze van gezichtspunten en de weging daarvan geen kwestie van theorievorming is in de zin van een zo adekwaat mogelijk weergeven van de werkelijkheid, maar een explicitering van waarden inhoudt. Deze explicitering kan principieel niet op wetenschappelijke gronden geschieden, maar moet gezien worden als een – eventueel in gezamenlijk overleg te nemen – beleidsbeslissing.

De mogelijkheid tot het leveren van een bijdrage van de gedragswetenschappen, in casu de psychologie, tot de ontwikkeling van een

meer verantwoorde werkclassificatie-methode ligt vooral op meettechnisch gebied: het opstellen van graderingsschalen per gezichtspunt. Zowel de Stichting van den Arbeid als ook de DC hebben herhaaldelijk gewezen op de noodzaak van wetenschappelijk onderzoek met betrekking tot de afweegfactoren. Onze conclusie is dat zo'n onderzoek het probleem kan verhelderen maar dat in laatste instantie de oplossing ervan niet langs wetenschappelijke weg bereikt kan worden.

De bewering, hierboven geciteerd, dat de GM een "objectiever wijze" van rangordebepaling van functies mogelijk maakt, veronderstelt dat elke taakanalist tot ongeveer dezelfde resultaten zal komen bij de toepassing van de GM. Bovendien zal men mogen eisen van een objectieve en wetenschappelijk verantwoorde methode, dat zij niet overmatig gevoelig is voor systematische maar externe factoren, dat wil zeggen voor factoren die met de aard van de functie hoogstens een indirect verband hebben. Om na te gaan of de GM aan deze eisen voldoet is in een experimentele situatie onderzocht of ervaren taakanalisten zich bij hun graderingen soms mede laten beïnvloeden door informatie die als externe informatie wordt gepresenteerd. Het bleek dat de invloed van deze informatie, die door de proefpersonen unaniem als extern werd beschouwd, bij één groep functies en bij twee soorten taakanalisten aantoonbaar was – en relatief groot. Bovendien viel het op dat de proefpersonen zich van deze invloed in het geheel niet bewust waren.

In het licht van deze resultaten behoeft het ons niet te verwonderen dat de GM in zijn stabiliserende functie 'goed voldoet'. In de inleiding (hoofdstuk I) werden drie voorwaarden genoemd, waaraan een werkclassificatie-methode moet voldoen om in grote lijnen met de praktijk overeen te komen. De eerste twee hadden betrekking op de methode zoals die in voorschriften is vastgelegd. De derde voorwaarde had betrekking op een mogelijk gebrek aan exactheid, waardoor de taakanalist naar een gewenst resultaat, bijvoorbeeld overeenstemming met de praktijk, kan toewerken. Het in hoofdstuk IV beschreven experiment maakt het wel zeer waarschijnlijk dat aan deze laatste 'voorwaarde' voldaan is. Deze voorwaarde is in die zin van de eerste twee verschillend, dat deze binnen bepaalde grenzen een voldoende voorwaarde is: zelfs al is aan de eerste twee niet voldaan, dan nog kan een kunstmatige overeenkomst met de bestaande loonstructuur bereikt worden.

Overzien we het resultaat van deze studie als geheel dan is de conclusie met betrekking tot de waarde van de GM als wetenschappelijke methode van functie-beoordeling ongunstig. Meettheoretisch is ze

ondoorzichtig en ongezond, praktisch is ze te sterk gevoelig voor externe invloeden. Deze veroordeling, gefundeerd op een theoretische analyse en een experimenteel en daardoor nooit geheel bewijskrachtig onderzoek, zal velen te ver gaan. Vooral zij die dagelijks met de methode te maken hebben zullen, zelfs wanneer zij geen houdbare tegenargumenten kunnen formuleren, het gevoel hebben dat dit beeld de GM onrecht doet. Dit is begrijpelijk en ten dele terecht: de GM heeft zeker een positieve functie bij de loonbepaling, zo niet als meetinstrument dan toch als analyse-methode. De gezichtspunten vormen een begrippensysteem waardoor het mogelijk wordt om meer gedifferentieerd dan voorheen over functie-verschillen te discussiëren. De pretentie van de methode is echter altijd meer geweest en het is deze pretentie die in de bovensamengevatte veroordeling bestreden wordt.

Wil men de methode handhaven en haar zowel wetenschappelijk verantwoord als praktisch bruikbaar maken, dan zijn er toch nog wel ontwikkelingsmogelijkheden. Deze kunnen echter alleen gerealiseerd worden langs de weg van veel en grondig voorbereid wetenschappelijk onderzoek. Een begin is reeds gesuggereerd: de ontwikkeling van intervalschalen per gezichtspunt. In de volgende paragraaf zullen nog enkele voorstellen gedaan worden die in dit kader niet verder in detail uitgewerkt kunnen worden.

5.2 PERSPECTIEF

Bij de gesprekken met de proefpersonen bleek dat sommige taakanalisten vooral globaal, andere vooral analytisch graderen. Wij hebben geen duidelijk verschil geconstateerd in gevoeligheid voor de invloed van externe informatie (T-informatie) tussen deze twee werkwijzen. Een vergelijking van deze methoden met betrekking tot betrouwbaarheid was met deze onderzoek-opzet niet mogelijk. Zo'n onderzoek zou uitgevoerd kunnen worden met een – zo homogeen mogelijke – groep taakanalisten als proefpersonen. De helft van deze groep krijgt dan de instructie om, na lezing van de functiebeschrijvingen, schattingen te maken van de T-scores. Nadat ze deze schattingen gegeven hebben graderen ze per gezichtspunt. De andere helft krijgt de uitdrukkelijke opdracht om gezichtspunt na gezichtspunt afzonderlijk te graderen zonder rekening te houden met de gevolgen van deze graderingen voor de hoogte van het totaalpuntental. Aan deze instructie kan bijvoor-

beeld toegevoegd worden: "U kunt later altijd correcties aanbrengen als blijkt dat de T-score niet aanvaardbaar is". Door deze toevoeging zal men eerder geneigd zijn om los van de konsekwenties voor de T-score te graderen. Op grond van deze gezichtspuntgraderingen worden dan de T-scores berekend. De onderlinge overeenkomst binnen de globaal werkende en binnen de analytisch werkende groep kan dan berekend worden met betrouwbaarheidsformules bij voorkeur van het variantie-analytische type (zie bijvoorbeeld RAJARATNAM 1960 en CRONBACH, RAJARATNAM en GLESER 1963).

Als men dit experiment zou uitvoeren zowel met een aantal taakanalisten die gewoonlijk analytisch te werk gaan (zoals onze vakbondsproefpersonen) als met een groep die gewoonlijk globaal werkt (zoals onze proefpersonen van Binnenlandse Zaken) zou men enige controle hebben op de effectiviteit van de experimentele instructies. Nog beter zou zijn, als men enkele proefpersonen uit beide groepen de instructie gaf om te graderen zoals ze dat gewoon zijn. Deze zouden fungeren als controle-groepen.

Als de in hoofdstuk III gedane voorstellen tot wijziging van de methode aanvaard zouden worden zou het beter zijn dit experiment uit te voeren met de nieuwe methode als werkclassificatie-instrument. Bovendien wordt het dan van belang om het beoordelings- of graderingsproces bij de taakanalist te bestuderen. Een bij uitstek geschikte methode van onderzoek hiernaar zou de 'hardop-denk-methode' zijn, zoals bijvoorbeeld beschreven door DE GROOT en zijn medewerkers (DE GROOT 1946 en 1966). Bij dit soort experiment wordt de proefpersoon gevraagd om bij alles wat hij doet zoveel mogelijk hardop te denken, zodat de psycholoog het proces min of meer kan volgen. Op deze wijze kan men enig inzicht krijgen in de mate waarin de taakanalist bij toepassing van de gewijzigde GM erin slaagt om gezichtspunten afzonderlijk te graderen. Is hij bijvoorbeeld geneigd om discrepanties in graderingsniveau tussen gezichtspunten, die hij als verwant ziet, weg te werken of te voorkomen? Gebruikt hij de 'inconveniënten' om een functie op een wat hoger niveau te brengen als het totaalcijfer wat erg ongunstig uitvalt? Werkt hij met bepaalde globale beoordelingen omtrent het niveau van de te graderen functies? Dit zijn vragen die slechts langs de weg van een hardop-denk-experiment op enigszins betrouwbare wijze te onderzoeken zijn.

Als men een werkclassificatie-methode als de GM van voldoende belang acht voor een verantwoorde loonvorming om de voorgestelde (of een andere) wijziging aan te brengen en in de praktijk te toetsen,

dan zal zeker ook bij deze nieuwe versie onderzocht moeten worden of systematisch werkende maar externe factoren het graderingsproces (kunnen) beïnvloeden. Men kan te dien einde een experiment, als in hoofdstuk IV beschreven, uitvoeren of ook hardop-denk-experimenten ontwikkelen die speciaal op dit verschijnsel gericht zijn.

In paragraaf 4.14 werd erop gewezen dat het toewerken naar een eindresultaat onder de invloed van externe factoren (motivatie, informatie) ook bij andere analytische beoordelingsmethoden aangetoond is. Men kan hieraan de conclusie verbinden dat een analytische methode geen garantie biedt tegen dit soort ongewenste invloeden. Het wordt dan ook niet onwaarschijnlijk geacht dat, ook na de boven voorgestelde wijziging van de methode, een invloed van T-informatie op de graderingscijfers aantoonbaar zal zijn. In zo'n geval is het noodzakelijk dat maatregelen genomen worden om te verhinderen dat taakanalisten – waarschijnlijk onbewust – meer doen dan alleen graderen. Een nogal rigoureuze uitweg uit deze impasse is op meerdere gronden te verdedigen. Gedacht wordt aan het afschaffen van een universeel en constant systeem van afweegfactoren. Hierdoor wordt voor de taakanalist veel minder voorspelbaar wat de konsekwenties van zijn graderingen voor de loonhoogte zijn. De afweegfactoren kunnen in onderhandelingen per bedrijfstak (of zelfs per bedrijf) vastgesteld worden. Men kan zich ook voorstellen dat de graderingscijfers de feitelijke basis vormen voor de onderhandelingen, waarbij per bedrijfstak vastgestelde afweegfactoren slechts als richtlijnen dienen en dus op meer flexibele wijze gebruikt worden. Het zal de onderhandelingen niet gemakkelijker maken, maar dat is de prijs voor een minder rigide en daardoor in potentie beter aan situatie en rechtsgevoel aangepast systeem. Een laatste argument voor deze functie van de afweegfactoren is gelegen in de conclusie, in hoofdstuk III getrokken, dat in tegenstelling tot het graderingsproces, het bepalen van de afweegfactoren zich voor wetenschappelijk onderzoek nauwelijks leent. Een systeem van afweegfactoren dat op wetenschappelijke gronden niet te verdedigen valt moet uit dien hoofde met voorzichtigheid en soepel gebruikt worden om willekeur te vermijden.

Tot slot zij hier met nadruk gesteld dat het belangrijkste bij de invoering van een nieuwe methode is dat dan een wetenschappelijk onderzoek van de werkclassificatie niet weer zo lang uitgesteld wordt dat het voor velen bijna onaanvaardbaar wordt, omdat dit onderzoek weer discutabel dreigt te stellen, waarin men zoveel geïnvesteerd had. Men trekke lering uit de historische beschouwingen van LIJFTOGT

(1966) en houde voortaan de communicatielijnen tussen de praktijk en de (gedrags-)wetenschappen open. Dit zal de ontwikkeling van een goed instrumentarium mogelijk maken en een zo doorzichtig mogelijk overleg met betrekking tot de toepassing hiervan bij de loonvorming bevorderen.

Summary

The Normalized Method of Job Evaluation, the subject of this study, has been designed as a standard job evaluation procedure, its main purpose being to make a comparison of jobs in different industrial enterprises possible, so as to enable governmental agencies to decide on demands for wage increases. These decisions had to be made because of a government-directed wage policy, which was established around the year 1950. (See LIJFTOGT 1966)

Primarily job evaluation has always been a practical affair. The adjective 'scientific' has been frequently applied to methods of job evaluation, but rarely with good reason. This study is part of a larger research project, originally conceived by Professor A. D. DE GROOT of the Municipal University of Amsterdam, in order to find out, to what extent this specific instance of a job evaluation method, the Normalized Method, could be considered to be scientifically sound.

The first part of the project has been a historical enquiry into the development of the Normalized Method, conducted by LIJFTOGT, a sociologist (LIJFTOGT 1966). Starting-point of the present study was the conception that a scientific job evaluation method should be regarded as a measuring procedure. In the chapters I through III some ensuing theoretical questions are discussed:

- to what extent can decisions on practical issues be based on data, engendered by a measuring instrument?
- what type of instrument is the Normalized Method: is it a simple predictor, i.e. a mere substitute for a more fundamental criterion variable, or should it rather be considered as a criterion measure in its own right? (For this distinction, see DE GROOT 1961)
- if one has chosen, as the present author did, for the second alternative, then what are the measurement theoretical implications of this point of view?

Chapter IV is a report of an experiment, designed to explore how far the Normalized Method, as it is used to-day, leaves room for external

information to influence the measurement outcomes. The expression "external information" refers to information the task analyst is not supposed to use, even though it may be available to him. Such information may be about the position of the job to be analyzed in the hierarchical and wage structure of the plant, the status of the job in general, etc. According to the principles and pretensions of the Method, its outcomes ought to be independent of such external information, primarily because such information is not necessarily in agreement with an independent evaluation.

In the experimental situation, one particular type of "external information" was simulated, namely fore-knowledge of the general level of the total score to be expected.

Summary of chapter I: Introduction

An analogy is drawn between the situation, where psychological advice is given and the use of a job evaluation method. It is pointed out that in both cases one should warn the user that a "scientific" method is not unfailing or errorless. In both cases two questions of relevance were shown to be pertinent:

1. Is the population on which the measuring procedure is tried out and validated relevant for the questions asked?
2. Is the system of values implied by the procedure relevant?

The problem of the relevance of a chosen population is especially important for a *normalized* method. Apart from urgent political considerations like a government-directed wage policy, it is considered indefensible to use one and the same method for many different job categories.

The problem of the relevance of value systems was shown to be hard to explore in the case of the Normalized Method, because the factual, measuring aspects and the value aspects of the method are inextricably confounded.

Summary of chapter II: Measurement theoretical criticism

The chapter begins with the observation, that the question whether the operations performed on job evaluation scales (multiplication of the so-called 'factors' by weights and adding the products) are meaningful,

very rarely has been raised. As the ordinal character of the scales has been frequently emphasized, this omission is surprising, to say the least of it.

These considerations are pertinent only, if a job evaluation system is regarded as a criterion measure. If, however, it is regarded as an efficient substitute for some other 'ultimate' variable, which is harder to measure, questions of scale type are less relevant and may be ignored, if the substitute has been shown to be a valid predictor of the criterion.

Several reasons are brought forward for considering a job evaluation method as a criterion variable, rather than as a predictor, the most important argument being, that such a method actually is a bargaining instrument. It should, therefore, clarify issues. This is only possible, if two functions of the method, i.e. its measuring function and its evaluation function, are kept separate. The measuring function is performed by the different factor scales (e.g. experience, physical demands, etc.), the evaluation function is represented by a set of factor weights. If the method is used as a predictor, the function of the weights may be compared to that of beta-weights in a regression equation. Regression weights, however, cannot be indicators of value at the same time. As an extreme example: negative regression weights have sometimes been obtained in a multiple regression analysis of a job evaluation system. Negative weights are, of course, unacceptable as value indicators in a bargaining situation.

An outline of STEVENS' treatment of measurement scales is presented and two types of criticism are discussed, one type based on statistical reasoning (see e.g. LORD 1953), the other based on an operational approach (see e.g. COMREY 1951 or DE GROOT 1962). The first form of criticism is shown to miss the point, the second is not relevant, if one chooses to consider a job evaluation method to be a criterion measure.

Accepting in essence STEVENS' stand-point, an analysis was made of the operations performed during the application of the Normalized Method of Job Evaluation. The analysis followed the presentation used by SUPPES and ZINNES to point out some of the properties and implications of their definition of 'empirical meaningfulness'. (SUPPES and ZINNES 1963)

The conclusions are that the weighted sums, assigned to jobs, can only be used to create a rank order, if the different factor scales have the properties of an absolute scale (no transformation allowed, as, e.g. in counting). In case the factor scales have no more than interval properties, addition is only meaningful, if the different factors are 'similar

variables'. Job demands are called 'similar', if they are judged according to a dollar criterion, that is, if the demands are expressed in something comparable with cardinal utilities (see, e.g. CRONBACH and GLESER 1957). Differences in 'monetary units' could then be corrected by appropriately chosen factor weights and the resulting weighted sum would be empirically meaningful.

It was suggested that the Normalized Method would be only acceptable as a criterion measure, if the three following conditions were met:

a. the different job factor scales should be 'similar', i.e. differ only in the unit of measurement.
b. there should be a unit of measurement, i.e. the job factors should at least have the properties of an interval scale.
c. it should be possible to estimate the ratios between the units of the job factor scales.

It is judged to be very unlikely that these conditions are met at present. The so-called Committee of Experts which developed the Normalized Method (LIJFTOGT 1966) has always insisted on the ordinal character of the factor scales and the weights have served more as regression weights, developed by trial and error, than as value indicators (measures of 'monetary units' or cardinal utility).

Summary of chapter III: Attempt at a constructive contribution

Chapter II was concluded on a rather sombre note: "Addition of the weighted job factor scores is a meaningless procedure, if the job factors are measured on an ordinal scale". Two solutions to this unpleasant situation were pointed out: one can either establish interval scales for the job factors or look for methods to combine variables with ordinal properties only. In the literature on decision theory different ways of combining such variables have been described (see e.g. the treatment of so-called "welfare functions" in LUCE and RAIFFA 1957). In this chapter some of these combination methods are discussed. It is shown that, for different reasons, none of these combination rules are applicable to job evaluation scales.

However, transformation of a job factor scale into ranks results in an absolute scale (cf. KENDALL 1962, p. 1). This seemed a way out, since a (weighted) sum of ranks is a meaningful number. It represents the total number of jobs which are lower in rank – on any job factor –

than the job under consideration. To solve the weighting problems it was suggested to count the jobs, ranked lower on important factors, more than once.

The question is, however, what kind of relation exists between this unambiguous piece of information and the demands, a particular job makes on the man who performs it. It seemed that our reasoning had been fallacious in the same way as the argument of LORD's statistician (LORD 1953). For our purpose, that numerical assignment system is relevant, where certain degrees of components of job demands (called factors) are the objects and ranks are the numerical values assigned to these objects. Within *this* system, the empirical relational system, which has the set of all possible levels of job demands as its domain, is isomorphic to a numerical system with ranks as its elements, *if these ranks are considered to be ordinal numbers* (cf. SUPPES and ZINNES 1963 and ADAMS et al. 1965). Again the conclusion had to be drawn that addition, even of ranks, is not allowed.

The other possible solution mentioned above, namely to establish interval scales for each job factor, is discussed. A simple method, mentioned by TORGERSON (1958, p. 137) is suggested as a way to develop such scales.

As soon as the decision is made to develop interval scales for each job factor, the question arises, whether the choice of components of job demands and of weights for these factors cannot be made in a more systematic manner than has been done by the Committee of Experts (see LIJFTOGT 1966).

Many psychologists would tend to answer questions like these in the affirmative. In most cases they would point out the possibility of determining type and number of job factors by way of factor analysis. Using factor analysis this way implies that highly correlated components fundamentally are the same. However, this assumption is not necessarily true. Fundamentally similar variables must be highly correlated but the reverse doesn 't hold good. Other weak points in factor analysis were mentioned: its lack of a unique solution and the dependence of the resulting factors on the population chosen. In the opinion of the author these features of the method make factor analysis unfit as a theoretical basis for the choice of job factors. At the most it might be useful to improve the efficiency of an existing job evaluation procedure by offering an approximation (see e.g. LAWSHE 1945).

It became clear that an adequate choice of job components is no simple matter. A correlational analysis can provide danger-signals:

two highly intercorrelated job factors *might possibly* be duplicates. In the case of the Normalized Method, for instance, 'knowledge' and 'responsibility' (the literal translation would be 'independence') are highly correlated and a comparison of the factor descriptions strongly suggests a considerable overlap. It was stated, however, that conclusions about overlap or identity of components should never be based solely on a correlational analysis.

Another problem, which cannot be solved by statistical means only, is the question of the weights to be given to job factors, once they are chosen. If a job evaluation method could be considered a predictor, determination of regression weights would be a straightforward solution of the problem. However, job evaluation methods being regarded as criterion measures in their own right – or rather, as sets of criterion measures: the job evaluation scales – weights must be considered value indicators. A set of such value indicators bears a certain resemblance to the set of axioms LUCE and RAIFFA (1957, p. 121 ff.) are referring to when discussing arbitration schemes. It is emphasized that one should examine the consequences of the choice of a particular set of reasonable weights carefully, just as one should check whether the arbitration scheme chosen yields results which we deem 'fair' (op. cit., p. 122).

It is pointed out that there is a certain similarity between testing a theory and testing the quantification of a set of values. Special attention was given to the difference between these two testing procedures. In the first case, tests are done by checking whether predictions deduced from the theoretical hypotheses have materialized. In the second case, tests are performed by checking whether the consequences of a particular value quantification are still acceptable. The outcome of the first type of test is more or less unambiguous: one can argue about it in rational terms. The second type of test does not allow for a conclusion on the truth or validity of the values as quantified. Rather, it serves to clarify value issues and the discussion about these issues. This is regarded as a convenient characterization of the purpose of job evaluation as a whole.

Summary of chapter IV: The experiment

In the theoretical part of this study, the Normalized Method was severely criticised. The present chapter is based on the assumption that users of the method would counter this criticism with arguments like:

"your reasoning may be sound, but the method works, so we are not impressed".

The purpose of the experiment reported in this chapter is, therefore, to show how the method could 'work' (that is, how it could give results which agreed, in the main, with existing wage structures) and still be a poor measuring instrument. The gist of the argument is that, for the very reason of the unusually high agreement between job evaluation results and wage structures, job measurements are likely to be contaminated by knowledge about the existing wage structure or by other 'external information' closely related to it.

In particular, the experiment was designed to show that external information, even where it is not in agreement with the job demands as defined by the method, may strongly contaminate the resulting ratings.

To this end, such information was simulated. The subjects taking part in the experiment were given descriptions of 32 jobs. They were to rate these jobs on the job factors defined by the Normalized Method and to compute a total score (weighted sum of the factor scores). Moreover, half of the subjects were given purely fictitious information about the general level of each job. This information was given in the form of a total score interval in which the job was said to have been put at some other occasion. Of a particular job the experimenter would say, for example, that its total score had been somewhere in between 40 and 54 points. This was offered as side-information, which the subject was free either to use or to ignore. The experimenter said, in fact: "Of course, this level does not limit your freedom of rating. We want *your* judgment on the job".

The other subjects served as controls: they were given identical tasks but did not receive the 'external information'. The subjects in the experimental situation reported that they felt free to disregard the (fictitious) information and that this was exactly what they intended to do, as this information generally was not compatible with their own judgment. These explicit statements from all of the experimental subjects made the results all the more impressive.

In the experiment, two sets of 32 job descriptions have been used (about half of them were common to both sets). Four subjects (two experimental, two control) were presented with one set, six subjects (three experimental, three control) were presented with the other one. The results with the first set (four subjects) were inconclusive: some of the ratings in the experimental situation were highly correlated with the fictitious information, but so were the ratings of the control subjects.

The slight difference between the two situations was not significant. The results of an analysis of covariance, with the ratings in the control situation as covariates, were negative.

The results with the other set of job descriptions (presented to six subjects) were surprisingly clear-cut. In figure 4.3 curves have been drawn representing the relationship between the fictitious information (T-info) and the ratings on the job factor *knowledge* (Ke) (which is, because of its large weight, heavily represented in the total score and accordingly, highly correlated with it). While the control curves (indicated by a C) are almost level (C-I-2 is an exception due to a flaw in the experimental set-up), the experimental curves are all clearly slanting upwards.

The results of an analysis of variance (see table 4.3) show that the influence of the external information (T-informatie) is unexpectedly great, as measured in terms of amount of variance accounted for. In one case (E-III) it is even greater than the influence of the other independent variable (ratings on the factor *knowledge* (Kennis) as given by the committee that wrote the job descriptions), which was used as an operationalization of the 'internal' or 'legitimate' information about the job.

The difference between the experimental and the control situations is already evident from the analysis of variance results (tables 4.3 and 4.4). An analysis of covariance has been performed, with the ratings of matched control subjects as covariates, to correct for any relation between external information and the ratings of the subjects in the control situation. This was another way to compare the results of the experimental situation with those from the control situation and to estimate the significance of the difference (see table 4.5). The results were positive for all three experimental subjects.

It was concluded that the high correlation between the job evaluation results and the existing wage structure (which made people state that 'the method works') must be due to contamination of the purportedly objective rating procedure. The experimental results confirmed the hypothesis that job analysts – even highly experienced ones – are apt to determine their ratings partly in accordance with preconceived ideas about the general level of the jobs to be rated. Even those subjects, who clearly worked analytically (i.e. they rated each job factor separately and they did not correct the total score afterwards, even when this differed from their expectations), got total scores which were significantly correlated with the fictitious experimental information received.

Although the subject of the experiment was a specific job evaluation procedure, the findings are of wider import. A purportedly 'analytical' and additive rating procedure was shown to be highly contaminated by fictitious information. What is more, the subjects were instructed in actual words not to trust this information more than their own judgment. They took this instruction seriously, but, all the same, the information had its effect, without any of the subjects being aware of it.

Summary of chapter V: Final considerations

After a summary of the preceding chapters, some suggestions for future research are given. As one possible problem to be investigated a comparative reliability study is recommended, in which subjects who rate analytically are compared with subjects who begin with an estimate of the total score.

As an alternative approach to the study of the rating process in job evaluation procedures the methods of 'systematic introspection' and 'thinking aloud' (see e.g. DE GROOT 1966) were proposed as particularly suited to an investigation into the process of rating and evaluating jobs.

As a final statement of policy it was recommended that there should always be a close co-operation between practitioners and social scientists during the developmental phases of methods like job evaluation procedures.

BIJLAGE 1

GRADERINGSFORMULIER

FUNCTIES	KENNIS 5	ZELF-STANDIGH. 6	CONTACT 2	GEZAG 2	UITDRUKK. VAARDIGH. 1	BEWEGINGS-VAARDIGH. 2	MATERIAAL & MACH. GEV. 1	ZWAARTE 1	HOUDING 1	OPLET-TENDH. 1	SFEER 3	PERSOONL. RISICO 1	AFBREUK RISICO 4	TOTAAL AANTAL PUNTEN
1														1
2														2
3														3
4														4
5														5
6														6
7														7
8														8
9														9
10														10
11														11
12														12
13														13
14														14
15														15
16														16
17														17
18														18
19														19
20														20

BIJLAGE 2

KENNISGRADERINGEN, T-SCORES EN T-INFORMATIE (PRAKTIJKCIJFERS EN PROEFPERSOON-GRADERINGEN) VOOR DE FASEN I EN III

Functie-nummers*	T-informatie	Praktijk		C-I-1		C-I-2		E-I-1		E-I-2		C-III		E-III	
		K**	T***	K	T	K	T	K	T	K	T	K	T	K	T
1	C	4½	75	4½	64	4	65½	3½	57	3	49	3	69	3½	58½
2	C	3	80	3½	75	4	72½	4	68	2½	57	3½	77½	2½	54½
3	C	1½	33	1	22	1½	29	1½	31	1½	34	1½	29½	1½	42½
4	A	4	68	4½	70	5	75	4	78	4	63	3½	70	5	82
5	B	2½	58	3½	62	2½	41½	4	58	3	58	4	84	3	65
6	D	3	77	3	71	3½	66	2	35	2	46	2½	56½	1½	47½
7	D	1½	35	1	24	1	19	1	22	1	29	1	26	½	22½
8	C	1	27	1½	21	1	18½	1½	22	1	18	1½	26	1	29
9	D	2	33	2½	39	2	32	3½	39	1½	26	2	48	1½	34½
10	B	1	34	1½	36	1	21½	1	21	1	29	1½	35½	1	30
11	D	4	51	3	43	2½	37	3½	45	3	49	4½	75	1½	28
12	C	4½	62	3½	57	3½	52½	4	47	3	48	3	61	2	44
13	D	1	37	2½	51	1½	28	2	32	1½	37	2½	58½	1½	32½
14	B	3½	62	4½	74	4½	72½	4½	66	3½	60	4	81	4½	67
15	A	2	38	2½	46	1½	25½	4	52	2½	47	3	59½	2½	50½
16	B	3	61	3½	63	4	58½	4½	54	3	52	3	69½	2½	51½
17	D	3½	81	4	76	3½	60	2	40	2	37	3½	84½	2	52
18	A	2½	55	3	57	3½	49	4	58	2	38	3	66	4	66
19	D	2½	44	2½	45	1½	19½	2	27	2	34	3	55½	½	16½
20	D	4	77	3	62	3	41½	3½	50	2½	51	5	100	1	32½
21	C	3½	61	3½	66	3	42	2½	40	2½	39	4	81	2½	49½
22	B	4½	86	4	69	3½	55½	4	59	3½	54	4½	82½	2½	53½
23	A	1½	38	2	49	2½	49	2½	46	3	47	2½	55½	2	49
24	B	1½	38	2	43	1½	28½	3	47	2	44	1½	42½	1½	46½
25	B	4	75	3½	69	3	45½	5	72	3½	58	4	87½	3	63
26	A	3½	65	3	65	4	68	4	65	2½	51	3½	71	2½	50½
27	C	2½	38	2½	32	2	34½	2½	37	2	34	2½	53½	1½	24
28	A	3	46	3	42	3½	45½	4½	61	3½	56	3	61½	4	64½
29	B	2	33	2½	36	3	41	3½	45	2½	33	3	56	4	57½
30	A	4	49	4	55	4½	64	5	66	4	52	4	76	6½	89½
31	C	2	45	2½	44	2½	38	3	43	2	43	3	62	1½	40
32	A	1½	36	2½	39	1½	27½	2½	29	3½	55	2	41	2	28½

* De functienamen behorende bij de nummers zijn gegeven in Bijlage 4.

** K = graderingscijfer voor Kennis

*** T = T-score

BIJLAGE 3

KENNISGRADERINGEN, T-SCORES EN T-INFORMATIE (PRAKTIJKCIJFERS EN PROEFPERSOON-GRADERINGEN) VOOR DE FASE II

Functie-nummers*	T-informatie	Praktijk		C-II-1		C-II-2		E-II-1		E-II-2	
		K**	T***	K	T	K	T	K	T	K	T
1	C	4½	75	2½	60½	2	55	4	71	4½	66½
2	D	3½	62	2	49	1½	47	3½	66	2	37
3	C	1½	34	1	34	1½	39	3	55	1	37
4	A	3½	60	3	58½	2½	59½	4	57	4	52½
5	B	2	49	2	45	2½	50	—	—[1]	2	37
6	D	3	77	3	72	3	62	3½	85	2½	55½
7	D	1½	35	1	26½	1	26½	1½	36	½	28½
8	C	2½	49	2	47	1½	38	2½	47	2	46
9	D	2	33	2	38	1½	33½	2	41	1½	33½
10	B	1	34	1	27	1	25	1½	39	2	51½
11	A	5	75	4	69	4½	69½	5½	81	5	77
12	C	4½	62	2½	47½	2½	42	4½	68	3	45½
13	C	4½	73	3	54½	3½	58½	4½	78	3	47
14	B	3½	62	3½	65½	3½	64	3½	63	5	65
15	A	2	34	2½	45½	2½	45	3	46	2½	45½
16	B	3	61	3	62	3	61	3	69	3½	64½
17	A	2½	62	3½	79½	3	74½	3	86	4	81
18	B	4	76	3	57½	2½	57	4	94	2½	54½
19	D	3	60	1½	45	1½	43	3	63	2	54
20	D	4	77	2	46½	2½	49½	3½	77	3	53
21	C	3½	61	2	39	2½	41½	3½	64	3	55
22	A	2	32	3	50	2½	38½	2	37	3	41
23	D	2	46	1	31½	1	29	2½	57	1	32½
24	A	4	76	2½	57½	3	51½	5	92	4½	71
25	B	4	75	3	63	3	58½	4	79	4	65
26	B	2	45	1½	31½	1½	35½	3	61	2	40½
27	D	2	49	1½	37½	1½	35½	2	52	2	45
28	A	3	46	2	40	2½	41½	3	48	3	49
29	B	2	33	2	39	2½	40	2½	39	2	29½
30	A	4	49	4½	65	5	58	4½	59	4½	62½
31	C	2	45	1½	41½	2	45½	2	49	2	34
32	C	2	33	2	37	3	43½	2	38	1½	25½

* De functienamen behorende bij de nummers zijn gegeven in Bijlage 4.

** K = graderingscijfer voor Kennis

*** T = T-score

[1] De proefpersoon heeft bij deze functie geweigerd een oordeel te geven. Bij de variantie- en covariantie-analyses is het gemiddelde van de functies met dezelfde T-informatie gebruikt.

BIJLAGE 4

NAMEN VAN DE FUNCTIES WAARVAN DE BESCHRIJVINGEN IN HET EXPERIMENT AAN DE PROEFPERSONEN TER GRADERING WERDEN AANGEBODEN

Functie-nummer	Fasen I en III	Fase II
1	Carrosserieschilder	idem
2	Chauffeur hoofdrioolzuiger	Buitenbandenreparateur
3	Controleur toegangsbewijzen	Syphonpomper fabrieksterrein
4	Blesser-houtmeter	Timmerman III
5	Putboorder	Meteropnemer
6	Ontsmetter	idem
7	Straatreiniger	idem
8	Railreiniger	Motormaaier
9	Filterwachter (snelfilters)	idem
10	Steenschuurder	idem
11	Magazijnbediende	Machinist (coagulatiegebouw)
12	Wachtsman hoofdrioolgemaal	idem
13	Wachtsman cokesverdeelhuis	Magazijn-loketbediende (chemische afd.)
14	Grofbankwerker	idem
15	Bediener mantrolley	Lichtdrukker
16	Stoker centraal-generatoren	idem
17	Boomsnoeier	Pontvoerder kabelpont (handkracht)
18	Grondboorder	Zweminstructeur bij een onoverdekt zwembad
19	Wasmachine-bediener	Opperman (bij een metselaar)
20	Bediende ontsmettings- en wasinstallatie	idem
21	Waterstandopnemer	idem
22	Watergasmaker	Zalfbereider
23	Bestuurder locomotor	Reiniger (kleine gemeente)
24	Bediende veemarkt	Houtsorteerder-bosarbeider
25	Spoormaker-thermietlasser	idem
26	Chauffeur ladderauto	Meethulp
27	Binnenbandenreparateur	Wegwerker (trambaan)
28	Filterwachter (langzame filters)	idem
29	Meterkastenreparateur	idem
30	Carburateur-reviseur	idem
31	Stalknecht	idem
32	Reiniger kabelmoffen	Bordenzetter en -verzorger

LITERATUUR

ADAMS, E. W., R. F. FAGOT en R. E. ROBINSON (1965): A theory of appropriate statistics. Psychometrika *30* (1965) 99-127.

ALLPORT, G. W. (1937): Personality; a psychological interpretation. New York, Henry Holt and Company, 1937.

BEHAN, F. L., en R. A. BEHAN (1954): Comment. Football numbers (continued). American Psychologist *9* (1954) 262-3.

BRODBECK, M. (1963): Logic and scientific method in research on teaching. Blz. 44-93 (chapt. 2) in: GAGE, N. L. (ed.), Handbook of research on teaching. Chicago, Rand McNally and Company, 1963.

CATTELL, R. B., en K. DICKMAN (1962): A dynamic model of physical influences demonstrating the necessity of oblique simple structure. Psychological Bulletin *59* (1962) 389-400.

CATTELL, R. B., en W. SULLIVAN (1962): The scientific nature of factors: a demonstration by cups of coffee. Behavioral Science *7* (1962) 184-93.

CHESLER, D. J. (1948): Reliability of abbreviated job evaluation scales. Journal of Applied Psychology *32* (1948) 622-8.

CHESLER, D. J. (1949): Abbreviated job evaluation scales developed on the basis of "internal" and "external" criteria. Journal of Applied Psychology *33* (1949) 151-7.

COMMISSIE WERKCLASSIFICATIE GEMEENTEN. Voorbeelden van gradering ten behoeve van de classificatie van gemeentelijke functies, dl. A, B en C. 's-Gravenhage, Vereniging van Nederlandse Gemeenten, z.j.

COMREY, A. L. (1950): An operational approach to some problems in psychological measurement. Psychological Review *57* (1950) 217-28.

COMREY, A. L. (1951): Mental testing and the logic of measurement. Educational and Psychological Measurement *11* (1951) 323-34.

COOMBS, C. H. (1953): Theory and methods of social measurement. Blz. 471-535 (chapt. 11) in: FESTINGER, L., and D. KATZ (eds.), Research methods in the behavioral sciences. New York, The Dryden Press, 1953.

COOMBS, C. H. (1964): A theory of data. New York; London, John Wiley and Sons, Inc., 1964.

CRONBACH, L. J. (1954): Report on a psychometric mission to Clinicia. Psychometrika *19* (1954) 263-70.

CRONBACH, L. J., en G. C. GLESER (1957): Psychological tests and personnel decisions. Urbana, University of Illinois Press, 1957.

CRONBACH, L. J., N. RAJARATNAM en G. C. GLESER (1963): Theory of generalizability: a liberalization of reliability theory. British Journal of Statistical Psychology *16* (1963) 137-63.

DESKUNDIGENCOMMISSIE VOOR WERKCLASSIFICATIE (1952): V3000: Ontwerp genormaliseerde methode van werkclassificatie, dl. I: Gezichtspunten; dl. II: Graderingsschema's. Hoofdcommissie voor de Normalisatie in Nederland (HCNN), september 1952.

DESKUNDIGENCOMMISSIE VOOR WERKCLASSIFICATIE (1959): NEN 3000: Genor-

maliseerde methode van beschrijving en gradering van werkzaamheden ten behoeve van werkclassificatie en andere doeleinden, dl. I: Gezichtspunten; dl. II: Graderingsschema's. Nederlands Normalisatie-Instituut (NEN), april 1959.

GROOT, A. D. DE (1946): Het denken van den schaker; een experimenteel-psychologische studie. Amsterdam, N.V. Noord-Hollandse Uitgevers Maatschappij, 1946. (Dissertatie G.U. Amsterdam).

GROOT, A. D. DE (1953): Genormaliseerde werkclassificatie. Mens en Onderneming *7* (1953) 401-15.

GROOT, A. D. DE (1954a): Scientific personality diagnosis. Acta Psychologica *10* (1954) 220-41.

GROOT, A. D. DE (1954b): De discussie over werkclassificatie. Mens en Onderneming *8* (1954) 443-7.

GROOT, A. D. DE (1961): Methodologie; grondslagen van onderzoek en denken in de gedragswetenschappen. 's-Gravenhage, Mouton & Co., 1961.

GROOT, A. D. DE (1962): Quelques réflexions sur les problèmes de la psychométrie. Blz. 61-71 in: Les problèmes de la mesure en psychologie – Symposium de l'Association de psychologie scientifique de langue française, Amsterdam 1961. Paris, Presses Universitaires de France, 1962.

GROOT, A. D. DE (1964): De kernitem-methode voor de bepaling van de caesuur voldoende/onvoldoende. Paedagogische Studiën *41* (1964) 425-40.

GROOT, A. D. DE (1966): Perception and memory versus thought: some old ideas and recent findings. Blz. 19-50 in: KLEINMUNTZ, B. (ed.), Problem solving. New York; London, John Wiley and Sons, Inc., 1966.

GUTTMAN, L. (1955): The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. British Journal of Statistical Psychology *8* (1955) 65-81.

HARMAN, H. H. (1960): Modern factor analysis. Chicago, University of Chicago Press, 1960.

HAYS, W. L. (1963): Statistics for psychologists. New York; Toronto; London, Holt, Rinehart and Winston, 1963.

HAZEWINKEL, A. (1964): Enkele beschouwingen betreffende de bijdrage van een component tot een samengestelde variabele. Statistica Neerlandica *18* (1964) 325-39.

HUNDT, D. (1965): Die Arbeitsplatz- und persönliche Bewertung als Kriterien zur Bestimmung des Leistungslohnes; Ergebnisse und Auswertungen einer Untersuchung in der chemischen Industrie. Bern; Stuttgart, Verlag Hans Huber, 1965. (Schriften zur Arbeitspsychologie in Verbindung mit der Schweizerischen Stiftung für Angewandte Psychologie herausgegeben von Hans Biäsch, nr. 7)

ITZ, H. J. (1960): Werkclassificatie. Tijdschrift voor Efficiëntie en Documentatie *30* (1960) 305-7.

JAHODA, M., M. DEUTSCH en S. W. COOK (1951): Research methods in social relations, with especial reference to prejudice, I: Basic processes. New York, The Dryden Press, 1951.

JONGE, H. DE (1964): Inleiding tot de medische statistiek, 2: Klassieke methoden; 2e herz. dr. Leiden, Nederlands Instituut voor Praeventieve Geneeskunde, 1964. (Verhandeling van het N.I.P.G., nr. 48)

JONGE, H. DE, en G. WIELENGA (1963): Statistische methoden voor psychologen en sociologen; 2e dr. Groningen, J. B. Wolters, 1963.

KAISER, H. F. (1958): The varimax criterion for analytic rotation in factor analysis. Psychometrika *23* (1958) 187-200.

KALVERAM, K. TH. (1965): Die Veränderung von Faktorenstrukturen durch "simultane Uberlagerung". Archiv für die gesamte Psychologie *117* (1965) 296-305.

KEMENY, J. G. (1959): Mathematics without numbers. Daedalus *88* (1959) 577-91.

KEMENY, J. G., en J. L. SNELL (1962): Mathematical models in the social sciences. Boston; Toronto, Ginn and Company, 1962.

KENDALL, M. G. (1962): Rank correlation methods; 3rd ed. London, Ch. Griffin and Co., Ltd., 1962.

KOHNSTAMM, PH. (1942): Over het probleem van "psychische metingen" in 't algemeen en van "intelligentiemetingen" in 't bijzonder. Tijdschrift voor Philosophie *4* (1942) 323-44 en 547-80.

KOUWER, B. J. (1955): Gewetensproblemen van de toegepaste psychologie. Groningen; Djakarta, J. B. Wolters, 1955. (Inaugurele rede, Groningen)

LAWSHE JR., C. H. (1945): Studies in job evaluation, 2: The adequacy of abbreviated point ratings for hourly-paid jobs in three industrial plants. Journal of Applied Psychology *29* (1945) 177-84.

LAWSHE JR., C. H., en P. C. FARBRO (1949): Studies in job evaluation, 8: The reliability of an abbreviated job evaluation system. Journal of Applied Psychology *33* (1949) 158-66.

LAWSHE JR., C. H., en G. A. SATTER (1944): Studies in job evaluation, 1: Factor analyses of point ratings for hourly-paid jobs in three industrial plants. Journal of Applied Psychology *28* (1944) 189-98.

LIJFTOGT, S. G. (1966): De genormaliseerde methode van werkclassificatie; waardering en kritiek. Den Haag, Drukkerij Albani, 1966. (Dissertatie Utrecht).

LEVE, G. DE (1956): Enige statistische aspecten van de factoranalyse. Amsterdam, Mathematisch Centrum, Statistische Afdeling, september 1956. (Rapport S 210)

LORD, F. M. (1953): Comment. On the statistical treatment of football numbers. American Psychologist *8* (1953) 750-1.

LUCE, R. D., en H. RAIFFA (1957): Games and decisions; introduction and critical survey. New York; London, John Wiley and Sons, Inc., 1957.

MEEHL, P. E. (1954): Clinical versus statistical prediction; a theoretical analysis and a review of the evidence. Minneapolis, University of Minnesota Press, 1954.

MILLER, G. E. (1964): Evaluation in medical education; a new look. Journal of Medical Education *39* (1964) 289-97.

MYERS, J. H. (1958): An experimental investigation of "point" job evaluation systems. Journal of Applied Psychology *42* (1958) 357-61.

NAGEL, E. (1955): Principles of the theory of probability. Blz. 343-422 in: NEURATH, O., R. CARNAP, CH. MORRIS, International encyclopedia of unified science, Vol. 1, Part 2. Chicago, University of Chicago Press, 1955.

NEDERLANDS VERBOND VAN VAKVERENIGINGEN: Werkclassfiicatie. Amsterdam, N.V.V. z.j. (Reeks voorlichtingsbrochures op het gebied van bedrijfsorganisatie en beloningstechniek, nr. 1)

OTIS, J. L., en R. H. LEUKART (1954): Job evaluation; a basis for sound wage administration; 2nd ed. Englewood Cliffs, N. J., Prentice-Hall Inc., 1954.

OVERALL, J. E. (1964): Note on the scientific status of factors. Psychological Bulletin *61* (1964) 270-6.

RAJARATNAM, N. (1960): Reliability formulas for independent decision data when reliability data are matched. Psychometrika *25* (1960) 261-71.

RITTER, J. H. (1958): Het begrip waarderen in de werkclassificatie. Het PTT-bedrijf *8* (1958) 195-225.

SANDEE, J., en R. RUITER (1957): Beroepseisen van de industriearbeider. Economisch-Statistische Berichten *42* (1957) 944-6.

SATTER, G. A. (1949): Method of paired comparisons and a specification scoring key in the evaluation of jobs. Journal of Applied Psychology *33* (1949) 212-21.

SCHAAFSMA, W. (1961): Faktoranalyse. Groningen, Mathematisch Instituut van de Universiteit, 1961. (Report TW-8)

SIEGEL, A. E. (1964): Sidney Siegel: a memoir. Blz. 1-23 (Introduction) in: MESSICK, S., and A. H. BRAYFIELD (eds.), Decisions and choice; contributions of Sidney Siegel. New York, London; Toronto, McGraw-Hill Book Company, 1964.

SILVA, D. J. DA (1951): 9e Internationaal Efficiency Congres, Brussel, 5-11 juli 1951. Werkclassificatie. Tijdschrift voor Efficiëntie en Documentatie *21* (1951) 186-8.

SNIJDERS, J. TH. (1965): Het rapport van de psycholoog. Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden *20* (1965) 554-73.

SNYGG, D., en A. W. COMBS (1949): Individual behavior; a new frame of reference for psychology. New York, Harper and Brothers, Publishers, 1949.

STEVENS, S. S. (1946): On the theory of scales of measurement. Science *103* (1946) 677-80.

STEVENS, S. S. (1951): Mathematics, measurement, and psychophysics. Blz. 1-49 in: STEVENS, S. S. (ed.), Handbook of experimental psychology. New York, John Wiley and Sons, Inc.; London, Chapman and Hall, Ltd., 1951.

STEVENS, S. S. (1959): Measurement, psycho-physics, and utility. In: CHURCHMAN, C. W. and PH. RATOOSH (eds.), Measurement: definitions and theory. New York, John Wiley and Sons, Inc., 1959.

STICHTING VAN DEN ARBEID: Werkclassificatie als hulpmiddel bij de loonpolitiek. 's-Gravenhage, Stichting van den Arbeid, z.j. (1952)

STOUTHARD, PH. C. (1965): Methodologie en onderzoek. Sociale Wetenschappen *8* (1965) 249-65.

SUPPES, P., en J. L. ZINNES (1963): Basic measurement theory. Blz. 1-76 in: LUCE, R. D., R. R. BUSH and E. GALANTER (eds.), Handbook of mathematical psychology, Vol. I, chapters 1-8. New York; London, John Wiley and Sons, Inc., 1963.

THURSTONE, L. L. (1947): Multiple-factor analysis. A development and expansion of "The vectors of mind". Chicago, University of Chicago Press, 1947.

TIFFIN, J. (1951): Industrial Psychology. London, George Allen & Unwin Ltd., 1951.

TORGERSON, W. S. (1958): Theory and methods of scaling. New York, John Wiley and Sons, Inc.; London, Chapman and Hall, Inc., 1958.

TRYON, R. C. (1935): A theory of psychological components; an alternative to "mathematical factors". Psychological Review *42* (1935) 425-54.

VERNON, PH. E. (1965): Ability factors and environmental influences. American Psychologist *20* (1965) 723-33.

VISSER, M. J. DE (1963): Rechtvaardige beloningen, II. Economisch-Statistische Berichten *48* (1963) 1176-8.

VITELES, M. S. (1941): A psychologist looks at job evaluation. Personnel *17* (1941) 165-76 (Reprint. New York, American Management Association, z.j.)

WERKCLASSIFICATIE (1962): Praktijk van de werkclassificatie. Loontechnische Informaties (Algemene Werkgevers-Vereniging) 1962, 6, 25-8.

WERKCLASSIFICATIE (1965): Herziening van de werkclassificatie volgens de genormaliseerde methode (G.M. '64). Loontechnische Informaties (Algemene Werkgevers-Vereniging) 1965, 3, 13-9.

WESSEM, A. VAN (1952): Opzet en verantwoording van een werkclassificatie methode. Haarlem, H. D. Tjeenk Willink en Zoon N.V., 1952.

WIEGERSMA, S. (1954): Enkele problemen betreffende de genormaliseerde werkclassificatie. Mens en Onderneming *8* (1954) 81-9 en 193-9.

WIEGERSMA, S. (1958): Gezichtspunten en factoren in de genormaliseerde werkclassificatie. Mens en Onderneming *12* (1958) 200-8.

WILDE, G. J. S. (1963): Neurotische labiliteit gemeten volgens de vragenlijst-methode. Amsterdam, Uitg. F. van Rossen, 1963.

WRIGLEY, C. F. (1957): The distinction between common and specific variance in factor theory. British Journal of Statistical Psychology *10* (1957) 81-98.

WRIGLEY, CH. (1958): Objectivity in factor analysis. Educational and Psychological Measurement *18* (1958) 463-76.

OVER DIT BOEK:

In 'Werkclassificatie: een wetenschappelijk instrument?' wordt de genormaliseerde methode opgevat als meetinstrument en vanuit dat standpunt aan een theoretische analyse onderworpen. Deze analyse leidt tot de conclusie dat de methode als meetinstrument niet kan voldoen. De vraag, hoe het dan mogelijk is dat deze methode in de praktijk tot zulke bevredigende resultaten blijkt te leiden, wordt beantwoord door middel van een experimenteel onderzoek, waarbij wordt aangetoond dat ervaren taakanalisten in sommige gevallen in sterke mate in hun beoordeling van functies volgens de genormaliseerde methode beïnvloed kunnen worden door wat zij reeds van deze functies weten of menen te weten. Dit levert een gerede verklaring voor het feit dat een theoretisch niet verantwoorde methode toch tot resultaten voert die met de praktijksituatie overeenkomen.

OVER DE AUTEUR:

Dr. A. HAZEWINKEL studeerde aan de Universiteit van Amsterdam, waar hij in 1957 het doctoraal-examen in de psychologie aflegde. Na een half jaar werkzaam te zijn geweest op de psychologische dienst van de Koninklijke Nederlandse Hoogovens en Staalfabrieken te IJmuiden zette hij zijn studie in de Verenigde Staten van Amerika voort, en wel aan de University of Illinois.

Sinds 1960 is hij als wetenschappelijk medewerker verbonden aan het Nederlands Instituut voor Praeventieve Geneeskunde TNO te Leiden. In maart 1967 promoveerde hij aan de Universiteit van Amsterdam. De titel van zijn proefschrift, waarvan dit boek een handelsuitgave is, luidt 'De genormaliseerde methode van werkclassificatie als meetinstrument'. (Promotor Prof. Dr. A. D. de Groot).