Robustheit metrischer und nicht-metrischer Conjoint-Analyse auf der Grundlage simulierter Präferenzdaten

Diplomarbeit

Oliver Schilling

Betreuung und Begutachtung: Prof. Dr. J. Werner
Prof. Dr. H.J. Ahrens / Dr. K.E. Rogge

Psychologisches Institut der Universität Heidelberg
November 1993

Ich danke Herrn Prof. Dr. Werner für großzügige Betreuung und Unterstützung.

Inhaltsverzeichnis

Teil 1: Theoretische Grundlagen

1.1 Conjoint Measurement

Die Untersuchung, die auf den folgenden Seiten dargestellt wird, gehört zu jenem inhaltlichen Bereich sozialwissenschaftlicher Forschungsmethodik, der in der Literatur allgemein unter dem Oberbegriff *Conjoint Measurement* (CM) bzw. — auf deutsch — *verbundene Messung* zusammengefaßt wird. Während nun CM normalerweise als Teilbereich der Meßtheorie abgehandelt wird (vgl. Orth 1974; Roberts 1979), sind innerhalb dieses Bereich selbst wieder Teilbereiche zu unterscheiden, deren Fragestellungen oft nichts mehr miteinander zu tun zu haben scheinen. Es soll deshalb im folgenden zunächst ein kurzer Überblick über das CM versucht werden, der als Einführung in die Thematik dienen und den „Standort" der im Anschluß daran vorgestellten Untersuchung innerhalb dieser Thematik verdeutlichen soll.

Am allgemeinsten läßt sich das CM durch die ihm zugrundeliegende methodologische Problemstellung charakteriesieren: Es geht um Messung bzw. Analyse sogenannter *Produktstrukturen* (vgl. Roberts 1979, S. 197ff). Damit sind solche Meßobjekte (Stimuli) gemeint, die als mehrdimensional oder multiattributiv aufgefaßt werden können. Oder, um es in der Terminologie der Meßtheorie auszudrücken: Dem CM liegt ein empirisches Relativ $\langle A, R\rangle$ zugrunde, dessen Stimulusmenge A als cartesisches Produkt $\langle A_1 \times A_2 \times \ldots \times A_t\rangle$ aufgefaßt werden kann. Jede der Mengen A_l $(l = 1, 2, \ldots, t)$ gibt ein Attribut bzw. eine Dimension bzw. einen Faktor (alles Bergiffe, die in der einschlägigen Literatur gebraucht werden) der in A enthaltenen Stimuli an und enthält alle k möglichen Ausprägungen a_{lk} der(des) jeweiligen Dimension(Attributs/Faktors).

Es könnten z.B. verschiedene Berufsalternativen als multiattributiv aufgefaßt werden: Berufe unterscheiden sich hinsichtlich mehrerer Dimensionen, wie etwa „Einkommen", „Aufstiegschancen", „Freizeitmöglichkeiten" usw., ein bestimmter Beruf ist als Kombination von Ausprägungen dieser Attribute darstellbar, so daß man sagen kann, die Menge der Berufsalternativen habe eine Produktstruktur.

Ausgangspunkt einer typischen CM-Studie wären nach diesem Beispiel die Präferenzurteile einer Versuchsperson über verschiedene Berufsalternativen (nicht notwendig über alle möglichen): *Die Datenbasis des CM bilden ordinale Urteile über multiattributive Alternativen.*

Dieser letzte Satz könnte allerdings auf Widerspruch stoßen: So unterscheiden z.B. Green & Srinivasan (1978, S.111f) zwischen nonmetrischen und metrischen Skalen für die abhängige Variable in der Conjoint-Analyse (zum Begriff

der Conjoint-Analyse siehe Abschnitt 1.1.2). Auf das hier gegebene Beispiel übertragen bedeutet das, daß die Versuchsperson ihre Urteile über die Berufsalternativen sowohl als ordinales Präferenzurteil im engeren Sinne — z.B. durch Rangreihung der Alternativen oder durch Paarvergleich von jeweils 2 Berufen — abgeben kann, oder aber auch z.B. auf Ratingkskalen, was, gemäß der allgemein üblichen Interpretation von Ratingskalen, bedeutet, daß die Urteile zumindest näherungsweise auf Intervallskalenniveau vorliegen. Dennoch halte ich die Behauptung für gerechtfertigt, daß letztere Möglichkeit sozusagen nicht die Eigentümlichkeit des CM erfaßt. So schreiben z.B. Krantz & Tversky (1971, S. 152) in einer Arbeit, die durchaus als Klassiker des CM bezeichnet werden kann: „*The key feature of the conjoint-measurement approach is that only the ordinal aspects of the data are required to be compatible with the proposed composition principle.*“ Dieses Zitat verweist auf die Axiome der verbundenen Meßstrukturen, die der Gegenstand des Bereichs des CM sind, der nachfolgend als „Axiomatisches CM“ dargestellt werden wird: Tatsächlich sind die gebräuchlichsten, in einschlägigen Darstellungen wiedergegebenen Axiomatisierungen verbundener Meßstrukturen (vgl. das in Abschnitt 1.1.1 angegebene Beispiel einer zweidimensionalen, additiv-verbundenen Struktur) mathematisch präzise Formulierungen der Eigenschaften einer Relation im empirischen Relativ, welche auf „ordinalen Urteilen“ beruht, d.h. einer Relation, die lediglich Präferenzen zwischen verschiedenen Stimuli des empirischen Relativs wiedergibt.

Die ordinalen Urteile werden in der CM-Literatur oft als *Nutzenurteile* bezeichnet (z.B. Backhaus 1990, S. 345ff). Viele Anwendungen des CM — v.a. der Conjoint-Analyse und häufig in Marktforschungsstudien — behandeln tatsächlich einen inhaltlichen Nutzenaspekt der multiattributiven Stimuli, d.h. mit den Präferenzurteilen soll der unterschiedliche „Gesamtnutzen“ verschiedener Stimuli erfaßt werden. Es ist aber festzuhalten, daß man es bei empirischen Fragestellungen nicht unbedingt mit Beurteilungen des Nutzens verschiedener Stimulusalternativen im (inhaltlichen) Wortsinn zu tun haben muß, sondern eben nur mit einer empirischen Relation R, die von der Versuchsperson zwischen diesen Stimuli hergestellt wird und die als mindestens ordinal im Sinne von Präferenzurteilen interpretiert werden kann. Ich werde jedoch im folgenden der Nutzen-(Utility-)Terminologie treu bleiben.

Ziel des CM ist Skalierung: Ausgehend von den ordinalen Gesamtnutzenurteilen sollen Intervallskalen für die „Teilnutzenwerte“ der Attribute/Dimensionen konstruiert werden, aus deren „Komposition“ sich dann auch intervallskalierte Gesamtnutzenwerte berechnen lassen. Das CM kann deshalb als dekompositionelles Verfahren bezeichnet werden, der ordinal erfaßte Gesamtnutzen eines Stimulus wird zunächst in Teilnutzenbeiträge der jeweils diesem Stimulus zugehörigen Ausprägungen der Attribute „dekomponiert“. Dabei sind zwei Gesichtspunkte zu unterscheiden:

1. Die *„Kompositionsregel“*, d.h. die Frage, welches (mathematische) Modell die Zusammensetzung des Gesamturteils aus den Bewertungen der einzelnen Attribute wiedergibt: Sind — z.B. — die Attribute einfache additiv verknüpft oder gilt eine kompliziertere Verknüpfungsfunktion?
2. Die *Skalierung* als solche, d.h. das Auffinden geeigneter numerischer Lösungen, die die empirisch ermittelte Präferenzordnung am besten wiedergeben.

Entsprechend dieser beiden Fragestellungen zerfällt das CM — in der Literatur seit den siebziger Jahren — in 2 Teilbereiche, die von vielen Autoren (z.B. Emery & Barron 1979; Nickerson & McClelland 1984) mit den Begriffen *„axiomatisches“* und *„numerisches“* CM unterschieden werden.

Dagegen wählen Green & Srinivasan (1978) „Conjoint Measurement“ und „Conjoint Analysis“ als begriffliches Gegensatzpaar für diese beiden Teilbereiche und scheinen so die Unabhängigkeit der beiden voneinander unterstreichen zu wollen. Sie befinden sich damit im Gegensatz zu der schon erwähnten Arbeit von Krantz & Tversky (1971): Dort wird argumentiert, daß gerade in der simultanen Beantwortung der beiden obigen Fragestellungen ein besonderer Wert des CM für psychologische Forschung läge. Denn psychologische Variablen werden oft durch physikalische Messungen repräsentiert, die als Indikator der Ausprägung der zugrundeliegenden psychologischen Variablen angesehen werden, und es kann kaum kontrolliert werden, inwieweit die physikalische Skala tatsächlich die zu messende psychologische Variable repräsentiert: *„Hence, the best one can hope for is to find a monotonic relation between the controllable physical scale and the relevant psychological variable“* (ebd., S. 151). D.h. die Skalen, auf denen diese psychologischen Variablen solchermaßen indirekt gemessen werden, besitzen allenfalls ordinales Niveau, und damit entsteht ein grundsätzliches Problem bei der Überprüfung psychologischer Theorien, in denen funktionale Beziehungen zwischen diesen Variablen postuliert werden. Als Beispiel verweisen Krantz & Tversky auf die Theorie von Hull (1952), wonach zwischen den psychologischen Variablen R(„response strength“), D(„drive“), H(„habit strength“) und K(„incentive“) die funktionale Relation $R = D \times H \times K$ gelte, und deren Modifikation durch Spence (1956), die dagegen die Beziehung $R = H(D+K)$ postuliert. Welche der beiden Relationen — welche Kompositionsregel — gilt nun? Eine empirische Prüfung, die anhand empirisch erhobener Werte für R, D, H, K zu zeigen sucht, welche der beiden Gleichungen den Werten genügt, ist eigentlich nicht möglich, solange die beteiligten Variablen nur auf Ordinalskalen gemessen werden. Das Problem, wie auf der Basis ordinaler Messungen psychologischer Variablen funktionale Zusammenhänge zwischen diesen konfundiert werden können, kann durch die verbundene Messung umgangen werden: *„An answer to this question is given by the conjoint-measurement approach, which attempts to solve the*

measurement and the composition problems simultaneously, by constructing measurement scales for the relevant variables so that the proposed composition principle is satisfied“ (ebd., S. 152).

Verläßt man die innerhalb der CM-Literatur gebräuchliche Terminologie (und auch die meßtheoretische Perspektive), dann kann CM als ein Verfahren bezeichnet werden, das eine abhängige Variable, deren Ausprägungen auf mindestens ordinalem Niveau erfaßt wurden — die ordinalen Gesamtnutzenurteile über multiattributive Stimuli —, mit mehreren unabhängigen Variablen, die mindestens in kategorialen Ausprägungen vorliegen — die Ausprägungen des jeweiligen multiattributiven Stimulus auf den einzelnen Dimensionen —, in Zusammenhang bringt.

Es folgt zunächst die allgemeine Darstellung der beiden Teilbereiche des CM, ehe die spezifischeren theoretischen Grundlagen der Untersuchung, die den eigentlichen Gegenstand dieser Arbeit bildet, erörtert werden.

1.1.1 Axiomatisches Conjoint Measurement

Axiomatisches CM (ACM) behandelt sozusagen den meßtheoretischen Teil des CM bzw. den ersten der beiden oben erwähnten Gesichtspunkte.

Die **Meßtheorie** untersucht Bedingungen für die Meßbarkeit von Eigenschaften und damit die Frage, welche empirischen Sachverhalte durch welche numerischen Strukturen adäquat abgebildet werden können. Dabei bedeutet *Messen* die *homomorphe Abbildung eines empirischen Relativs in ein numerisches Relativ*. Ein empirisches Relativ besteht aus (mindestens) einer Menge empirischer Entitäten — den Meßobjekten — und mindestens einer empirischen Relation, die zwischen diesen Entitäten entsprechend der zu messenden Eigenschaft hergestellt werden kann. Das numerische Relativ dazu besteht aus (mindestens) einer Zahlenmenge und mindestens einer numerischen Relation, die zwischen den Elementen der Zahlenmenge gilt[1]. Formal: Gegeben ein empirisches Relativ $\langle A, R\rangle$ (wobei A die Menge empirischer Meßobjekte und R die Relation zwischen denselben benennt) und ein numerisches Relativ $\langle Z, S\rangle$ (Z steht für die Zahlenmenge, S für die auf derselben gültige numerische Relation) und eine Abbildung f von $A \rightarrow Z$ — dann ist f homomorph wenn gilt (für alle $a, b \in A$): aRb gdw $f(a)Sf(b)$. Damit eine homomorphe Abbildung eines empirischen in ein numerisches Relativ gegeben ist, müssen die Relationen des empirischen Relativs bestimmte Eigenschaften erfüllen: Diese Eigenschaften können mathematisch exakt in Form von *Axiomen* formuliert und aus denselben kann dann das sogenannte *Repräsentationstheorem* gefolgert

[1]Eine ausführlichere Erläuterung der in diesem Absatz eingefürten meßtheoretischen Grundbegriffe gibt Orth (1974, S.14ff)

werden. Dieses besagt, daß bei Gültigkeit der genannten Axiome eine homomorphe Abbildung des empirischen ins numerische Relativ (d.h. eine Skala) existiert. Darüber hinaus wird aus den Axiomen das *Eindeutigkeitstheorem* bewiesen: Es gibt die zulässigen Transformationen an, durch die verschiedene homomorphe Abbildungen (Skalen), die zu dem gegebenen empirischen Relativ existieren, ineinander überführt werden können (d.h. das Skalenniveau der jeweiligen Messung). Das Ganze (empirisches Relativ, numerisches Relativ, Axiome, Repräsentationstheorem, Eindeutigkeitstheorem) wird als *Meßstruktur* bezeichnet.

Im Fall des CM hat das empirische Relativ also die Form $\langle A_1 \times A_2 \times \ldots \times A_t, R\rangle$. Es folgt das Beispiel einer zweidimensionalen additiv-verbundenen Meßstruktur (nach Roberts 1979):

Definition: Es seien A_1 und A_2 nichtleere Mengen und R eine binäre Relation auf $A_1 \times A_2 = A$. Das Relativ $\langle A_1 \times A_2, R\rangle$ ist eine additiv verbundene Struktur wenn folgende Axiome erfüllt sind:

A1 : $\langle A, R\rangle$ ist eine strikt schwache Ordnung, d.h. R ist asymmetrisch und negativ transitiv[2].

A2 : $\langle A, R\rangle$ erfüllt Unabhängigkeit, d.h. $(a_1, a_2)R(b_1, a_2) \Rightarrow (a_1, b_2)R(b_1, b_2)$ und $(a_1, a_2)R(a_1, b_2) \Rightarrow (b_1, a_2)R(b_1, b_2)$ für alle $a_1, b_1 \in A_1$ und $a_2, b_2 \in A_2$.

A3 : $\langle A, R\rangle$ erfüllt die Thomsen-Bedingung, d.h. $(a_1, b_2)E(b_1, c_2)$ und $(b_1, a_2)E(c_1, b_2) \Rightarrow (a_1, a_2)E(c_1, c_2)$ (wobei E als Äquivalenzrelation definiert ist: $aEb \Leftrightarrow\sim aRb$ und $\sim bRa$).

A4 : Jede streng begrenzte Standardfolge auf jeder der beiden Mengen A_1 und A_2 ist endlich (Archimedisches Axiom)[3].

A5 : $\langle A, R\rangle$ ist beschränkt lösbar auf beiden Komponenten, d.h. für alle $x, \underline{y}, \overline{y} \in A_1$ und $q, r \in A_2$ gilt: wenn $(\overline{y}, r)R(x, q)R(\underline{y}, r)$, dann existiert ein $y \in A_1$, so daß $(y, r)E(x, q)$ (entsprechendes gilt für die Komponente A_2).

A6 : Beide Komponenten von $\langle A, R\rangle$ sind wesentlich, d.h. (für A_1) es existieren $a_1, b_1 \in A_1$ und $a_2 \in A_2$ derart, daß $\sim (a_1, a_2)E(b_1, a_2)$ (entsprechend für die andere Komponente).

Repräsentationstheorem: Es sei $\langle A_1 \times A_2, R\rangle$ eine additiv-verbundene Struktur. Dann existieren reelle Funktionen f_1 auf A_1 und f_2 auf A_2 derart,

[2] R ist asymmetrisch, wenn gilt $aRb \Rightarrow\sim bRa$, und negativ transitiv, wenn gilt $\sim aRb$ und $\sim bRc \Rightarrow\sim aRc$ für alle $a, b, c \in A$ ($\sim aRb$ bedeutet: „nicht aRb“).

[3] Erklärung siehe Roberts (1979, S. 217).

daß für alle (a_1, a_2) und $(b_1, b_2) \in A_1 \times A_2$ gilt:

$$(a_1, a_2)R(b_1, b_2) \Leftrightarrow f_1(a_1) + f_2(a_2) > f_1(b_1) + f_2(b_2).$$

Eindeutigkeitstheorem: Sind g_1 und g_2 zwei andere reelle Funktionen auf A_1 bzw. A_2, die obige Eigenschaft (Repräsentationstheorem) erfüllen, dann existieren reelle Zahlen α, β, γ derart, daß gilt:

$$g_1(a_1) = \alpha f_1(a_1) + \beta \text{ und } g_2(a_2) = \alpha f_2(a_2) + \gamma.$$

Diese Axiomatisierung ist sozusagen die klassische für die zweidimensionale Struktur, die — mit leichten Abweichungen — in allen einschlägigen Lehrbüchern zu finden ist. Es ist jedoch nicht die einzig mögliche für diesen Fall (vgl. z.B. Roberts 1979, S. 222ff) und natürlich müssen im Fall mehrer Attribute die obigen Axiome z.T. neu formuliert werden. Für andere, nicht-additive Kompositionsregeln wurden entsprechende Axiomatisierungen entwickelt.

Die „axiomatische“ Analyse gegebener empirischer Präferenzdaten ist nun hauptsächlich mit dem Testen von Axiomen beschäftigt. Z.B. könnte von einer Versuchsperson eine Rangfolge über 16 zweidimensionale Stimulusalternativen mit jeweils 4 Stufen pro Dimension hergestellt worden sein, wie sie durch die 3×3-Tafel in Abbildung 1.1 wiedergegeben wird. Die 4 Spalten der Figur entsprechen den 4 Stufen des Attributs A_1, die 4 Zeilen denen des Attributs A_2, in den Kästchen stehen die Rangwerte der zweidimensionalen Stimuli. Zum Test (z.B.) des Unabhängigkeitsaxioms müssen nun alle Zeilen und Spalten miteinander verglichen werden: A_2 ist unabhängig von A_1, wenn für jedes Paar von Spalten gilt, daß in allen Zeilen (dieses Spaltenpaares) diesselbe Präferenzrichtung besteht. A_1 ist unabhängig von A_2, wenn umgekehrt innerhalb eines jeden Zeilenpaares dieselbe spaltenweise Präferenzrichtung gegeben ist. Diese letztere Bedingung ist in obiger Rangordnungsmatrix an einer Stelle verletzt! Mathematisch exakt bestünde die Überprüfung des Unabhängigkeitsaxioms in der Durchführung bzw. Testung *aller* möglichen Paarvergleiche, die sich aufgrund des Axioms **A2** ergeben. Die Verletzung der Unabhängigkeit im obigen Beispiel wirkt sich auf 3 dieser Paarvergleiche aus[4].

ACM behandelt — wie schon gesagt wurde — den Gesichtspunkt der *Kompositionsregel*: Durch die Testung von Axiomen wird gezeigt, daß eine bestimmte, im Repräsentationstheorem spezifizierte Repräsentation (=Kompositionsregel) der empirischen Daten durch die Skalen f_l für die einzelnen Attribute möglich ist. So behandeln z.B. Krantz & Tversky (1971) 4 verschiedene Kompositionsregeln für dreidimensionale Strukturen — nämlich die additive, die multiplikative, die distributive und die dual-distributive. Für ein empirisches Relativ $\langle A_1 \times A_2 \times A_3, R\rangle$ ergeben sich nach diesen Kompositionsregeln folgende For-

[4]Es gilt für die Rangwerte r_{ij} (i soll hier der Zeilen-, j der Spaltenindex sein): $r_{24} < r_{34}$, während in allen anderen Spalten zwischen der 2. und 3. Zeile die umgekehrte Präferenzrichtung besteht, so daß alle Paarvergleiche, die sich nach **A2** aus $r_{24} < r_{34}$ ergeben, falsch werden ($r_{24} < r_{34} \Rightarrow r_{21} < r_{31}$, $r_{24} < r_{34} \Rightarrow r_{22} < r_{32}$, $r_{24} < r_{34} \Rightarrow r_{23} < r_{33}$).

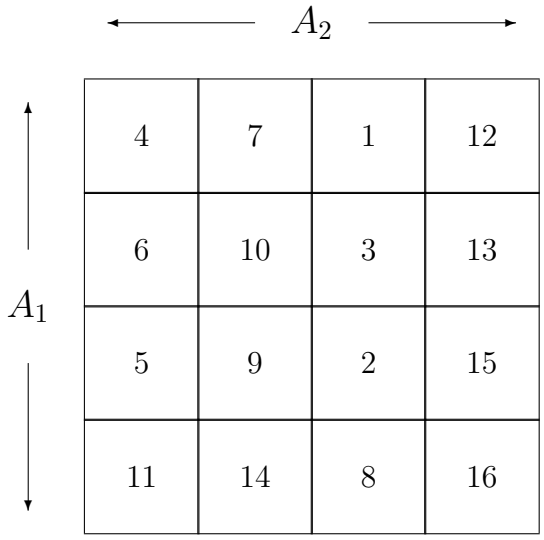

Abbildung 1.1: Beispiel einer Rangordnung über die 16 Stimuli einer 4^2-Struktur (4×4-Tafel).

mulierungen des Repräsentationstheorems:

> Additive Regel: $(a_1, a_2, a_3)R(b_1, b_2, b_3) \Leftrightarrow f_1(a_1) + f_2(a_2) + f_3(a_3) > f_1(b_1) + f_2(b_2) + f_3(b_3)$.
> Multiplikative Regel: $(a_1, a_2, a_3)R(b_1, b_2, b_3) \Leftrightarrow f_1(a_1) \times f_2(a_2) \times f_3(a_3) > f_1(b_1) \times f_2(b_2) \times f_3(b_3)$.
> Distributive Regel: $(a_1, a_2, a_3)R(b_1, b_2, b_3) \Leftrightarrow (f_1(a_1) + f_2(a_2)) \times f_3(a_3) > (f_1(b_1) + f_2(b_2)) \times f_3(b_3)$.
> Dual-distributive Regel: $(a_1, a_2, a_3)R(b_1, b_2, b_3) \Leftrightarrow f_1(a_1) \times f_2(a_2) + f_3(a_3) > f_1(b_1) \times f_2(b_2) + f_3(b_3)$[5].

Die Entscheidung, welche der 4 Regeln eine gegebene Produktstruktur (am besten) repräsentiert, wird nun dadurch getroffen, daß eben die zur jeweiligen Repräsentation notwendigen Axiome getestet werden[6]. Krantz & Tversky (S. 159 und S. 163) geben hierfür diagnostische Flußdiagramme an, die eine be-

[5]Es sind selbstverständlich noch andere Formulierungen der distributiven und dual-distributiven Regel möglich, je nach Permutation der Funktionen f_1, f_2, f_3 in den obigen Ausdrücken.

[6]Es müssen allerdings nur 3 Kompositionsregeln getestet werden, da die additive in die multiplikative Repräsentation transformiert werden kann: Wenn $g_1 = e^{f_1}, g_2 = e^{f_2}, g_3 = e^{f_3}$, dann folgt aus dem additiven Repräsentationstheorem: $(a_1, a_2, a_3)R(b_1, b_2, b_3) \Leftrightarrow g_1(a_1) \times g_2(a_2) \times g_3(a_3) > g_1(b_1) \times g_2(b_2) \times g_3(b_3)$ — (vgl. Roberts 1979, S.213).

stimmte Abfolge der zu testenden Axiome festlegen und so zur gewünschten Entscheidung führen.

Dabei ist allerdings zu bedenken, daß der Nachweis einer Repräsentation über die axiomatische Testung keinen Beweis für die Struktur des tatsächlichen psychologischen Urteilsprozesses liefert. D.h. es ist nicht nachzuweisen, daß der Pb „in seinem Kopf“ einzelne Nutzenwerte z.B. addiert, um zu einem Urteil über den Gesamtnutzen zu gelangen. Es ist durchaus möglich, daß empirische Präferenzordnungen die Axiome für mehrere Repräsentationen erfüllen!

Gleichwohl ergibt sich der praktische Nutzen des ACM für psychologische Forschung v.a. dann, wenn das theoretische Interesse der Struktur des kognitiven Prozesses, der zur Bildung der Präferenzurteile führt, gilt — und nicht so sehr der Vorhersage dieser Urteile, wozu dann konkrete numerische Lösungen für die im Repräsentationstheorem angeführten Funktionen f_l gefunden werden müssen (das Thema des numerischen CM). ACM kann zeigen, daß eine bestimmte Theorie über den funktionalen Zusammenhang der Attribute im psychologischen Urteilsbildungsprozeß — eine bestimmte Kompositionsregel — von den empirischen Daten nicht widerlegt wird, und es kann damit dazu dienen, zwischen alternativen Kompositionsregeln zu entscheiden.

Es hat nun allerdings eine Kontroverse darum gegeben, ob für diesen Zweck tatsächlich die axiomatische Testung — und damit das ACM überhaupt — notwendig sei: Emery & Barron (1979) argumentieren, daß auch „numerische“ Prozeduren (vgl. Abschnitt 1.1.2) in der Lage seien, sowohl die datengenerierende Kompositionsregel zu finden, als auch zwischen verschiedenen alternativen Regeln zu entscheiden. Die Autoren führen dazu eine Simulation durch, bei der 4^3-Strukturen (d.h. 3 Attribute mit jeweils 4 Ausprägungen/Stufen) auf Additvität, Distributivität und Dual-Distributivität untersucht werden. Dazu werden zunächst Werte für die Stufen der Attribute „künstlich“ erzeugt und dann nach den 3 Kompositionsregeln zu „wahren“ Gesamtnutzenwerten verrechnet. Letztere werden in Rangwerte umgewandelt, an diesen Rangwerten (der multiattributiven Stimuli) wird erstens eine axiomatische Testung unter Zuhilfenahme des oben erwähnten diagnostischen Flußdiagramms nach Krantz & Tversky (1971) und zweitens eine „numerische“ Analyse vermittels des Programms MONANOVA (Näheres dazu S. 19) bzw. Modifaktionen desselben für die distributive und dual-distribitive Kompositiosregel durchgeführt. MONANOVA funktioniert über die Minimierung eines sogenannten Streß-Wertes — die Hypothese der Autoren lautet, daß sowohl minimaler Streß, als auch ein weiteres Kriterium namens PRECAP („predictive capability“), nämlich der Prozentsatz korrekter Paarvergleiche im skalierten Datensatz (relativ zum ursprünglichen, die Datenbasis bildenden Datensatz), zur Identifikation der Kompositionsregel ausreichen. Die Autoren fanden ihre Hypothese durch die Ergebnisse weitestgehend bestätigt.

Die Arbeit von Emery & Barron hat jedoch einigen Widerspruch nach sich gezogen. Nickerson & McClelland (1984) kritisieren Emery & Barron in 3 Punkten: Erstens begrenzt die Verwendung fehlerfreier Daten die Aussagekraft der Ergebnisse, da man es in der Praxis psychologischer Forschung in der Regel eben mit fehlerhaften Daten zu tun hat. Zweitens ist auch die Entscheidung zwischen additiver, distributiver und dual-distributiver Kompositionsregel von geringer praktischer Bedeutung, da die letzteren beiden Regeln in der bisherigen psychologischen Theorienbildung nur sehr selten zur Anwendung kamen und aus praktischer Perspektive v.a. der Nachweis einer additiven Repräsentation wichtig sei. Drittens aber sind Goodness-of-fit-Kriterien, wie die von Emery & Barron verwendeten, nur mit Vorsicht zu interpretieren, wenn es um die Entscheidung für eine Kompositionsregel geht, da solche Kriterien dazu tendieren, auch solchen Modellen einen guten „Fit“ zu attestieren, die mit der tatsächlichen Kompositionsregel, nach der die Daten generiert wurden, nicht übereinstimmen. Zur Begründung dieses letzeten Punktes verweisen Nickerson & McClelland auf mehrere Arbeiten (Anderson & Shanteau 1977; Birnbaum 1973; Shanteau 1977; Zeleny 1976), in denen gezeigt wurde, daß in Regressions- und Varianzanalysen Korrelationen zwischen vorhergesagten und tatsächlichen Werten Modelmißspezifikationen oft durch hohe Werte verdecken, v.a. dann, wenn das theoretische Modell ein lineares ist.

In ihrer eigenen Arbeit konzentrieren sich Nickerson & McClelland auf den Vergleich von ACM und numerischem CM hinsichtlich zweier Aufgaben: (1) korrekte Identifikation einer additiven Kompositionsregel bei fehlerhaften Daten und (2) Zurückweisung der additiven Kompositionsregel, wenn sie dem datengenerierenden „wahren“ Modell nicht entspricht. In der hier gebotenen Kürze kann das wesentliche Ergebnis ihrer Simulation damit zusammengefaßt werden, daß bei leichten Verletzungen des additiven Modells bei der Datengenerierung, die bei fehlerbehafteten Daten vorkommen sollten, dennoch durch ACM und numerisches CM gleichermaßen korrekt Additivität diagnostiziert werden kann, während jedoch MONANOVA auch bei schwersten Verletzungen des additiven Modells mit äußerst niedrigen Streß-Werten die Gültigkeit des additiven Modells anzuzeigen scheint, so daß die numerische Methode also praktisch immer die einmal unterstellte Kompositionsregel bestätigt.

Erwähnt sei schließlich auch noch die Arbeit von Timmermans (1980), der empirische Daten über die Päferenz von Konsumenten bezüglich Einkaufszentren untersuchte: Untersucht wurde eine 3^3-Struktur (3 Attribute von Einkaufszentren, jeweils auf 3 Stufen variiert) und getestet wurden die Daten mittels des (numerischen) Programms UNICON (Roskam 1974) nach 8 verschiedenen Kombinationsregeln. Anhand der durch UNICON erzielten Streß-Werte war es praktisch unmöglich, für jede der 18 Versuchspersonen eine klare Entscheidung zu treffen, welche Kombinationsregel die jeweils angemessene sei — die Streß-Werte waren allemal sehr niedrig. Timmermans schließt: *„However,*

this experiment has also clearly indicated the weakness of numerical conjoint measurement models to diagnose the composition rules individuals apply in decision-making tasks“ (S. 299).

Faßt man die Veröffentlichungen zu dieser Kontroverse zusammen — was hier freilich nur sehr verkürzt geschehen ist, weil anderenfalls der thematische Rahmen der vorliegenden Arbeit zu weit verlassen worden wäre — so läßt sich wohl sagen: Es bleibt dabei, daß der axiomatischen Testung die Aufgabe zukommt, Entscheidungen über die Angemessenheit bestimmter (v.a. der additiven) Kompositionsregeln zu treffen. Axiomatisches und numerisches CM sind nicht zwei konkurrierende Methoden der Analyse multiattributiver Präferenzdaten, sondern sozusagen 2 Seiten der einen Medaille CM. Im idealtypischen Fall einer CM Studie folgen diese beiden Seiten schrittweise aufeinander (vgl. Krantz & Tversky 1971, S.166f; Nickerson & McClelland 1984, S.195). In konkreten, unter ökonomischen Einschränkungen stattfindenden Untersuchungen wird es — wie bereits gesagt — vom spezifischen wissenschaftlichen Interesse abhängen, ob Axiome getestet werden oder ob numerische Lösungen für die Skalierung gesucht werden. Dieser Zusammenhang bzw. die relative Bedeutung des ACM soll hier deshalb noch einmal hervorgehoben werden, weil sich ab dem nachfolgenden Abschnitt 1.1.2 die vorliegende Arbeit ganz und gar im Bereich des numerischen CM bewegen wird.

Das wohl schwerwiegendste Problem des ACM besteht im *Fehlen einer Fehlertheorie*: Die Frage ist, wieviel Verletzungen von Axiomen toleriert werden, bevor ein Axiom als nicht erfüllt betrachtet wird. Es wird in der psychologischen Forschung kaum der Fall fehlerfreier Daten auftreten, d.h. man wird immer davon ausgehen müssen, daß eine Versuchsperson Axiomverletzungen auch dann produziert, wenn die zu belegende Repräsentation eigentlich die „richtige“ ist. Eine Fehlertheorie, welche es z.B. erlauben würde, bei Annahme der Nullhypothese keiner Axiomverletzung in den „wahren Werten“ eine gegebene Anzahl von Axiomverletzungen auf Signifikanz zu überprüfen, wäre also notwendig. Dieser Punkt wurde in der oben geschilderten Kontroverse zwischen Anhängern der axiomatischen vs. Anhängern der numerischen Überprüfung von Kompositionsregeln oft von letzteren für die numerische Vorgehensweise ins Feld geführt (z.B. wieder Emery & Barron 1979, S. 204)[7]. Jedenfalls scheint es bis heute keine allgemeingültigen Regeln für die Anzahl zu tolerierender Axiomverletzungen zu geben, so daß darüber vom jeweiligen Untersucher mehr oder weniger durch willkürlich „nach Gefühl“ entschieden wird.

[7]Ein allerdings sehr schwaches Argument, denn — wie auch schon Nickerson & McClelland (1984, S.184) bemerken — es existiert auch für das numerische Vorgehen keine entsprechende Fehlertheorie: „*Procedures such as LINMAP and MONANOVA yield index-of-fit measures, C^* and stress, respectively, but do not provide statistical significance*“ (Umesh & Mishra 1990, S.34).

Es gibt aber immerhin Arbeiten, aus denen sich sozusagen gewisse Richtwerte für dieses Gefühl ableiten lassen und die vielleicht als Ansätze zur Entwicklung einer solchen Fehlertheorie dienen könnten. So liefern Arbuckle & Larimer (1976) in einer vielzitierten Untersuchung zwar noch nicht die gewünschte Fehlertheorie, aber immerhin Berechnungen über die Wahrscheinlichkeit, daß unter der Voraussetzung zufälliger (nicht additiver oder sonst einer Kompositionsregel folgender) Datengenerierung bestimmte Axiome erfüllt sind. Sie bestimmen bzw. schätzen die Anzahl derjenigen (aus allen möglichen) Rangordnungen einer zweidimensional Struktur, die bestimmte Axiome der Additivität erfüllen. Z.B. gibt Abbildung 1.1 eine von $(4 \times 4)!$ möglichen $4{\times}4$-Tafeln wieder, die bei systematischer Permutation der Werte in der Tafel gebildet werden könnten. Arbuckle & Larimer berechnen nun:

> $N(r, c)$ — d.i. die Anzahl aller möglichen $r{\times}c$-Tafeln, die alle Axiome der additiven Repräsentation für eine $r{\times}c$-Struktur (ein Attribut mit r, ein weiteres mit c Stufen) erfüllen.
> $N_i(r, c)$ — d.i. die Anzahl aller möglichen $r \times c$-Tafeln, die alle Cancellation-Axiome der i-ten Ordnung erfüllen[8].

Dabei kann nur $N_1(r, c)$ — die Anzahl der Tafeln, die das Unabhängigkeitsaxiom erfüllen — algorithmisch berechnet werden, für $N(r, c)$ und $N_2(r, c)$ geben die Autoren lediglich Schätzungen (bei gegebenen r, c) auf der Basis von Zufallsziehungen an. Alle diese Berechnungen können aber nur Indikatoren dafür liefern, wie „stark“ eine positiv verlaufener axiomatischer Test ist, d.h. es sind mittels dieser Zahlen Angaben über die Wahrscheinlichkeit, daß das entsprechende Axiom bzw. die Additivität insgesamt per Zufall durch die vorgefundene Rangordnung erfüllt wird, möglich.

Noch interessanter hinsichtlich der in Frage stehenden Schlußfolgerungen, die aus gegebenen Axiomverletzungen zu ziehen sind, scheinen die Arbeiten von Nygren (1985a; 1985b; 1986): Sein Ansatz verfolgt zweierlei, nämlich einmal Berechnungen der bei zufälliger Datengenerierung zu erwartenden Axiomverletzungsraten für bestimmte Axiome und zum andern — und v.a. — eine differenziertere Betrachtung der Axiomverletzungen, aus der dann Schlüsse über

[8]Der Begriff „cancellation“ — deutsch „Aufhebung“ bzw. „Kürzung“ — benennt eine Gruppe von Axiomen, diese lassen sich nach der Anzahl der in ihrer Wenn-Bedingung enthaltenen Relationen ordnen. Von den am Anfang dieses Kapitels gezeigten Axiomen sind 2 Cancellation-Axiome: Das Unabhängigkeitsaxiom ist das Cancellation-Axiom 1. Ordnung, die Thomsen-Bedingung stellt das Cancellation-Axiom 2. Ordnung dar und wird in der Literatur oft als Double Cancellation bezeichnet (genaugenommen wird die Thomsen-Bedingung erst dann zur Double Cancellation, wenn die Äquivalenzrelation E durch die Ordnungsrelation R ersetzt wird — wie Roberts (1979, S. 220) bemerkt, folgt die Thomsen-Bedingung aus der Double Cancellation wenn R eine strikt schwache Ordnung ist). Es können noch weitere Cancellation-Axiome höherer Ordnung formuliert werden, diese haben jedoch kaum praktische Bedeutung für das CM.

deren wahrscheinliche Ursache abzuleiten wären. Nygren behandelt dreidimensionale (3^3-, 4^3-) Strukturen und das Hauptaugenmerk seiner Arbeit (v.a. Nygren 1985b) gilt der Unterteilung der möglichen Verletzungen des Unabhängigkeitsaxioms in qualitativ unterschiedliche Fälle (sogenannte „dominant failures“ und „tradeoff failures“) — woraus dann Rückschlüsse darüber möglich sein sollen, ob vorgefundene Verletzungen darauf zurückzuführen sind, daß der Faktor, dessen Unabhängigkeit getestet wurde, irrelevant in dem Sinne, daß er keinen Einfluß auf das ordinale Gesamtnutzenurteil hat, ist, oder darauf, daß der Einfluß dieses Faktors nicht additiv und also die additive Repräsentation nicht angemessen ist. Eine genauere Darstellung dieser nicht unkomplizierten Untersuchungen würde allerdings den Rahmen der vorliegenden Arbeit sprengen.

1.1.2 Numerisches Conjoint Measurement

Das numerische CM behandelt nun den Skalierungsaspekt: Es werden numerische Lösungen $f_l(a_{lk})$ für jede Stufe k jedes Attributs l gesucht (vgl. das Repräsentationstheorem S. 5). Parallel zum Begriff „numerisches CM“ wird auch — wie eingangs erwähnt — die auf Green & Srinivasan (1978) zurückgehende Bezeichnung *„Conjoint-Analyse“* für alle die Verfahren, die dieses leisten, verwendet: *„Die Conjoint-Analyse ist ein Verfahren, das auf der Basis empirisch erhobener Gesamtnutzenwerte versucht, den Beitrag der einzelnen Komponenten zum Gesamtnutzen zu ermitteln“* (Backhaus 1990, S.345/6).

Es ist weitgehend dem persönlichen Geschmack überlassen, welche der beiden Bezeichnungen man wählt. „Numerisches CM“ hat den Vorteil, daß damit der inhaltliche Zusammenhang der bezeichneten Verfahren zum CM überhaupt deutlicher wird, weswegen dieser Titel hier für die allgemeine Darstellung des CM bzw. zur Gegenüberstellung der beiden Teilbereiche gewählt wurde. „Conjoint-Analyse“ scheint dagegen besser zu beschreiben, was im numerischen CM getan wird, indem allein schon der Begriff die gemeinten Verfahren in den Rahmen bekannter multivariater *Analyse*methoden — wie z.B. die Varianzanalyse — stellt. Es wird deshalb im folgenden der Begriff Conjoint-Analyse (CA) verwendet.

Die CA übergeht — wie im letzten Abschnitt schon dargestellt — das Problem der Kompositionsregel bzw. sie setzt eine bestimmte Repräsentation (meist die additive) voraus und modelliert die Daten nach diesem Modell. Backhaus (1990) unterscheidet bei der CA die in Abbildung 1.2 dargestellten 5 Schritte, die im folgenden dem Überblick über die mit der CA verbundenen Problemstellungen zugrundegelegt werden sollen. Detailfreudiger als Backhaus führen Grenn & Srinivasan (1978) durch das Gebiet, die Autoren haben ihre Darlegungen auch in einem „Update“ auf den neusten Stand gebracht (Green &

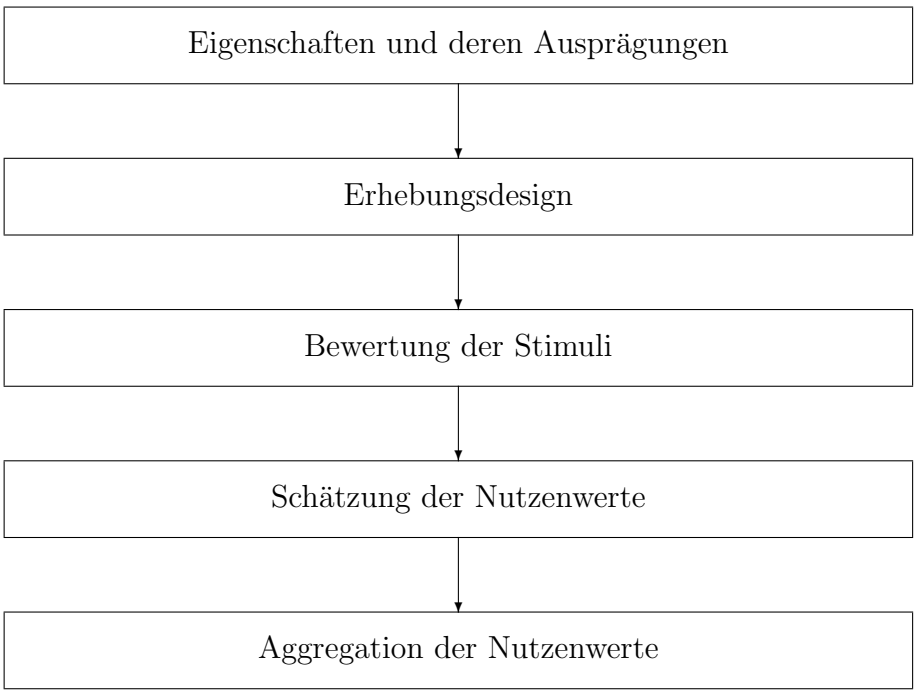

Abbildung 1.2: Die Schritte der CA (nach Backhaus 1990, S. 347).

Srinivasan 1990) — diese beide Arbeiten sind auch eine wertvolle Quelle bei der Literatursuche. Es ergibt sich natürlich aus dem im vorangegangenen Abschnitt Dargestellten, daß die ersten beiden dieser Schritte genausogut für das ACM von Belang sind: Der Platz der axiomatischen Testung läge in der indealtypischen CM-Studie zwischen dem 2. und 3. Schritt in Abbildung 1.2.

Es müssen also zunächst RELEVANTE EIGENSCHAFTEN bzw. Attribute der zu untersuchenden Produktstruktur bzw. der Stimulusmenge A (vgl. S. 1) ausgewählt und für jede dieser Eigenschaften Ausprägungen bzw. Stufen bestimmt werden, aus deren Kombinationen sich diejenigen Stimuli aus A ergeben, die der Versuchsperson dann tatsächlich präsentiert und über die von dieser die (ordinalen) Urteile gefällt werden. Man könnte auch sagen: Es muß ein theoretisches Modell gebildet werden, das unabhängige Variablen spezifiziert, welche die abhängige Variable — nämlich die ordinalen Urteile über die Stimuli A — beeinflussen, und es müssen Abstufungen dieser unabhäbgigen Variablen ausgewählt werden, anhand derer das Modell getestet werden soll. Es muß allerdings an dieser Stelle noch nicht eine bestimmte Kompositionsregel festgelegt werden (wenngleich die CA praktisch sowieso immer unter der Voraussetzung der additiven Repräsentation arbeitet, wie noch zu zeigen sein wird)! Mehr ist zu diesem Punkt im Grunde nicht zu sagen, Regeln wie die von

Backhaus (1990, S. 348), daß z.B. die Anzahl der Eigenschaften und ihrer Ausprägungen aus erhebungstechnischen Gründen begrenzt werden sollte, da der Befragungsaufwand exponentiell mit der Zahl der Eigenschaftsausprägungen wachse, verstehen sich eigentlich von selbst[9].

Sodann ist ein ERHEBUNGSDESIGN festzulegen. Dieser Punkt betrifft die Umsetzung des im obigen ersten Schritt gebildeten theoretischen Modells in der Konstruktion und Auswahl multiattributiver Stimuli, die zur Datenerhebung präsentiert werden sollen. Es ist hier zu unterscheiden:

(1) Hinsichtlich der Art der zu präsentierenden Stimuli zwischen der Profilmethode und der Zwei-Faktor-Methode. Bei der **Profilmethode** werden der Versuchsperson vollständige Stimulusprofile präsentiert, d.h. jeder Stimulus besteht aus der Kombination je einer Ausprägung aller Eigenschaften. Bei der **Zwei-Faktor-Methode** werden sogenannte Trade-Off-Matrizen gebildet: Die Stimuli bestehen aus der Kombination je einer Ausprägung von nur 2 Eigenschaften, für jedes mögliche Paar von Eigenschaften wird eine Trade-Off-Matrix gebildet. Bei 4 Eigenschaften mit jeweils 3 Stufen erhielte man nach der Profilmethode $3^4 = 81$ zu präsentierende Stimuli, die jeweils aus der Kombination der Stufen aller 4 Attribute gebildet werden. Bei der 2-Faktoren-Methode erhält man $\binom{4}{2} = 6$ 3×3-Matrizen, bei denen jeweils die Kombinationen aus den Ausprägungen zweier Attribute bewertet werden müssen. Backhaus (1990, S.351) stellt fest, daß bei praktischen Anwendungen meist die Profilmethode bevorzugt werde. Der Grund dafür dürfte in der größeren Realitätsnähe der nach der Profilmethode gebildeten Stimuli liegen: Zwar besteht ein Vorteil der Zwei-Faktor-Methode darin, daß die Aufgabe für die Versuchsperson leichter zu bewältigen ist, da sie sozusagen jeweils nur 2 Attribute im Auge behalten muß, jedoch wird dieser Vorteil um den Preis sehr dürftiger und unrealistischer Stimulusbeschreibungen erzielt, die ihrerseits Probleme aufwerfen: Man könnte z.B. fragen, inwieweit der psychologische Urteilsbildungsprozeß bei solchen zweidimensionalen Stimulusbeschreibungen denjenigen „abbildet“, der bei realen multidimensionalen Stimuli abläuft bzw. ob letzterer nicht qualitativ verschieden vom erste-

[9]Dies gilt allerdings nicht für Backhaus' dritten Gesichtspunkt (a. a. O.), daß nämlich die ausgwählten Eigenschaften unabhängig sein sollten. Zwar ist die Begründung, daß eine Verletzung dieser Bedingung dem additven Modell widerspräche, richtig, aber genau die Frage der Kompositionsregel ist — wie schon in Abschnitt 1.1.1 gezeigt wurde — ein Problem der CA. Eine „Lösung“ dadurch, daß einfach noch vor der Datenerhebung solche Daten gefordert werden, die mit der additiven Repräsentation kompatibel sind, ignoriert das Problem eher, als daß sie es tatsächlich löst. Auch beschneidet eine solche Einschränkung möglicherweise die Einsatzmöglichkeiten der CA in unnötiger Weise: In der im Anschluß an Teil 1 dargestellten Untersuchung wird gerade auch die Effizienz der CA unter Bedingungen, in denen die Unabhängigkeitsforderung verletzt ist (durch Interaktion zwischen Attributen) getestet werden.

ren ist. Auch ergibt sich bei vielen Anwendungen das ganz praktische Problem, daß zweidimensionale Stimuli nur durch verbale Beschreibungen und nicht etwa durch realistische Abbildungen von Gegenständen etc. dargeboten werden können (was wiederum einen gewissen Realitätsverlust beinhaltet). Eine ausführliche Diskussion der Vor- und Nachteile der beiden Methoden findet sich bei Green & Srinivasan (1978, S. 107f).

(2) Hinsichtlich der Anzahl der zu präsentierenden Stimuli zwischen vollständigen und reduzierten Designs. **Vollständige Designs** wären die soeben beschriebenen: Z.B. die Präsentation aller 81 Stimuli der 3^4-Struktur. Wie dieses Beispiel zeigt, führen vollständige Designs bei wachsender Zahl von Attributen und/oder Stufen schnell zu einer sehr großen Anzahl von Stimuli und damit zu der Gefahr, daß die Versuchsperson überfordert wird und in hohem Maße „fehlerbehaftete" Antworten liefert. Präsentiert man die Stimuli nach der Profilmethode, so bietet hier der Rückgriff auf **reduzierte faktorielle Designs** einen Ausweg: Es existiert eine große Zahl von Veröffentlichungen — v.a. aus den fünfziger und sechziger Jahren — die die Konstruktion solcher experimenteller (varianzanalytischer) Versuchspläne behandeln, die eine Reduzierung der Anzahl der Zellen des Experiments um den Preis eines Informationsverlustes ermöglichen. Während vollständige experimentelle Designs — welche soviele Zellen enthalten, wie sich aus den Kombinationen der experimentell variierten Abstufungen der unabhängigen Variablen ergeben — unkorrelierte Schätzungen sämtlicher Haupteffekte und Wechselwirkungen der unabhängigen Variablen zulassen, werden bei reduzierten Designs einzelne Effekte, nämlich ein Teil der bzw. alle Interaktionen, ignoriert. Beim Einsatz reduzierter faktorieller Designs wird also von vorneherein davon ausgegangen, daß bestimmte (Interaktions-) Effekte vernachlässigbar sind. Damit liegt es nahe, diese in der CA einzusetzen, denn die dort in den meisten Fällen vorausgesetzte additive Kompositionsregel enthält ja lediglich die Haupteffekte der Attribute und keinerlei Interaktionen zwischen denselben. Die „kürzestmöglichen" reduzierten Designs, die lediglich eine unkorrelierte Schätzung der Haupteffekte erlauben, sind die sogenannten „orthogonal arrays" (vgl. Green 1974, S. 63)[10]. Z.B. kann für die 3^4-Struktur (anstatt des vollständigen Designs mit 81 Stimuli) das in Abbildung 1.3 gezeigte auf 9 Stimuli reduzierte Design verwendet werden. Die Arbeit, der es entnommen ist (Addelman 1962a), liefert praktisch das gesamte Rüstzeug für den überwiegenden

[10]Bose & Bush (1952, S. 508) geben die folgende exakte Definition: „*A $k\times N$ matrix A, with entries from a set Σ of $s \geq 2$ elements, is called an orthogonal array of strength t, size N, k constraints and s levels if each $t\times N$ submatrix of A contains all possiple $t\times 1$ column vectors with the same frequency λ. The array may be denoted by (N, k, s, t).*" Dies aber nur der Vollständigkeit wegen, eine ausführliche Diskussion dieser Definition würde wohl vom Thema der vorliegenden Arbeit weit weg führen.

	A_1	A_2	A_3	A_4
Stimulus 1:	0	0	0	0
Stimulus 2:	0	1	1	2
Stimulus 3:	0	2	2	1
Stimulus 4:	1	0	1	1
Stimulus 5:	1	1	2	0
Stimulus 6:	1	2	0	2
Stimulus 7:	2	0	2	2
Stimulus 8:	2	1	0	1
Stimulus 9:	2	2	1	0

Abbildung 1.3: Reduziertes faktorielles 3^4-Design (nach Addelman 1962a: „basic plan 2“) mit den Attributen $A_1 \ldots A_4$. 0,1,2 stehen für die jeweils 3 Stufen der 4 Attribute.

Teil der CA-Anwendungen: Es wird dort nicht nur gezeigt, wie orthogonale reduzierte Designs für symmetrische Experimente (d.h. solche, in denen die Anzahl der Stufen pro unabhängiger Variable jeweils gleich ist) konstruiert werden, Addelman liefert darüber hinaus in mehreren „basic plans“ einige der grundlegendsten solcher Designs und gibt schließlich Regeln an, nach denen aus den „basic plans“ reduzierte Designs für asymetrische Experimente (mit ungleichen Anzahlen von Stufen über die Faktoren) gewonnen werden können.
Werden die Stimuli nach der Zwei-Faktor-Methode präsentiert, so sind die Möglichkeiten, die Aufgabe für den Probanden zu verkürzen, begrenzt. Eine Reduzierung kann hier mittels unvollständiger Blockdesigns („incomplete block designs“) erreicht werden. Da dieser Fall aufgrund der geringeren Praxisrelevanz der Zwei-Faktor-Methode aber eher selten auftritt und für die vorliegende Arbeit ebenfalls nicht von Belang ist, sei hierzu auf die Darstellung von Green (1974, S.64f) verwiesen.

Die Bewertung der Stimuli durch die Versuchsperson kann auf verschiedene Arten erfolgen: Nämlich **(1)** durch **Rangreihung**, **(2)** über **Ratingskalen** oder **(3)** mit **Paarvergleichen**. Es wurde vorne (S. 1) bereits eine gewisse Problematik, die die Verwendung von Ratingskalen für die CA darstellt, angesprochen: Der Fall einer auf diese Weise erzeugten metrischen (Intervall-) Skala (vgl. Green & Srinivasan 1978, S. 111) für die abhängige Variable beraubt sozusagen die CA ihrer Eigentümlichkeit, die eben in der Analyse von lediglich ordinalen Gesamtnutzenurteilen besteht. Werden diese Urteile mit Ratingskalen erhoben und sind somit als metrisch aufzufassen, dann steht einer Auswertung mittels bekannter metrischer Verfahren — bei denen eine intervallskalier-

te abhängige Variable und lediglich nominales Skalenniveau der unabhängigen Variablen vorausgesetzt ist (Varianzanalyse bzw. OLS-Regression mit Dummy-Variablen, siehe unten) — nichts mehr im Wege. Man könnte fragen, ob der pompöse Titel „Conjoint-Analyse“ überhaupt noch gerechtfertigt ist, bezeichnet er doch so nur noch einen speziellen Fall der Anwendung dieser metrischen Analysemethoden, der eben dadurch gekennzeichnet ist, daß hier die verschiedenen Ausprägungen der abhängigen Variablen an ein und derselben Versuchsperson erhoben wurden. Und man kann gleich weiter fragen, warum nicht prinzipiell die Urteile auf Ratingskalen erhoben werden, so daß sich der besondere Aufwand der nachfolgend noch darzustellenden nonmetrischen Verfahren gar nicht erst ergibt. Unter den Gründen, die Green & Srinivasan (1978, S. 112) für die nichtmetrischen Bewertungsmethoden nennen, scheint v.a. der folgende von Gewicht: *„Ranked data are likely to be more reliable, since it is easier for a respondent to say which he/she prefers more as compared to expressing the magnitude of his/her preference“*. Dieser Grund dürfte maßgeblich sein für die insgesamt häufige Anwendung der Rangreihung[11]. Es scheint so, daß das Für und Wider der einzelnen Bewertungsmethoden gar nicht allgemein zu erörtern ist, sondern von den inhaltlichen und situativen Gegebenheiten der jeweiligen Anwendung abhängt — einen Überblick über mögliche Vorteile der einen oder anderen Methode liefern Green & Srinivasan (1978, S. 112).

CA im engeren Sinne bedeutet nun Schätzung der Nutzenwerte. Es werden auf der Basis der ermittelten (Rang-) Werte zunächst *Teilnutzenwerte* für die Stufen der Attribute ermittelt, aus denselben lassen sich dann metrische (intervallskalierte) *Gesamtnutzenwerte*[12] Stimuli und die *relative Wichtigkeit der Attribute* ableiten. Ein additives Modell der CA kann so formuliert werden:

[11] Wittink & Cattin (1989) liefern Zahlen: Sie untersuchten 698 kommerzielle Anwendungen der CA (in der Marktforschung) aus dem Zeitraum 1971–1980 und 1062 solche Projekte zwischen 1981–1985. Sie fanden Rangreihung in 47% der Studien vor und 36% nach 1980, gegenüber 34% mit Ratingskalen vor und 49% nach 1980 — immerhin scheint sich hier sozusagen ein Trend für die Ratingskalen abzuzeichnen.

[12] Backhaus (1990, S. 346) schreibt:*„Die Auskunftsperson gibt also ordinale Gesamtnutzenurteile ab, aus denen durch die Conjoint-Analyse metrische Teilnutzenwerte abgeleitet werden“* — und gibt damit die übliche Auffassung bezüglich der Skalenqualität der CA-Parameterschätzungen wieder. Genaugenommen aber sind die geschätzten Teilnutzenwerte nur annähernd intervallskaliert:*„. . . it should be noted that even though the dependent variable is nonmetric, the estimated parameters tend to satisfy close to intervall-scaled properties, for typical values of n* [= Anzahl der Stimuli] *and T, the number of estimated parameters“* (Green & Srinivasan 1978, S. 112). Green & Srinivasan verweisen hier auf Colberg (1978), der dieses zeigt, dessen Arbeit aber mit nicht zugänglich war. Man kann es wohl bei der Auffassung, daß die geschätzten Teilnutzenwerte metrische Qualitäten besitzen, belassen, solange man sich nur der Vereinfachung, die in ihr enthalten ist, bewußt bleibt.

$$y_j = \sum_{l=1}^{t} \sum_{k=1}^{s_l} \beta_{lk} x_{lk} \tag{1.1.1}$$

wobei:

$$\begin{array}{rcl} y_j & = & \text{Gesamtnutzenwert für Stimulus } j. \\ \beta_{lk} & = & \text{Teilnutzenwert für Ausprägung } k \text{ von Attribut } l. \\ x_{lk} & = & \begin{cases} 1 & \text{falls bei } y_j \text{ das Attribut } l \text{ in der Ausprägung } k \text{ vorliegt} \\ 0 & \text{sonst.} \end{cases} \\ s_l & = & \text{Anzahl der Stufen des Attributs } l. \end{array}$$

Unter der (üblichen) Annahme nicht fehlerfreier Daten — d.h. wenn eine stochastische Komponente in das Modell integriert werden soll — ist Gleichung (1.1.1) um einen Fehlerterm zu ergänzen. Danach ergeben sich also die Gesamtnutzenwerte durch einfache Addition der Teilnutzenwerte β_{lk} [13]. Letztere sollen so geschätzt werden, daß die resultierenden Schätzungen der Gesamtnutzenwerte $\hat{y}_j$ möglichst gut den empirischen Rangwerten entsprechen. Es gibt nun eine gewisse Vielfalt von Rechenverfahren, die angewendet werden könnten:

(1) Wenn man unterstellt, daß die empirisch ermittelten Rangwerte Intervallskalenniveau besitzen, dann kann man die Teilnutzenwerte β_{lk} durch **Varianzanalyse** bzw. **OLS-Regression** schätzen — die sogenannten *metrischen Verfahren* der CA. Die Bekanntheit dieser Verfahren darf wohl hier vorausgesetzt werden, so daß sie an dieser Stelle nicht detailliert dargestellt zu werden brauchen (eine genauere Darstellung der OLS-Regression folgt in Abschnitt 1.2.2). Man könnte ein wenig erstaunt darüber sein, daß in der Literatur die Besonderheit eines mit Gleichung (1.1.1) gegebenen varianzanalytischen Designs keinerlei Beachtung findet. Es handelt sich schließlich um den Fall einer Beobachtungseinheit pro Faktorenstufenkombination mit Meßwiederholung über alle Faktoren! Die Besonderheiten eines solchen Designs wirken sich auf die Bestimmung der Interaktions- und Fehlervarianz aus (vgl. Bortz 1979, S. 396ff und S. 431ff) — da in der CA eigentlich nur die Schätzungen für die β_{lk} von Interesse sind, fallen diese Besonderheiten wohl hier nicht ins Gewicht. Bei der OLS-Regression sind die Stufen der Attribute als Dummy-Variablen zu kodieren, wie es in den x_{lk} in Gleichung (1.1.1) bereits ausgedrückt ist (zu

[13] Green & Srinivasan (1978, S.105) unterscheiden grundsätzlich 3 Modelle: Nämlich neben dem durch Gleichung (1.1.1) wiedergegebenen „part-worth function model“ noch ein „vector model“ und ein „ideal-point model“ — die letzteren beiden unterscheiden sich vom ersteren dadurch, daß sie nicht mehr nur kategoriale Ausprägungen der Attribute zu Voraussetzung haben, sondern kontinuierliche Skalen, auf denen diese Ausprägungen gemessen werden. Insofern sind diese Modelle als Spezialfälle des allgemeineren „part-worth function models“ aufzufassen. Sie haben in der Literatur nicht viel Widerhall gefunden, sieht man einmal von den umfangreichen Publikationen der beiden Autoren und davon, daß der genannte Artikel ein vielzitierter ist, ab. Sie werden im Rahmen der Darstellung des Verfahrens LINMAP in Abschnitt 1.2.1 erläutert werden.

Fragen der Kodierung vgl. Abschnitt 1.2.2, S. 50). In der bei Darstellungen der OLS-Regression üblichen Matrizenschreibweise lautet obiges Modell:

$$\mathbf{y} = (\mathbf{X}_1|\mathbf{X}_2|\ldots|\mathbf{X}_t)\left(\begin{array}{c}\boldsymbol{\beta_1}\\\hline\boldsymbol{\beta_2}\\\hline\vdots\\\hline\boldsymbol{\beta_t}\end{array}\right) + \boldsymbol{\epsilon} \tag{1.1.2}$$

wobei:

$\mathbf{y}$	=	der $n\times1$-Vektor der (metrischen) Gesamtnutzenwerte y_j $(j = 1, 2, \ldots, n)$.
$\mathbf{X}_l$	=	die $n\times(s_l-1)$-Kodiermatrix des Attributs l, welche spaltenweise die Dummy-Kodiervektoren für die Stufen des Attributs enthält.
$\boldsymbol{\beta_l}$	=	der $(s_l-1)\times1$-Vektor der Koeffizienten des Attributs l.
$\boldsymbol{\epsilon}$	=	$n\times1$-Vektor der Fehler der einzelnen Beobachtungen j.

Anders als in Gleichung (1.1.1) enthält diese Gleichung für jedes Attribut l mit s_l-Stufen nur $s_l - 1$ Kodiervariablen: Dies ergibt sich aus der Notwendigkeit, eine redundante Variable pro Attribut zu eliminieren, welche sich bei vollständiger Dummy-Kodierung — so wie es Gleichung (1.1.1) nahelegt — ergeben würde. Dadurch verändert sich auch die Bedeutung der β-Werte (gegenüber Gleichung (1.1.1)), die nun nicht mehr so einfach als Teilnutzenwerte interpretierbar sind (vgl. dazu S. 50). Es gelten hier die üblichen Annahmen über die Fehler als Minimalvoraussetzung für erwartungstreue Schätzer: der Erwartungswert der Fehler $E(\boldsymbol{\epsilon}) = \mathbf{0}$ und $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I}$, d.h. die Varianz-Kovarianz-Matrix der Fehler ist eine Diagonalmatrix mit konstanter Fehlervarianz in der Diagonalen (vgl. Werner 1993, S. 74).

(2) Die allgemeine Form der Problemstellung des CM bezieht sich aber auf ordinale Daten. Es wurden in der CA sogenannte *nonmetrische Verfahren* entwickelt, die lediglich die Annahme ordinalskalierter empirischer Daten voraussetzen. Das — zumindest in den siebziger Jahren — am häufigsten verwendete nonmetrische Verfahren ist **MONANOVA**, die monotone Varianzanalyse[14]. Es wurde in den sechziger Jahren von J. B. Kruskal entwickelt (Kruskal 1964a; 1964b; 1965) und soll hier — da es einerseits bei der in den folgenden Kapiteln beschriebenen Untersuchung nicht verwendet wird, andererseits aber innnerhalb der Gruppe der nonmetrischen Verfahren eine herausragende Rolle gespielt hat — nur in groben Zügen dargestellt werden. MONANOVA ist

[14]Vgl. z.B. wieder Wittink & Cattin (1989): Zwischen 1971 und 1980 zählten sie bei 24% der kommerziellen Anwendungen MONANOVA als Auswertungsmethode, zwischen 1981 und 1985 nur noch bei 11% — gegenüber 16% OLS/ANOVA vor 1980 und 54% danach. MONANOVA führt damit deutlich bei den nonmetrischen Verfahren.

ein iteratives Verfahren, dessen Grundprinzip durch die folgende Gleichung wiedergegeben werden kann:

$$p_j \xrightarrow{f_M} z_j \cong \hat{y}_j = \sum_{l=1}^{t} \sum_{k=1}^{s_l} b_{lk} x_{lk} \tag{1.1.3}$$

wobei:

p_j	=	empirische Rangwerte der Stimuli.
z_j	=	monoton angepaßte Rangwerte.
$\hat{y}_j$	=	vorhergesagte Gesamtnutzenwerte.
b_{lk}	=	geschätzter Teilnutzenwert für Ausprägung k von Attribut l.
f_M	=	monotone Transformation zur Anpassung der z-Werte an die y-Werte.
$\cong$		bedeutet Anpassung im Sinne des Streß-Kriteriums.

MONANOVA paßt also nicht die aus den ermittelten Teilnutzenwerten vorhergesagten Gesamtnutzenwerte $\hat{y}_j$ direkt den empirisch ermittelten Werten p_j an, sondern deren monotoner Transformation z_j. Diese Anpassung zwischen z_j und $\hat{y}_j$ geschieht wechselseitig und iterativ, solange, bis ein Zielkriterium — minimaler Streß — erreicht ist. Das Streßmaß hat die folgende Form (nach Backhaus 1990, S. 357 — vgl. aber auch Kruskal 1965, S.252):

$$S = \sqrt{\frac{\sum_{j=1}^{n}(z_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2}} \tag{1.1.4}$$

$\bar{y}$ meint hier natürlich den Mittelwert der $\hat{y}_j$. Die Anpassung der z_j-Werte an die $\hat{y}_j$-Werte erfolgt durch eine sogenannte monotone Regression: Gesucht wird eine monotonen Transformation f_M, die die folgende Monotoniebedingung erfüllt: $p_j < p_{j'} \Rightarrow z_j \leq z_{j'}$ (vgl. Kruskal 1965, S. 252)[15]. Ausgangspunkt der monotonen Regression sind die $\hat{y}_j$-Werte, diese werden mit den p_j-Werten daraufhin verglichen, ob sie die obige Monotoniebedingung erfüllen. Erfüllt ein $\hat{y}_j$-Wert diese Bedingung in allen Paarvergleichen, dann wird $z_j = \hat{y}_j$ gesetzt. Verletzen 2 oder mehrere $\hat{y}_j$-Werte die Monotoniebedingung, dann wird deren Mittelwert gebildet und dieser den entsprechenden z_j-Werten zugeordnet.

Bei dem Verfahren, durch das die Anpassung der $\hat{y}_j$-Werte an die z_j-Werte erreicht wird, handelt es sich um ein sogenanntes Gradientenverfahren: Es wird

[15]Hinsichtlich der Behandlung gleicher Rangwerte — also $p_j = p_{j'}$ — gibt es dabei zwei Möglichkeiten: Nämlich Beibehaltung der obigen Einschränkung für die z_j-Werte oder $z_j = z_{j'}$ (vgl. ebd., S.253).

der Vektor der partiellen ersten Ableitungen (Gradientenvektor) von Gleichung (1.1.4) nach **b** gebildet (**b** ist natürlich der Vektor der geschätzten β_{lk}-Werte aus Gleichung (1.1.1)) und nach der „Methode des steilsten Abstiegs“ („method of steepest descent“, vgl. Kruskal 1964b, S. 117f) nach einem Minimum gesucht: Aus dem Gradientenvektor für ein gegebenes **b** sind die Richtungen abzulesen, nach welchen die einzelnen b_{lk} verändert werden müssen, damit sich der Streßwert einem Minimum nähert. Verändert man die b_{lk} um einen kleinen Betrag in diese Richtungen, so resultiert ein niedrigerer Streßwert. Dies kann man solange wiederholen, bis ein Minimum erreicht ist, d.h. bis der Gradientenvektor zum Nullvektor geworden ist (eine weitere Bedingung für ein Minimum ist nach der Methode des steilsten Abstiegs nicht nötig). Tatsächlich werden die Iterationen abgebrochen, wenn der Rückgang des Streßwertes, der durch eine Iteration erzielt wird, einen festgelegten Minimalwert unterschreitet. Ein Problem dieses Gradientenverfahrens besteht darin, daß es nicht garantieren kann, daß das globale Minimum erreicht wurde, es kann sich bei dem aufgefundenen Minimum um ein lokales handeln.

MONANOVA läuft also wie folgt ab: Es startet mit einer ANOVA-Lösung für **b** und berechnet auf Basis dieser ersten Lösung $\hat{y}_j$-Werte. Sodann erfolgt die monotone Regression, es werden also an die $\hat{y}_j$-Werte angepaßte z_j-Werte gebildet. Dann werden für die bisher ermittelten Werte Streß und der Gradientenvektor ermittelt. Ist noch kein Minimum erreicht, dann werden nun, nach der Methode des steilsten Abstiegs, neue Lösungen für **b** ermittelt und der soeben beschriebene Ablauf wird erneut durchlaufen — die Prozedur wiederholt sich so lange, bis ein Minimum erreicht ist. Über Details informieren v.a. Kruskal (1964b) und Kruskal (1965)[16].

(3) Ein weiteres nonmetrisches Verfahren, das in den achtziger Jahren größeres theoretisches Interesse — gemessen an der Anzahl wissenschaftlicher Veröffentlichungen, die sich damit beschäftigen — hervorgerufen hat, wenngleich es auch bisher noch seltener praktische Anwendung zu finden scheint (vgl. wieder Wittink & Cattin 1989) ist **LINMAP**. Der Titel entstand aus der Abkürzung von „*LIN*ear programming techniques for *M*ultidimensional *A*nalysis of *P*references“, es wurde von V. Srinivasan und Allan D. Shocker entwickelt (Srinivasan & Shocker 1973a; 1973b). LINMAP wird im Abschnitt 1.2.1 ausführlich behandelt werden. Das gesteigerte wissenschaftliche Interesse an der Methode dürfte nicht unwesentlich mit einer vergleichenden Simulationsstudie von Wittink & Cattin (1981) zusammenhängen, in welcher LINMAP unter bestimmten Bedingungen die besten Resultate unter den getesteten Methoden erzielte, während in allen anderen Bedingungen sich die metrische Methode (ANOVA) als überlegen erwies (und auch MONANOVA in keiner Bedingung besonders günstig abschnitt) — auch diese Studie wird

[16]Insbesondere zur Frage der „Schrittgröße“ beim Übergang von einer Lösung von **b** zur nächsten vgl. Kruskal (1965, S. 261f).

später (Abschnitt 2.1.1) noch genauer dargestellt werden.

(4) Interessant ist auch das nonmetrische **Verfahren nach Johnson** (1975), da es auch zur Analyse von Trade-Off-Matrizen entwickelt wurde. Für dieses Verfahren hat sich keine spezielle Bezeichnung eingebürgert. Es ist insofern MONANOVA sehr ähnlich, als es sich ebenfalls um ein iteratives Gradientenverfahren handelt, ist andererseits aber einfacher, da es ohne die monotone Transformation f_M (vgl. Gleichung (1.1.3)) auskommt und so wie die metrischen Verfahren direkt die geschätzten Werte $\hat{y}_j$ an die empirischen Rangwerte p_j anpaßt. Dazu definiert Johnson θ — als Maß für den „lack of fit“:

$$\theta^2 = \frac{\sum_{j,j'} \delta_{jj'}(\hat{y}_j - \hat{y}_{j'})^2}{\sum_{j,j'}(\hat{y}_j - \hat{y}_{j'})^2} \quad (\text{für } j \neq j') \tag{1.1.5}$$

wobei:

$$\delta_{jj'} \quad = \quad \begin{cases} 1 & \text{wenn Vorzeichen } (\hat{y}_j - \hat{y}_{j'}) \neq \text{Vorzeichen } (p_j - p_{j'}) \\ 0 & \text{sonst} \end{cases}$$

Gleichung (1.1.5) wurde von Johnson (1975, S. 164) übernommen und der Schreibweise von Gleichung (1.1.3) angepaßt. Der Zähler von θ^2 enthält die Summe der quadrierten Differenzen zwischen allen Paaren vorhergesagter Gesamtnutzenwerte, die sich „in der falschen Rangfolge“ — verglichen mit den empirisch ermittelten Rangwerten — befinden. Über die Bedeutung von θ^2 schreibt Johnson: *„It can be shown that θ^2 has a natural interpretation as the proportion of the variation among the $\hat{y}$'s which is 'inconsistent' with the y's“* (ebd., S. 165). Wie schon bei MONANOVA der Streß S (Gleichung (1.1.4)), so wird hier θ nach **b** abgeleitet und nach der Methode des steilsten Abstiegs ein Minimum gesucht.

Johnson hat — wie schon gesagt — sein Verfahren auch zur Analyse von Trade-Off-Matrizen entwickelt (Johnson 1974). Das Prinzip dieser Analyse ist einfach: Die Paarvergleiche, die zur Berechnung des „lack of fit“ herangezogen werden, werden nun eben nicht mehr zwischen den vorhergesagten Gesamtnutzenwerten $\hat{y}_j$ der vollständigen Stimulusprofile gezogen, sondern zwischen den Rangwerten der Zellen der Trade-Off-Matrizen, wie sie aus einer gegebenen Lösung **b** vorhergesagt werden. Nimmt man z.B. eine 3^4-Struktur, so erhält man 6 3×3-Matrizen, in denen jeweils 2 der 4 Attribute gegeneinander abgewogen werden. Jede dieser Matrizen enthält 9 Zellen, die von der Versuchsperson in eine Rangfolge gebracht werden, und es ergeben sich für jede Trade-Off-Matrix 36 Paarvergleiche der Zellen. Aus einer gegebenen Lösung **b** lassen sich nun auch die Rangordnungen in den einzelnen Trade-Off-Matrizen vorhersagen (unter Voraussetzung der Gültigkeit des Unabhängigkeitsaxioms): Genauso, wie in Gleichung (1.1.5), geht ein Paarvergleich (zwischen den Zellen einer Trade-Off-Matrix) dann in den Zähler des Lack-of-fit-Wertes ein, wenn die vorhergesagte Rangfolge die empirisch ermittelte umkehrt. Es werden die

Paarvergleiche aller Trade-Off-Matrizen kumuliert, im Beispiel der 3^4-Struktur gingen also $6 \times 36 = 216$ Paarvergleiche in den Nenner des Lack-of-fit-Wertes ein. Johnson definiert diesen Wert für das Verfahren zur Analyse von Trade-Off-Matrizen allerdings abweichend von Gleichung (1.1.5) und nennt ihn hier ϕ: Wenn r das Verhältnis der Rangwerte jeweils zweier Zellen jeweils einer Trade-Off-Matrix ist, so gilt (Johnson 1974, S. 124):

$$\phi = \frac{\sum(r + (1/r) - 2)\delta}{\sum(r + (1/r) - 2)}$$

wobei δ wieder dieselbe Bedeutung hat, wie $\delta_{jj'}$ in Gleichung (1.1.5)[17]. Die Summierung — wie gesagt — erfolgt über alle Paarvergleiche aller Trade-Off-Matrizen.

Die Auflistung der Schätzverfahren zur CA ist mit den obigen Punkten **(1)** bis **(4)** nicht vollständig, soll aber hier nicht weitergeführt werden, um nicht allzu sehr das Ziel des vorliegenden Kapitels — Einführung in bzw. Übersicht über die theoretischen Grundlagen der im Anschluß daran dargestellten Untersuchung — aus den Augen zu verlieren. Die Auswahl der Verfahren erfolgte nach dem Eindruck relativer praktischer oder theoretischer Bedeutsamkeit, den ich aufgrund der durchgesehenen Literatur gewonnen habe. Über weitere Verfahren informieren Green & Srinivasan (1978, S. 112f), insbesondere sei noch auf eine dort aufgeführte Gruppe von Prozeduren hingewiesen, die sich von den metrischen und nonmetrischen Verfahren grundsätzlich unterscheiden — in den Worten der Autoren: *„Methods which relate paired-comparison data to a choice probability model“* (ebd., S. 113). Auch an die im Abschnitt 1.1.1 (S. 9) erwähnte nonmetrische Prozedur UNICON (Roskam 1974) sei noch einmal erinnert.

Es kann jetzt, mit Blick auf den Titel der vorliegenden Untersuchung, eine *Klärung der Begriffe* erfolgen: Eine **metrische CA** liegt dann vor, wenn die Daten mit einer metrischen Methode analysiert werden, dagegen wird als **nichtmetrische CA** eine Analyse mittels eines der nonmetrischen Verfahren bezeichnet. Die Unterscheidung richtet sich also danach, welcher Gebrauch von den Präferenzdaten gemacht wird: In der metrischen CA wird Intervallskalierung der abhängigen Variablen vorausgesetzt bzw. werden — genauer ausgedrückt — die abhängigen Variablen so behandelt, als seien sie intervallskaliert! Es hat sich nämlich in praktischen Anwendungen der CA mehr und mehr eingebürgert, die metrischen Verfahren auch bei Rangwerten über multiattributive Stimuli anzuwenden. Dabei wird dann Gebrauch von metrischen Qualitäten der Rangwerte gemacht, die diese vielleicht gar nicht besitzen

[17] Die Notwendigkeit dieser von θ abweichenden Definition von ϕ fand ich in den mir zugänglichen Arbeiten von Johnson allerdings nicht erläutert. Vielleicht befindet sich diese Erläuterung bei Johnson (1973b), ein Text, der bei Green & Srinivasan (1978) zitiert wird, der mir aber nicht zur Verfügung stand.

(d.h. die numerischen Abstände zwischen den Rangwerten werden als sinnvolle Größen interpretiert, in denen sich das tatsächliche Ausmaß des „subjektiven Abstands“ zweier Stimuli in der Präferenz des Individuums widerspiegelt) — ein Gebrauch, der theoretisch nicht gerechtfertigt ist. Diese Rechtfertigung kann z.B. mit Simulationsstudien versucht werden, in denen untersucht wird, „wie gut“ metrische Verfahren auf der Basis ordinaler Daten im Vergleich zu nonmetrischen Verfahren schätzen. Die in den nachfolgenden Teilen vorgestellte Untersuchung ist eine solche Simulationsstudie, sie wurde u.a. mit dem Ziel entworfen, die Ergebnisse der OLS-Regression bei ordinalen Rangwerten als abhängiger Variable mit denen, die durch LINMAP erzielt werden, zu vergleichen. Zu Beginn von Teil 2 werden die Simulationsstudien vorgestellt werden, die der eigenen als Vorbild dienten (Abschnitt 2.1.1).

Bleibt noch der letzte Punkt der Abbildung 1.2, die Aggregation der Nutzenwerte. CA bzw. CM überhaupt ist zunächst immer Individualanalyse, d.h. Analyse individueller Präferenzordnungen jeweils einer Person. In den meisten praktischen Anwendungen besteht aber Interesse an allgemeineren Aussagen über Gruppen von Individuen. Z.B. könnte danach gefragt werden, ob verschiedene soziale Gruppen auch verschiedene „typische Präferenzstrukturen“ aufweisen (d.h. ob es gruppentypische Gewichtungen der Attribute und deren Stufen gibt), oder es könnte umgekehrt ein Interesse bestehen, innerhalb einer Gesamtheit von Individuen Gruppen ähnlich präferierender Personen zu identifizieren. Es müssen also die Ergebnisse individueller Analysen aggregiert werden und dieses wird erreicht unter Rückgriff auf bekannte und übliche statistische Berechnungen und Prozeduren. So besteht die einfachste Möglichkeit einer Aggregation darin, über eine Stichprobe getesteter Personen hinweg die Mittelwerte und Standardabweichungen der ermittelten Teilnutzenwerte zu bilden[18]. Green & Srinivasan (1978, S. 117) verweisen auf die Bedeutung von Clusteranalyse und Diskriminanzanalyse für die Aggregation individueller Nutzenwerte: Die Clusteranalyse erlaubt die Identifizierung typischer Personengruppen, die hinsichtlich der relevanten Attribute der in Frage stehenden multiattributiven Stimuli ähnliche Präferenzen besitzen, in der Diskriminanzanalyse können die Teilnutzenwerte als Prädiktorvariablen und die Zugehörigkeit zu bestimmten, vorher definierten Gruppen als Kriterium verwendet werden, mit dem Ziel, die Zugehörigkeit zu solchen a priori definierten Gruppen aus den individuellen Präferenzstrukturen vorhersagen zu können. Ein jeder kann sich wohl mit etwas Phantasie weitere Möglichkeiten, die Ergebnisse mehrerer individueller Analysen zu aggregieren, vorstellen — der Rückgriff auf statistische Verfahren hängt im Einzelfall der jeweiligen Anwendung von der Fragestellung ab. Damit aber wird deutlich, daß die Methoden der Aggregation nach der Individualanalyse im Grunde nicht zum spezifischen Themenbereich

[18] Zuvor müssen die Teilnutzenwerte jeder Person normiert werden. Backhaus (1990, S. 362) liefert dazu eine Formel.

der CA zählen, weshalb sie auch hier nicht weiter erläutert werden.

Letzteres gilt nicht für eine Art der Aggregation, die sich bei Anwendung metrischer Verfahren anbietet (analoge Vorgehensweisen sind auch für die nonmetrischen Verfahren formulierbar): Man kann die Aggregation sozusagen in die Schätzung der Nutzenwerte hineinziehen. Anstatt bei m Versuchspersonen m Varianzanalysen oder m Regressionen zu rechnen und danach die Mittelwerte der Teilnutzenwerte zu bilden, kann eine einzige Varianzanalyse bzw. Regression gerechnet werden, in die die Rangdaten aller m Personen eingehen. Für die ANOVA hieße das, daß nun nicht mehr ein Fall pro Zelle vorliegt, sondern m Fälle, und für die OLS-Regression müßte Gleichung (1.1.2) so erweitert werden:

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \dots & \mathbf{X}_t \\ \mathbf{X}_1 & \mathbf{X}_2 & \dots & \mathbf{X}_t \\ \vdots & & & \\ \mathbf{X}_1 & \mathbf{X}_2 & \dots & \mathbf{X}_t \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \vdots \\ \boldsymbol{\beta_t} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \\ \vdots \\ \boldsymbol{\epsilon_m} \end{pmatrix} \tag{1.1.6}$$

wobei:

$\mathbf{y}_i$ = der $n\times1$-Vektor der (metrischen) Gesamtnutzenwerte der Versuchsperson i ($i = 1, 2, \dots, m$).

$\boldsymbol{\epsilon_i}$ = $n\times1$-Vektor der Fehler der einzelnen Beobachtungen j ($j = 1, 2, \dots, n$) für Versuchsperson i.

Es ist gerade bei „einfachen“ Aggregationen, bei denen die einzelnen Präferenzstrukturen der Versuchspersonen einfach und ohne Differenzierung zu Stichprobenkennwerten aggregiert werden — sei es durch nachträgliche Mittelwertsbildung, oder nach der durch Gleichung (1.1.6) gegebenen Methode — Vorsicht geboten. Das Problem einer solchen Aggregation sind heterogene Präferenzstrukturen in der Stichprobe: Größere Varianzen der in individuellen Analysen ermittelten Teilnutzenwerte (über die Versuchspersonen hinweg) erhöhen den Standardfehler der Mittelwerte (vgl. Bortz 1979, S. 116: Gl. (3.3)), so wie auch bei einer OLS-Regression nach Gleichung (1.1.6) umso größere Standardfehler der geschätzten Koeffizienten zu erwarten sind, je heterogener die Stichprobe ist. Allerdings setzt die Darlegung des letzteren Zusammenhangs bereits einige Kenntnisse der OLS-Regression voraus, wie sie in dieser Arbeit erst an späterer Stelle in Abschnitt 1.2.2 vermittelt werden — sie sei trotzdem hier und möglichst knapp versucht, denn die einfache Aggregation über eine Stichprobe ist sicherlich ein häufiges Anliegen konkreter Anwendungen und auf den ersten Blick bietet sich gerade die Aggregation nach Gleichung (1.1.6) an, erspart diese doch den Aufwand vieler einzelner Analysen.

Ist also die Stichprobe sehr heterogen, so bedeutet das, daß in den verschiedenen Vektoren $\mathbf{y}_i$ in Gleichung (1.1.6) sehr unterschiedliche Abfolgen von

Rangwerten enthalten sind (der Einfachheit wegen sei angenommen, daß die empirisch erhobenen Werte als ganzzahlige Werte $1, 2, \ldots, n$ vorliegen). Der Anschaulichkeit halber soll dieses an einem einfachen fiktiven Beispiel von 2 sehr heterogenen Versuchspersonen und einer 2^2-Struktur (jeweils nur eine Dummy-Variable pro Attribut) gezeigt werden:

	$\mathbf{y}$	$\mathbf{X}$	
Vp1	1	1	1
	2	1	0
	3	0	1
	4	0	0
Vp2	4	1	1
	3	1	0
	2	0	1
	1	0	0

Die Figur zeigt den Vektor $\mathbf{y}$ und die Matrix $\mathbf{X}$ (dem Vektor $\mathbf{y}$ des OLS-Modells entspricht im Fall der Gleichung (1.1.6) der aus den einzelnen Vektoren $\mathbf{y}_i$ gebildete Supervektor, der $\mathbf{X}$-Matrix die aus den einzelnen Matrizen $\mathbf{X}_l$ gebildete Supermatrix). $\mathbf{X}$ enthält soviele unterschiedliche Zeilen, wie Stimuli bewertet wurden, pro Versuchsperson wiederholt sich die Abfolge dieser unterschiedlichen Zeilen, d.h. die Matrix enthält die unterschiedlichen Zeilen mehrmals. Nun gilt unter der Annahme der Voraussetzungen des OLS-Models $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ für die Verteilung der y-Werte (vgl. Werner 1993, S. 76): $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. D.h. Die Varianz der y-Werte bei einer bestimmten, fixen Konstellation der Prädiktorvariablen entspricht der Residualvarianz σ^2. Wie obiges Beispiel zeigt, bedingt die Heterogenität der Versuchspersonen, daß die y-Werte, die mit einer bestimmten Konstellation der x-Werte (mit einer bestimmten, mehrmals wiederholten Zeile von $\mathbf{X}$) verknüpft sind, manchmal stark variieren (vgl. z.B. die Zeilen 1 und 5 im obigen Beispiel). Diese Variation kann durch die Prädiktoren des Modells nicht „erklärt“ werden, sie ist auf die Residualvarianz zurückzuführen (bzw. sie erhöht die Residualvarianz). Mit der Größe der Residualvarianz aber wachsen auch die Standardfehler der geschätzten Koeffizienten $\mathbf{b}$, für deren Verteilung gilt (Werner 1993, S. 98: Gleichung (3.13.4)): $\mathbf{b} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

Heterogene Stichproben können also dazu führen, daß daraus aggregierte Nutzenwerte mit hohen Standardfehlern behaftet sind. Ein Ausweg liegt hier natürlich nahe: *„Bei starker Heterogenität lassen sich durch Anwendung einer Clusteranalyse (...) homogene(re) Teilgruppen bilden. Die Clusterung kann auf Basis der empirischen Rangdaten wie auch auf Basis der durch Einzelanalysen gewonnenen Teilnutzenwerte vorgenommen werden“* (Backhaus 1990, S. 363). Hagerty (1885, S. 169) gibt einen kurzen Überblick über einige Studien, in denen 3 verschiedene „Levels“ der Aggregation verglichen wurden (hinsichtlich des Prozentsatzes korrekt vorhergesagter Entscheidungen der Mitglieder

der jeweiligen Stichprobe). Es ergab sich stets dieselbe, wenig verblüffende, Rangordnung: Die wenigsten korrekten Vorhersagen wurden durch Aggregation über die Gesamtstichprobe erzielt, bessere Ergebnisse wurden bei Clusterung der Stichprobe (und Berechnung von Nutzenwerten für die Cluster) erzielt, die besten, wenn gar nicht aggregiert wurde und für jeden Respondenten die Vorhersage auf der Basis seiner individuellen Nutzenwerte gewonnen wurde. Es existieren aber 2 Ansätze, die darauf abzielen, vermittles einer Aggregation eine Verbesserung der Vorhersage gegenüber den unaggregierten individuellen Analysen zu erreichen.

Hagerty (1985) benutzt die Faktorenanalyse, um die Daten der Versuchspersonen zu kombinieren und dadurch die Effizienz der Schätzung zu erhöhen: „*The rationale we use is that when respondents are similar, their responses should be averaged in some way to obtain more reliable responses*“ (ebd., S. 168). Er zielt also auf den Fehler in den Antworten der Respondenten ab. Hätten z.B. 2 Personen exakt diesselben „wahren“ Präferenzen, so würde die Schätzung verbessert, wenn anstelle ihrer empirisch ermittelten Rangdaten bei beiden die Mittelwerte aus ihren Antworten gesetzt würden. Auf diese Weise würde dann im Gesamtmodell (s.u.) die Zahl der zu schätzenden Teilnutzenwerte reduziert (bei gleichbleibender Zahl von Beobachtungen) und diesselbe Überlegung, die oben zur Erläuterung des größeren Standardfehlers der OLS-Koeffizienten bei heterogenen Versuchspersonen herangezogen wurde, führt hier zur Erwartung eines kleineren Standardfehles der Parameterschätzung. Die Darstellung von Hagerty's komplizierten Berechnungen würde den Rahmen dieses Kapitels sprengen, es soll deshalb nur ein grober Blick auf sein Modell geworfen werden. In Anlehnung an Gleichung (1.1.2) und Gleichung (1.1.6) kann das Gesamtmodell, von dem Hagerty ausgeht, so formuliert werden:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U} \tag{1.1.7}$$

wobei:

$\mathbf{Y}$ = $(\mathbf{y}_1|\mathbf{y}_2|\ldots|\mathbf{y}_m)$.

$\mathbf{X}$ = $(\mathbf{X}_1|\mathbf{X}_2|\ldots|\mathbf{X}_t)$.

$\mathbf{B}$ = $\begin{pmatrix} \mathbf{b}_{11} & \mathbf{b}_{12} & \ldots & \mathbf{b}_{1m} \\ \mathbf{b}_{21} & \mathbf{b}_{22} & \ldots & \mathbf{b}_{2m} \\ \vdots & & & \\ \mathbf{b}_{t1} & \mathbf{b}_{t2} & \ldots & \mathbf{b}_{tm} \end{pmatrix}$
(d.h. die Matrix der Schätzungen von $(\boldsymbol{\beta_1}|\boldsymbol{\beta_2}|\ldots|\boldsymbol{\beta_t})$ für jede Versuchsperson i).

$\mathbf{U}$ = die Matrix der Residuen für jede Person (spaltenweise).

Gesucht wird nach einer Gewichtungsmatrix $\mathbf{A}$, die die empirisch erhobenen Rangwerte $\mathbf{Y}$ transformiert: $\mathbf{Z} = \mathbf{YA}$. Mit den transformierten Daten wird dann die eigentliche Analyse der Nutzenwerte durchgeführt. $\mathbf{A}$ ist durch eine

Matrix $\mathbf{S}$ definiert: $\mathbf{A} = \mathbf{S}(\mathbf{S}'\mathbf{S})\mathbf{S}'$. Die Spalten der Matrix $\mathbf{S}$ entsprechen den Clustern oder Faktoren, die über die Personen extrahiert werden, die Zeilen entsprechen den Personen. Hagerty veranschaulicht ihre Bedeutung am Fall nicht-überlappender Cluster: Wenn z.B. 3 Personen auf 2 Cluster aufgeteilt würden — die ersten beiden auf das erste, die dritte auf das zweite — dann resultiert

$$\mathbf{S} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Jede Person hat also ihre Zeile und dort in der Spalte eine 1, die für das Cluster steht, in das die Person fällt. Aus diesem Gewichtungsschema resultiert eine $\mathbf{Z}$-Matrix, in der nun anstelle der individuellen Antworten (Rangwerte) jeder Person die Mittelwerte dieser Antworten über die Mitglieder des Clusters stehen — im obigen Beispiel:

$$\mathbf{Z} = (\mathbf{y}_1|\mathbf{y}_2|\mathbf{y}_3)\mathbf{A} = \left(\frac{\mathbf{y}_1 + \mathbf{y}_2}{2}|\frac{\mathbf{y}_1 + \mathbf{y}_2}{2}|\mathbf{y}_3\right)$$

Liefert die Faktorenanalyse das Gewichtungsschema, dann stehen in den Zeilen von $\mathbf{S}$ die Ladungen der jeweiligen Versuchsperson auf den extrahierten Faktoren. Hagerty zeigt, daß die optimale Transformation aus der Hauptkomponentenanalyse der Korrelationsmatrix zwischen den Versuchspersonen resultiert. Der Erwartungswert des mittleren quadrierten Fehlers über das gesamte Modell nämlich wird durch diese Transformation minimiert. Unter den Möglichkeiten zur Bildung von $\mathbf{S}$, die in dieser Hinsicht schlechter abschneiden, gehört auch $\mathbf{S} = \mathbf{I}$, also die Möglichkeit gar keiner Transformation (und Aggregation)! Auf den eigentlichen „Trick“, mit dem dies erreicht wird, sei noch einmal mit Hagerty's eigenen Worten hingewiesen: *„The optimal weighting method may also be considered to reduce the total number of parameters estimated. It does so ... by reducing the number of independent respondents estimated. It reduces subject space from n dimensions (the number of independent responses) to r dimensions (the number of factors kept)“* (Hagerty 1985, S. 181).

Einen ähnlichen Ansatz verfolgt Kamakura (1988), allerdings über die Clusterung ähnlicher Versuchspersonen. Sowohl Hagerty, als auch Kamakura liefern in den genannten Studien Ergebnisse von Untersuchungen mit sowohl synthetischen, als auch empirischen Daten, die jeweils den Wert ihrer Methoden bei der Vorhersage von Präferenzentscheidungen belegen. Green & Helsen (1989) beziehen beide Ansätze in eine Vergleichsstudie mit ein: In keiner der dort getesteten Bedingungen kann eine der beiden Methoden die individuelle metrische CA (OLS-Regression) in der Vorhersage der Präferenzen für einen Satz von Validierungs-Stimulusprofilen wesentlich übertreffen. Green & Srinivasan (1990) kommen zu dem Schluß: *„Overall, it appears that conventional,*

individual-level-based conjoint analysis may be difficult to improve in a major way (at least when the number of stimulus evaluations is large in relation to the number of parameters being estimated)“ (ebd., S. 8).

1.2 Analyseverfahren

In diesem Kapitel sollen die beiden Analysemethoden genauer erläutert werden, deren prädiktive Qualtitäten in der im Anschluß an dieses Kapitel dargestellten Simulationsstudie überprüft werden: Die *OLS-Regression* als metrisches Verfahren einerseits und *LINMAP* als nonmetrisches andererseits. Dabei soll v.a. der Darlegung von letzterem größerer Raum eingeräumt werden, da LINMAP ein außerhalb der Gemeinde der CA-Experten noch relativ unbekanntes und selten dokumentiertes Verfahren ist, das sozusagen noch nicht zu den Standards sozialwissenschaftlicher Methodenlehre gehört — ganz im Gegensatz zur OLS-Regression, die dementsprechend etwas knapper behandelt werden wird.

1.2.1 LINMAP

LINMAP steht — wie schon in Abschnitt 1.1.2 erwähnt — für „*LIN*ear programming techniques for *M*ultidimensional *A*nalysis of *P*references“, es wurde von seinen Erfindern V. Srinivasan und Allan D. Shocker in 2 vielzitierten Arbeiten vorgestellt, die scheinbar die einzigen Quellen der Dokumentation des Verfahrens bilden (Srinivasan & Shocker 1973a; 1973b). Diese Darlegungen sind nicht ganz unkompliziert, so daß der Anspruch, bei meinen eigenen Erörterungen hier besondere Originalität walten zu lassen, wohl etwas vermessen wäre und höchstwahrscheinlich nur Verwirrung stiften würde (und wozu sollte solche Originalität überhaupt gut sein?): Ich werde mich deshalb im folgenden eng an Srinivasan & Shocker (1973a) anlehnen (alle in diesem Abschnitt mit „ebd.“ gekennzeichneten Referenzen beziehen sich auf diesen Artikel), insbesondere bei der Terminologie der formalen Ausdrücke nehme ich nur leichte Veränderungen vor, um eine gewisse Einheitlichkeit der Bezeichnungen über die verschiedenen Kapitel hinweg zu wahren. In dieser Arbeit ist LINMAP in seiner allgemeinsten Form und am detailliertesten dokumentiert, während in Srinivasan & Shocker (1973b) sozusagen eine Unterversion der Prozedur vorgestellt wird, welche die für die nonmetrische CA relevante ist — nichtsdestotrotz ist der Rückgriff auf die grundlegendere Darstellung im ersteren Artikel wohl notwendig zum Verständnis des Verfahrens.

In seiner allgemeinsten Form ist LINMAP ein Verfahren zur Parameterschätzung auf der Basis von Paarvergleichen multiattributiver Stimuli und

unter Zugrundlegung eines sogenannten „ideal-point models“. Ein *Idealpunktmodell* beruht auf der Vorstellung, daß eine Person auf den Attributen der Stimuli Punkte optimaler Präferenz besitzt und daß ein multiattributiver Stimulus umso mehr päferiert wird, je „näher“ die Ausprägungen auf den Attributen an diesen Idealpunkten liegen. Dieses Modell setzt die Existenz einer kontinuierlichen Skala für die Ausprägungen der Attribute voraus — was auf den ersten Blick der obigen Behauptung der Allgemeinheit gerade dieses Modells zu widersprechen scheint, benötigt dagegen doch das Modell der CA lediglich nominale Ausprägungskategorien auf den Attributen. Es wird aber gezeigt werden, wie das CA-Modell formuliert werden kann, so daß es mit dem LINMAP-Algorithmus kompatibel ist. LINMAP setzt nicht voraus, daß die Ausprägungen der Attribute vor Durchführung der Analyse tatsächlich auf einer kontinuierlichen Skala gemessen werden können: Srinivasan & Shocker unterscheiden zwischen externaler und internaler Analyse — der Unterschied besteht darin, daß bei ersterer die Ausprägungen der Attribute a priori festgelegt sind, während bei letzterer auch die Positionen der Ausprägungen auf der kontinuierlichen Skala (zusammen mit den übrigen Parametern) geschätzt werden. Die Darstellung des „Kerns“ der Methode erfolgt für die externale Analyse.

Zunächst sei J die Menge der multiattributiven Stimuli j ($j = 1, 2, \ldots, n$), über die die Präferenzurteile gefällt wurden, und L die Menge der Attribute l ($l = 1, 2, \ldots, t$). Ferner sei $X_j = \{x_{jl}\}, l \in L$, die Menge der Ausprägungen des Stimulus j auf den Attributen l — oder, anders ausgedrückt, die Lage des Stimulus j im t-dimensionalen (Attribut-)Raum.

Nun ist der *Idealpunkt* einer Person in diesem Raum gegeben durch die Menge $O = \{o_l\}$ der Idealpunkte optimaler Präferenz o_l auf den Skalen der Ausprägungen der Attribute L [19]. Die o_l können positiv, negativ oder gleich Null sein. Im t-dimensionalen Raum ist die *ungewichtete Euklidische Distanz* d_j^u des Stimulus j vom Idealpunkt:

$$d_j^u = \left(\sum_{l=1}^{t} (x_{jl} - o_l)^2 \right)^{1/2}$$

Nach dem klassischen Modell des multidimensionalen „Unfolding“, das die Autoren zitieren (ebd., S. 341; siehe auch Roberts 1979, S. 240), wird ein Stimulus j genau dann gegenüber einem Stimulus r präferiert, wenn gilt $d_j^u \leq d_r^u$. Es ist aber ratsam, die unterschiedlichen Attribute/Dimensionen unterschied-

[19] Srinivasan & Shocker benennen die Ausprägungen mit y und die Idealpunkte mit x. Diese Terminologie wurde verändert im Hinblick auf die bei der Darstellung solcher Parameterschätzverfahern übliche Benennung, bei der mit x im allgemeinen Prädiktoren, mit y Kriteriumsvariablen benannt werden.

lich zu gewichten, denn zum einen sind normalerweise (in konkreten Anwendungsproblemen) die verschiedenen Attribute in unterschiedlichen Einheiten skaliert, zum anderen muß angenommen werden, daß für ein Individuum die unterschiedlichen Dimensionen auch unterschiedliche Wichtigkeit besitzen. Srinivasan & Shocker führen also Gewichte für die Attribute ein, in die beides — unterschiedliche Skalen und unterschiedliche subjektive Wichtigkeit für das Präferenzurteil — miteingeht: $W = \{w_l\}$ ist die Menge der Gewichte w_l der Dimensionen aus L. Es ergibt sich die *gewichtete Euklidische Distanz*

$$d_j = \left(\sum_{l=1}^{t} w_l (x_{jl} - o_l)^2 \right)^{1/2} \tag{1.2.1}$$

und schließlich als *quadrierte Distanz* s_j des Stimulus j vom Idealpunkt:

$$s_j = \sum_{l=1}^{t} w_l (x_{jl} - o_l)^2 \tag{1.2.2}$$

Damit nach dem Modell der gewichteten Distanzen die Präferenz für einen Stimulus mit der Nähe zum Idealpunkt wächst, bedarf es noch der Einschränkung, daß die Gewichte nicht negativ werden können:

$$w_l \geq 0 \quad \text{für alle} \quad l \in L \tag{1.2.3}$$

Es sei nun $\Omega = \{(j, r)\}$ die *Menge der geordneten Stimuluspaare* (j, r): Hierbei ist j der im Paarvergleich gegenüber r präferierte Stimulus (einstweilen seien forcierte Entscheidungen vorausgesetzt, also keine Indifferenzen zwischen Stimuli). Die Menge Ω stellt die empirische Datenbasis des Verfahrens dar, es ist dazu nicht notwendig, daß sie alle möglichen Paarvergleiche enthält — jedoch: „*However, the estimated parameters w_p and x_p* [das sind w_l und o_l in meiner Notation] *(2t in number) are more reliable, the larger the number of pairs in Ω compared to $2t$*“ (ebd., S. 342). Bei einer gegebenen Lösung für die Parameter w_l und o_l müßte bei völliger Konsistenz des Modells der gewichteten Distanzen mit den empirisch ermittelten Präferenzdaten für jedes Stimuluspaar (j, r) aus Ω gelten:

$$s_r \geq s_j \tag{1.2.4}$$

Damit ist Richtung und *Ziel der LINMAP-Analyse* klar: Die zu schätzenden Parameter des Modells sind die Gewichte w_l und die Idealpunkte o_l, für sie ist — gegeben die Menge X_j der Ausprägungen der Stimuli in J auf den Attributen in L und die Menge geordneter Stimuluspaare Ω — eine Lösung

(W, O) zu finden, die Gleichung (1.2.3) erfüllt und durch die Gleichung (1.2.4) möglichst selten verletzt wird.

Zur Präzisierung dieses Ziels wird der „poorness of fit“ einer Lösung (W, O) definiert: Zunächst einmal ist

$$(s_r - s_j)^- = \max\{0, (s_j - s_r)\}$$

d.h. dieser Wert wird dann Null, wenn durch eine Lösung (W, O) das Modell mit einem Paar (j, r) aus Ω konsistent ist, andernfalls wird er zur Differenz der quadrierten Distanzen des Stimuluspaares. Es folgt für den *Poorness-of-Fit* B:

$$B = \sum_{(j,r)\in\Omega} (s_r - s_j)^- \tag{1.2.5}$$

Umgekehrt läßt sich auch der *Goodness-of-Fit* G einer bestimmten Lösung definieren:

$$(s_r - s_j)^+ = \max\{0, (s_r - s_j)\}$$

und

$$G = \sum_{(j,r)\in\Omega} (s_r - s_j)^+ \tag{1.2.6}$$

Nun existieren 2 *triviale Lösungen*, die es auszuschließen gilt

(1) Wenn alle $w_l = 0$ dann wird auch $B = 0$.

(2) Transformiert man eine Lösung (W, O) in (W', O) durch $w'_l = \lambda w_l, \lambda > 0$, dann folgt (vgl. Gleichung (1.2.2)): $s'_j = \lambda s_j \Rightarrow (s'_r - s'_j)^- = \lambda(s_r - s_j)^- \Rightarrow B' = \lambda B$. Man kann also jede beliebige Lösung für (W, O) als Ausgangspunkt wählen und durch ein hinreichend kleines λ den Poorness-of-Fit gegen Null verschieben — selbstverständlich wird auf diese Weise keine sinnvolle Lösung gefunden.

Srinivasan & Shocker lösen das Problem der trivialen Lösungen durch folgende Bedingung, die jede Lösung (W, O) erfüllen muß:

$$G - B = h \tag{1.2.7}$$

wobei h eine beliebige, jedoch strikt positive Zahl ist. Die Bedeutung von Bedingung (1.2.7) wird deutlich, wenn man B und G durch (1.2.5) und (1.2.6) ersetzt und umformt:

$$\sum_{(j,r)\in\Omega} (s_r - s_j)^+ - \sum_{(j,r)\in\Omega} (s_r - s_j)^- = \sum_{(j,r)\in\Omega} (s_r - s_j) = h \tag{1.2.8}$$

Die triviale Lösung $w_l = 0$ (für alle $l \in L$) führt zu $h = 0$ und also zum Widerspruch zur strikten Positivität von h. Wenn eine Lösung (W, O) Gleichung

(1.2.8) erfüllt, dann gilt für die durch $w'_l = \lambda w_l$ transformierte Lösung (W', O): $\sum(s'_r - s'_j) = \lambda \sum(s_r - s_j)$ — die transformierte Lösung erfüllt Gleichung (1.2.8) nur dann, wenn $\lambda = 1$.

Es folgen Umformungen, die zur Formulierung eines linearen Programms zur Minimierung von B (vgl. S. 34) notwendig sind. Aus Gleichung (1.2.2) folgt:

$$\begin{aligned} s_r - s_j &= \sum_{l=1}^{t} w_l (x_{rl} - o_l)^2 - \sum_{l=1}^{t} w_l (x_{jl} - o_l)^2 \\ &= \sum_{l=1}^{t} w_l (x_{rl}^2 - x_{jl}^2) - 2 \sum_{l=1}^{t} w_l o_l (x_{rl} - x_{jl}) \end{aligned} \tag{1.2.9}$$

Es seien

$$a_{jrl} = x_{rl}^2 - x_{jl}^2 \tag{1.2.10}$$

$$b_{jrl} = -2(x_{rl} - x_{jl}) \tag{1.2.11}$$

— für $(j, r) \in \Omega$ — und:

$$V = \{v_l\} = \{w_l o_l\} \tag{1.2.12}$$

Man erhält:

$$s_r - s_j = \sum_{l=1}^{t} w_l a_{jrl} - \sum_{l=1}^{t} v_l b_{jrl} \tag{1.2.13}$$

Aus (1.2.5) und (1.2.13) läßt sich die Definition des Poorness-of-Fit B für eine Lösung (W, V) so formulieren:

$$B(W, V) = \sum_{(j,r)\in\Omega} \max \left\{ 0, - \left(\sum_{l=1}^{t} w_l a_{jrl} - \sum_{l=1}^{t} v_l b_{jrl} \right) \right\} \tag{1.2.14}$$

Ferner seien:

$$A_l = \sum_{(j,r)\in\Omega} a_{jrl} \tag{1.2.15}$$

$$D_l = \sum_{(j,r)\in\Omega} b_{jrl} \tag{1.2.16}$$

Aus (1.2.13), (1.2.15) und (1.2.16) läßt sich Gleichung (1.2.8) umformulieren:

$$\sum_{l=1}^{t} w_l A_l + \sum_{l=1}^{t} v_l D_l = h \tag{1.2.17}$$

Bedingung (1.2.7) kann also durch Bedingung (1.2.17) ersetzt werden. Das Ziel der Prozedur, die optimale Lösung (W, O) zu finden, kann jetzt unter Zuhilfenahme der „neuen Parameter“ $v_l \in V$ so formuliert werden: Finde eine

Lösung (W, V) , die Gleichung (1.2.14) minimiert unter Berücksichtigung der Einschränkungen (1.2.3) und (1.2.17). Das Zielkriterium minimaler Poorness-of-Fit wird B^* benannt:

$$B^* = \min_{(W,V) erfüllt (1.2.3),(1.2.17)} \{B(W, V)\} \tag{1.2.18}$$

Srinivasan & Shocker liefern nun mit dem bisherigen formalen Rüstzeug die Formulierung eines linearen Programms („linear program“). Sie machen keine Angaben darüber, wie dieses lineare Programm im Detail abläuft, sondern begnügen sich mit den allgemein gehaltenen Feststellungen, daß lineare Programme eben die Optimierung linearer Funktionen unter Nebenbedingungen bzw. Restriktionen leisten, daß das folgende lineare Programm unter Rückgriff auf die Simplex-Methode nach Dantzig gelöst werden könne und daß ihre Formulierung zu einer Klasse linearer Programmierungsprobleme gehöre, die unter dem Begriff „goal programming“ bekannt seien. Ansonsten verweisen sie hier auf spezifische Literatur zur Theorie linearer Programmierung (vgl. ebd. S. 346) — ich folge den Autoren auch hier, denn der Versuch, das Thema zu vertiefen, würde wohl, allem Vorsatz zur detaillierten Beschreibung von LINMAP zum Trotz, zu weit vom eigentlichen Thema der vorliegenden Arbeit ablenken.

Die **Formulierung des linearen Programms** also lautet:

$$\text{Minimiere} \quad \sum_{j,r \in \Omega} z_{jr} = F \tag{1.2.19}$$

unter den folgenden Restriktionen:

$$\sum_{l=1}^{t} a_{jrl} w_l + \sum_{l=1}^{t} b_{jrl} v_l + z_{jr} \geq 0 \quad \text{für } (j, r) \in \Omega \tag{1.2.20}$$

$$\sum_{l=1}^{t} A_l w_l + \sum_{l=1}^{t} D_l v_l = h \tag{1.2.21}$$

$$w_l \geq 0 \tag{1.2.22}$$

und

$$z_{jr} \geq 0 \quad \text{für } (j, r) \in \Omega \tag{1.2.23}$$

Die Variablen z_{jr} können mit den Worten der Autoren als „artificial variables“ betrachtet werden. Wenn man Restriktion (1.2.20) umformuliert zu:

$$z_{jr} \geq -\left(\sum_{l=1}^{t} a_{jrl} w_l + \sum_{l=1}^{t} b_{jrl} v_l \right)$$

Dann folgt aus (1.2.23):

$$z_{jr} \geq \max\left\{0, -\left(\sum_{l=1}^{t} a_{jrl} w_l + \sum_{l=1}^{t} b_{jrl} v_l\right)\right\}$$

Da das durch (1.2.19) vorgegebene Ziel die Minimierung der Summe der z-Werte ist, läßt sich für die optimale Lösung folgern:

$$z_{jr} = \max\left\{0, -\left(\sum_{l=1}^{t} a_{jrl} w_l + \sum_{l=1}^{t} b_{jrl} v_l\right)\right\}$$

D.h. also $F = B(W, V,)$ und für die optimale Lösung:

$$F^* = B^* \tag{1.2.24}$$

Man könnte die Vorgehensweise des linearen Programms also in Worten so beschreiben: Das Programm versucht, positive Werte für die „künstlichen“ Variablen z_{jr} zu finden, so daß deren Summe minimiert wird und gleichzeitig die Restriktionen (1.2.20) bis (1.2.22) für die w_l und v_l gelöst werden können. Srinivasan & Shocker führen dazu im Anhang ihrer Arbeit 2 Beweise. Erstens wird noch einmal explizit bewiesen, daß Gleichung (1.2.24) gilt, daß also tatsächlich die optimale Lösung für F^* den Poorness-of-Fit $B(W, V)$ minimiert[20] (ebd., S.366: Theorem 1). Zweitens wird bewiesen, daß das Programm eine finite optimale Lösung besitzt[21].

Eine Lösung (W, V) könnte nun nach Gleichung (1.2.12) in die eigentlich gesuchte Lösung (W, O) transformiert werden — dies wird aber nicht immer möglich sein. Es werden durch das Programm ja die v_l-Werte geschätzt, ohne irgendeiner Restriktion, in die Gleichung (1.2.12) einwirkt, ausgesetzt zu sein. Es sind für jedes Attribut $l \in L$ folgende Fälle denkbar:

(1) $w_l > 0, v_l \neq 0$: Dann ist $o_l = v_l / w_l$.

(2) $w_l = 0, v_l = 0$: D.h. die Dimension l trägt nichts zum Distanzmaß s_j bei bzw. ist (für die Versuchsperson) irrelevant hinsichtlich der Bildung von Präferenzen.

[20] Der Beweis läuft in 2 Schritten ab: Zuerst wird gezeigt, daß für die optimale Lösung (W^*, V^*) des Programms gilt: $F^* = B(W^*, V^*)$. Dann wird gezeigt, daß (W^*, V^*) tatsächlich $B(W, V)$ über alle (W, V) hinweg minimiert.

[21] Dazu müssen nur die beiden anderen Möglichkeiten, die für Lösungen des linearen Programms nach der Simplex-Methode noch bestehen, ausgeschlossen werden — nämlich daß das Programm keine Lösung besitzt, oder daß es ein ungebundenes Optimum (eine Lösung für die Parameter, durch die die Funktion F irgendeiner arbiträren großen negativen Zahl gleichgesetzt werden kann) besitzt. Ersteres wird ausgeschlossen, indem die immer mögliche Lösung $w_l = 0$ für alle $l \in L$ und $v_l = 0$ für $l = 1, 2, \ldots, t-1$ und $v_t = h / D_t$ gezeigt wird. Letzteres ist bereits durch Restriktion (1.2.23) ausgeschlossen.

(3) $w_l = 0, v_l \neq 0$: o_l kann nicht nach (1.2.12) berechnet werden bzw. (1.2.12) führt zu einem infiniten Wert für o_l. Damit wird das Idealpunktmodell für Dimension l zum sogenannten Vektormodell (vgl. Green & Srinivasan 1978, S. 105). Nach dem Vektormodell ergibt sich der Gesamtnutzenwert für Stimulus j aus $\sum_{l \in L} u_l x_{jl}$, wobei u_l ein „Wichtigkeitsfaktor“ (die Gewichtung) für die Dimension l sein soll. Es stellt also den einfachsten Fall der Modellierung subjektiver Präferenzstrukturen dar, wonach eine lineare Beziehung zwischen kontinuierlich gemessener Ausprägung des Attributs und Präferenz besteht. Die Differenz zwischen 2 Stimuli j und r beträgt für das Vektormodell $u_l(x_{jl} - x_{rl})$ — betrachtet man den rechten Ausdruck von Gleichung (1.2.9) unter der Annahme von $w_l \to 0$ und $o_l \to \infty$, dann erhält man für die Differenz zwischen r und j nach dem Idealpunktmodell $-2v_l(x_{rl} - x_{jl})$, so daß also für diesen Fall durch $2v_l = u_l$ für die Dimension l eine Lösung nach dem Vektormodell gefunden ist.

Damit aber entspricht das lineare Programm einem *„mixed model“*, also einem Präferenzmodell, das es erlaubt, daß der Einfluß einzelner Attribute nach dem Idealpunktmodell, der andere nach dem Vektormodell modelliert wird (vgl. wieder Green & Srinivasan 1978, S. 105ff). Wenn $L_1 \subseteq L$ die Dimensionen enthält, für die gilt $w_l > 0$ oder $w_l = v_l = 0$ und $L_2 = L - L_1$ diejenigen mit $w_l = 0$ und $v_l \neq 0$, dann kann unter Bezugnahme auf die Ausdrücke im letzten Absatz eine modifizierte quadrierte Distanz $\tilde{s_j}$ so definiert werden (vgl. Gleichung (1.2.2):

$$\tilde{s_j} = \sum_{l \in L_1} w_l(x_{jl} - o_l) - 2 \sum_{l \in L_2} v_l x_{jl}$$

Man kann sich leicht davon überzeugen, daß $(\tilde{s_r} - \tilde{s_j})$ der Differenz $(s_r - s_j)$ in Gleichung (1.2.9) entspricht, wenn man für alle $l \in L_2$ v_l durch $w_l o_l$ ersetzt.

Diese Kompatibilität des Vektormodells mit dem für das Idealpunktmodell formulierten linearen Programm besteht aber nicht nur für den Fall „nachträglicher“ Entdeckung solcher Attribute, deren Idealpunkt gegen $\pm\infty$ geht, sondern es können auch a priori solche Dimensionen festgelegt werden: LINMAP erlaubt ein Vielzahl von Restriktionen für die Parameter, man könnte also z.B. von vorneherein einzelne der w_l gleich Null setzen. Ein von dieser Möglichkeit etwas abweichender Weg der Formulierung des linearen Programms für ein volles Vektormodell (für alle Attribute) wird in Srinivasan & Shocker (1973b) gezeigt: Der Wert s_j wird anstatt wie in Gleichung (1.2.2) als quadrierte Distanz, als sogenanntes „composite criterion“ so definiert:

$$s_j = \sum_{l=1}^{t} w_l x_{jl} \tag{1.2.25}$$

Die Autoren verwenden in diesem Artikel dieselbe Notation, was in der Zusammenschau verwirren kann, da eben innerhalb des allgemeineren Idealpunktmodells das Vektormodell sich aus $w_l = 0$ ergibt und so die w_l des Composite-Criterion sozusagen den $2v_l$ der quadrierten Euklidischen Distanzen entsprechen. Trotzdem hat die Analogie der Schreibweisen Vorteile: Zum einen haben die w_l in Gleichung (1.2.25) dieselbe Bedeutung, wie die in (1.2.2) als Gewichte der Attribute, in die sowohl unterschiedliche Wichtigkeit, als auch unterschiedliche Skalierung der Attribute miteingeht. Zum andern aber braucht so das gesamte Verfahren für das Composite-Criterion nicht neu dargestellt zu werden: Es gelten alle oben aufgeführte Gleichungen, nachdem in ihnen $o_l = 0$ gesetzt und die Quadrierung der x_{jl} entfernt wurde — in der Folge verschwinden die v_l, b_{jrl} und D_l ganz aus den Formulierungen. Ansonsten aber bleibt alles gegenüber der oben gezeigten Prozedur gleich.

Damit — mit der Möglichkeit mittels der LINMAP-Prozedur die Parameter eines Vektormodells zu schätzen — ist aber auch der Weg frei für das „part-worth model“, für das Modell also, das im Abschnitt 1.1.2 durch Gleichung (1.1.1) als das eigentliche, „typische“ Modell der CA vorgestellt wurde. Bei der **Anwendung von LINMAP für eine CA** mit kategorial abgestuften Attributen wird über eine Dummy-Kodierung der Stimulusprofile sozusagen ein Pseudovektormodell definiert (vgl. Srinivasan & Shocker 1973b, S. 490f): Dazu wird jede Stufe k eines jeden Attributs l zur neuen Dimension l' gemacht und die Ausprägungen $x_{jl'}$ der Stimuli aus J auf diesen Dimensionen werden in gewohnter Weise dummy-kodiert — in der Folge erhält man die Schätzungen der Teilnutzenwerte durch die optimale Schätzung W^* der Gewichte $w_{l'}$. Hat man z.B. 3 Attribute mit jeweils 3 Ausprägungen, so resultieren 9 Dimensionen l' und $x_{jl'} = 1$, wenn Stimulus j die durch Dimension l' gemeinte Ausprägung „besitzt“, ansonsten $x_{jl'} = 0$. Die 9 Gewichte $w_{l'}$ sind dann die Teilnutzenwerte für die 9 Stufen.

Nun hat zwar LINMAP mit G bereits ein Maß für den Goodness-of-Fit, jedoch ist der Wert von G — genauso wie der von B — abhängig vom gewählten Wert h: Wird ein Wert h in Restriktion (1.2.21) durch einen anderen Wert q ersetzt, so gilt: $F_q^* = (q/h)F_h^*$ (dies folgt aus der Theorie linearer Programmierung; vgl. Srinivasan & Shocker 1973a, S.348)). Nach Gleichung (1.2.24) folgt dann auch $B_q^* = (q/h)B_h^*$ und auch $W_q^* = (q/h)W_h^*$ und $V_q^* = (q/h)V_h^*$ können aus den aufgeführten Gleichungen gefolgert werden. Eine Veränderung des Wertes h bewirkt also eine Veränderung der absoluten Beträge B^*, W^*, V^*, nicht aber der relativen Beträge. Es wird nun noch ein *Index-of-Fit* C definiert, dessen Absolutbetrag gegenüber der Wahl von h invariant ist:

$$C = \frac{B}{(h + B)} \tag{1.2.26}$$

Da C eine streng monoton steigende Transformation von B ist, gilt auch:

$$C^* = \frac{B^*}{(h + B^*)} \tag{1.2.27}$$

Dieser Index-of-Fit bietet einige Vorteile: Er ist, wie schon gesagt, invariant gegenüber dem gewählten h und er kann nur Werte zwischen 0 und 1 annehmen (da $B \geq 0$), sodaß er Maßen wie dem Streß bei MONANOVA (vgl. Gleichung (1.1.4) und $(1-R^2)$ der multiplen Regression (vgl. Abschnitt 1.2.2) vergleichbar ist. Auch ist C unabhängig von der Anzahl der Paarvergleiche in Ω.[22]

Im „Algorithm 1" (ebd., S. 349) fassen Srinivasan & Shocker die Berechnungen ihrer Prozedur (der externalen Analyse) zusammen, er enthält nur Wiederholungen des bereits Dargestellten, so daß er hier nicht noch einmal wiedergegeben wird. Es ist allerdings darauf hinzuweisen, daß in diesem LINMAP-Algorithmus $h = 1$ gesetzt ist, was aufgrund der vorangegangen Überlegungen zu Index C^* keinen Verlust an Allgemeinheit für die Prozedur bedeutet.

Im folgenden (ebd., S. 349) diskutieren Srinivasan & Shocker verschiedene *Gesichtspunkte und Eigenschaften der (externalen) Analyse mit LINMAP*, von denen hier einige wichtig erscheinende Punkte kurz aufgegriffen werden sollen:

(1) Es können sich als Resultat der LINMAP-Analyse alternierende Optima („alternate optima") ergeben, d.h. es kann mehr als eine Lösung (W, V) herauskommen, durch die derselbe Wert F^* erreicht wird. Die Autoren zeigen ein Verfahren zur Auswahl zwischen diesen optimalen Lösungen (ebd., S. 350).

(2) Es wurde bereits darauf hingewiesen, daß die Gewichte w_l sowohl die subjektive Wichtigkeit der verschiedenen Attribute für die Präferenzurteile des Individuums, als auch Unterschiede der Maßeinheiten der Attribute wiederspiegeln. Wenn man Gewichte wünscht, die lediglich subjektive Wichtigkeit ausdrücken und den Einfluß der Maßeinheit nicht mehr enthalten, dann sollte man die Ausprägungen der Attribute in Einheiten der Standardabweichung ausdrücken: $x'_{jl} = (x_{jl} - \bar{x}_l)/\sigma_l$ ($\bar{x}_l$ und σ_l sind Mittelwert und Standardabweichung der Werte x_{jl} des Stimulus j auf Dimension l). Wenn man wünscht, die Gewichte verschiedener Personen zu vergleichen, dann sollte man die w_l und v_l pro Person so transformieren, daß jeweils ihre Summe 1 ergibt.

(3) Die Autoren liefern Modifikationen der Prozedur, die nötig sind, um Indifferenzen bei den Paarvergleichen zu integrieren. Diese werden hier

[22] Zur Unabhängigkeit von h: Wird h nach q verändert, dann gilt $C^*_q = B^*_q/(q + B^*_q) = (q/h)B^*_h/(q + (q/h)B^*_h) = B^*_h/(h + B^*_h) = C^*_h$. Zur Unabhängigkeit von der Zahl der Paarvergleiche: $C^* = B^*/(h + B^*) = B^*/G^*$ und die Zahl der Paare in Ω beeinflußt B^* und G^* in gleicher Weise (vgl. (1.2.5) und (1.2.6)).

nicht alle im Detail aufgeführt. Das Prinzip besteht darin, die Menge Ω aller Paarvergleiche in eine Teilmengen Ω' der streng geordneten Paare (j, r) und in eine Teilmenge Φ der indifferenten Paare zu unterteilen. Aus Φ wird dann eine Menge Ψ gebildet, indem für jedes Paar (e, f) aus Φ die Paare (e, f) und (f, e) in Ψ aufgenommen werden. Ausgehend von für die Paare aus Ω' und Ψ getrennten Formulierungen der Gleichung (1.2.4) ergeben sich dann die dort aufgeführten (ebd., S. 353) Veränderungen des linearen Programms.

(4) Die Prozedur erlaubt eine Vielzahl von Restriktionen für die Parameter. Die Darlegung aller Möglichkeiten, die die Autoren diskutieren, würde den gegenwärtigen Abschnitt wohl in einer Weise aufblähen, die dem Ziel der Darstellung des Verfahrens als theoretische Grundlage der ab Teil 2 wiedergegebenen Untersuchung nicht mehr angemessen wäre — es sei hier also auf die Orginalliteratur verwiesen (ebd., S. 356ff).

(5) Letzteres gilt auch für die Modifikationen der Prozedur für intervallskalierte Präferenzdaten. Man könnte solche natürlich auch in ordinale Paarvergleiche überführen, würde damit aber sozusagen Information verschenken. Da das Verfahren auch für intervallskalierte Präferenzdaten praktikabel sein soll, Srinivasan & Shocker aber diesen Informationsverlust vermeiden wollen, liefern sie dazu ein modifiziertes lineares Programm auf der Basis der Formulierung eines MSAE-Regressionsmodells (MSAE: „*m*inimizing the *s*um of *a*bsolute *e*rrors“) (ebd., S. 358ff).

(6) Liegen die ordinalen Präferenzurteile nicht in Form von Paarvergleichen vor, sondern als Rangwerte über die Stimuli, dann können diese Rangwerte über n Stimuli natürlich in $n(n-1)/2$ Paarvergleiche umgewandelt werden.

Wie eingangs erwähnt, formulieren Srinivasan & Shocker (ebd., S. 360f) auch einen *Algorithmus zur internalen Analyse*, bei der nicht nur die Gewichte w_l und die Idealpunkte o_l geschätzt werden, sondern auch die Lage der Stimuli im t-dimensionalen Attributraum, also auch die Ausprägungen x_{jl}. Dies bedingt eine gewisse Aggregation der individuellen Präferenzdaten, da es sich bei den Werten x_{jl} nicht mehr um individuelle Größen handelt und sie also für alle Mitglieder einer Stichprobe geschätzt werden. Es wird also zunächst die Menge der Individuen $I = \{1, 2, \ldots, m\}$ definiert. Für $i \in I$ kann dann das lineare Programm (1.2.19) bis (1.2.23) wie folgt neu formuliert werden:

$$\text{Minimiere} \quad \sum_{(j,r)\in\Omega_i} z_{ijr} = F_i \tag{1.2.28}$$

Unter folgenden Restriktionen:

$$\sum_{l=1}^{t} w_{il}\left(x_{rl}^2 - x_{jl}^2\right) - 2\sum_{l=1}^{t} v_{il}\left(x_{rl} - x_{jl}\right) + z_{ijr} \geq 0 \quad \text{für} \quad (j,r) \in \Omega_i \tag{1.2.29}$$

$$\sum_{l=1}^{t} w_{il}\Big(\sum_{(j,r)\in\Omega_i}(x_{rl}^2 - x_{jl}^2)\Big) - \sum_{l=1}^{t} v_{il}\Big(\sum_{(j,r)\in\Omega_i} 2(x_{rl} - x_{jl})\Big) = 1 \tag{1.2.30}$$

$$w_{il} \geq 0 \quad \text{für} \quad l \in L, \quad z_{ijr} \geq 0 \quad \text{für} \quad (j,r) \in \Omega_i \tag{1.2.31}$$

Ein Kriterium für den Fit einer Lösung (X) der internalen Analyse wird nun definiert als Mittel der individuellen optimalen Poorness-of-Fit Werte für die durch die Lösung gegebenen Werte x_{jl}:

$$D(X) = \frac{1}{m}\sum_{i=1}^{m} F_i^*(X) \tag{1.2.32}$$

$$D^* = \min D(X) \tag{1.2.33}$$

Es wird also versucht, den mittleren Poorness-of-Fit über alle Mitglieder der Stichprobe hinweg durch systematische Variation der Werte x_{jl} zu minimieren.

Srinivasan & Shocker schlagen dazu die folgende iterative Prozedur vor:

(i) Wähle (beliebige) Werte x_{jl}.

(ii) Berechne das Programm (1.2.28) bis (1.2.31) für die gegebenen Werte x_{jl} und alle $i \in I$.

(iii) Minimiere die F_i erneut durch systematische Veränderung (s.u.) der Werte x_{jl} — unter Beibehaltung der in (ii) erzielten Lösungen (W_i, V_i) — und berechne den Rückgang in D, der dadurch erreicht wird.

(iv) Beende die Prozedur, wenn die Reduktion von D einem festgelegten Wert unterschreitet.

(v) Verändere die x_{jl}-Werte entsprechend der Ergebnisse aus (iii) und kehre zurück zu (ii).

Der obige Schritt (iii) bedarf der Erläuterung[23]: Es seien Δ_{jl} die Veränderungen der Werte x_{jl} von einer Iteration zur nächsten:

$$x_{jl} = \dot{x}_{jl} + \Delta_{jl} \tag{1.2.34}$$

wobei $\dot{x}_{jl}$ die Ausprägungen der Stimuli j auf den Attributen l im vorangegangen Iterationsschritt sind. Wenn die Δ_{jl} sehr klein im Verhältnis zu den $\dot{x}_{jl}$ sind, dann können die x_{jl}^2 näherungsweise bestimmt werden:

$$x_{jl}^2 \simeq \dot{x}_{jl}^2 + 2\Delta_{jl}\dot{x}_{jl} \tag{1.2.35}$$

[23]Ich weiche hier von der Notation des Originaltexts (vgl. ebd., S. 361) stärker, als sonst in meiner Darstellung, ab, da diese Notation mir hier etwas verwirrend erscheint.

Die x-Werte in den Restriktionen (1.2.29) und (1.2.30) werden nun durch die rechtsseitigen Ausdrücke in (1.2.34) und (1.2.35) ersetzt und bei gegebenen Werten w_{il}, v_{il} und $\dot{x}_{jl}$ sucht das Programm (1.2.28) bis (1.2.32) Lösungen für z_{ijr} und Δ_{jl}. Um sicherzustellen, daß die Δ_{jl} hinreichend klein im Verhältnis zu den $\dot{x}_{jl}$ sind, werden noch die folgenden Restriktionen hinzugefügt[24]:

$$\lambda \, \mathcal{M}_{\substack{j\in J\\ l\in L}}\left(|\dot{x}_{jl} - \dot{x}_{\bullet l}|\right) \;\geq\; \Delta_{jl} \;\geq\; -\lambda \, \mathcal{M}_{\substack{j\in J\\ l\in L}}\left(|\dot{x}_{jl} - \dot{x}_{\bullet l}|\right)$$

wobei:

λ	=	ein kleiner Wert > 0 (die Autoren schlagen .05 vor).
$\dot{x}_{\bullet l}$	=	der Median der $\dot{x}_{jp}$ über j auf der Dimension l.
$\mathcal{M}_{\substack{j\in J\\ l\in L}}(\ldots)$:		Mittelwert über alle j und l.

Damit beende ich die Beschreibung von LINMAP. Die grundsätzliche Vorgehensweise ist nun, so hoffe ich, einigermaßen erschöpfend dargelegt und auf die vielfältigen, über die Anwendung als nonmetrisches Verfahren zur Lösung von „typischen“ CA-Problemen hinausgehenden Möglichkeiten des Verfahrens sind zumindest angesprochen worden.

1.2.2 OLS-Regression

Im Gegensatz zu LINMAP ist die multiple Regression mit Kleinstquadrateschätzung (OLS: „*o*rdinary *l*east *s*quares“) ein Verfahren, das man wohl zu den Standards der sozialwissenschaftlichen Methodenlehre zählen darf und dessen Bekanntheit hier im Grunde vorausgesetzt werden könnte. OLS wird, um der Vollständigkeit willen — sozusagen um die Symmetrie der Darlegung der Grundlagen für die ab Teil 2 vorgestellte Untersuchung zu wahren —, im folgenden dennoch dargestellt, allerdings nicht in der Ausführlichkeit wie LINMAP, was wohl mit Blick eben auf die Bekanntheit der Methode und auf die Unzahl von Veröffentlichungen, in denen der OLS-Algorithmus erläutert wird, zu rechtfertigen ist. Ich beziehe mich in meiner Darstellung v.a. auf Werner (1993), eine weitere sehr gute Beschreibung findet sich bei Hanushek & Jackson (1977).

Ihre zentrale Bedeutung (für die sozialwissenschaftliche Methodenlehre) erhält die OLS-Schätzung als Rechenalgorithmus für alle univariaten Submodelle des allgemeinen linearen Modells (ALM): Sowohl Regressionsmodelle mit kontinuierlich skalierten Prädiktoren, als auch unterschiedliche (univariate) varianz- und kovarianzanalytische Designs sind durch OLS-Regression schätzbar (vgl. Werner 1993). Die im Abschnitt 1.1.2 genannten metrischen Verfahren sind

[24] So jedenfalls habe ich die Gleichungen (112) und (113) auf S. 361 (ebd.) verstanden.

somit nicht als konkurrierende Möglichkeiten aufzufassen, sondern ANOVA[25] und OLS-Regression benennen dort nur unterschiedliche Prozeduren, die zu denselben Ergebnissen führen (vgl. Werner 1993, S. 191f: Einleitung zu Kapitel 4). Die OLS-Schätzung ist gegenüber der auf „traditionelle“ Art (durch Quadratsummenbildung) durchgeführten ANOVA also sozusagen der universellere Algorithmus, weswegen er auch in der ab Teil 2 dargestellten Untersuchung als Verfahren der metrischen CA verwendet wurde — man hätte aber genauso gut den spezifischen ANOVA-Algorithums verwenden können und man *kann* genauso gut die durchgeführten OLS-Regressionen als Varianzanalysen bezeichnen.

Bei der (univariaten) multiplen Regression werden zur Vorhersage einer abhängigen Variablen (Kriteriumsvariable) y mehrere unabhängige Variablen (Prädiktorvariablen) $x_1, x_2, \ldots, x_{p-1}$ verwendet. Für die Beobachtungen j der Variablenwerte läßt sich das *Grundmodell der multiplen Regression* so schreiben:

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_{p-1} x_{jp-1} + \epsilon_j \quad \text{für } j = 1, 2, \ldots, n \quad (1.2.36)$$

Das Modell enthält zwei „charakteristische Elemente“:

(1) Die Werte $\beta_0, \beta_1, \ldots, \beta_{p-1}$ sind die eigentlichen Parameter des Modells, die (Partial-)*Regressionskoeffizienten*. Aus Gleichung (1.2.36) läßt sich die inhaltliche Bedeutung dieser Parameter ableiten: Der Koeffizient einer Prädiktorvariablen gibt an, um wieviele Einheiten sich y verändert, wenn der Wert dieser Prädiktorvariablen um eine Einheit erhöht wird, und alle anderen Prädiktoren konstant gehalten werden. Besondere Beachtung verdient dabei β_0, die sogenannte (Regressions- oder additive) Konstante: Dieser Parameter resultiert aus dem Umstand, daß die verschiedenen Variablen in Gleichung (1.2.36) nicht standardisiert sind, daß also sozusagen ihre Werte aus unterschiedlichen numerischen Bereichen stammen — folglich entfällt β_0, wenn die Kriteriums- und Prädiktorvariablen hinsichtlich Mittelwert und Varianz standardisiert werden. Anschaulich wird die Bedeutung der Konstanten, wenn man die Regression geometrisch darstellt: β_0 gibt an, wo die Regressionsebene die y-Achse schneidet (siehe unten). Da die Variablen in Gleichung (1.2.36) nicht standardisiert sind, sind auch die verschiedenen β-Gewichte nicht miteinander vergleichbar: Man stelle sich z.B. vor, die Länge eines Gegenstandes fungiere als Prädiktorvariable einer Regression und der Wert des Kriteriums wachse um 0.5, wenn sich die Länge des Gegenstandes um einen Meter erhöht. Dann erhält dieser Prädiktor den Koeffizienten 0.5, wenn die Länge in Metern, und 0.005, wenn sie in Centimetern gemessen wurde. Dieses einfache Beispiel zeigt, daß die Höhe eines

[25] Im Falle individueller metrischer CA ist immer die orthogonale ANOVA mit der Zellhäufigkeit 1 gemeint.

Koeffizienten von der Varianz des Pädiktors abhängt: Wird die Länge in kleineren Einheiten gemessen, dann muß der numerische Betrag der Varianz, der ja in diesen Einheiten ausgedrückt wird, größer werden: Größere Varianz bedingt kleinere Koeffizienten. Um also verschiedene Koeffizienten vergleichbar zu machen, muß eine Standardisierung erfolgen: Entweder die Standardisierung der Variablen des Modells, oder eine direkte Standardisierung der Koeffizienten nach der Formel $\beta^s = \beta s_x / s_y$ (β^s ist der standardisierte Koeffizient, s_x die Standardabweichung des zugehörigen Prädiktors, s_y die des Kriteriums). Standardisierte Koeffizienten geben an, um wieviel Standardabweichungen sich das Kriterium verändert, wenn der Wert des Prädiktors um eine Standardabweichung nach oben gesetzt wird. Sie sind gegenüber unstandardisierten Parametern insofern ungenauer, als sie „stichprobenabhängiger" sind, da jede Standardisierung auf Schätzungen der Stichprobenkennwerte beruht.

(2) Der Ausdruck ϵ_j bezeichnet den „*Fehler*" der Beobachtung j. Unter dem Fehler werden alle Einflüsse auf die (empirisch ermittelten) Werte der Kriteriumsvariablen subsumiert, welche zusätzlich zu den durch die Prädiktoren spezifizierten noch wirksam sind. Inhaltlich können das sein: In der Modellgleichung nicht enthaltene Prädiktoren (die gemeinsame Varianz mit dem Kriterium aufweisen), „echte" Zufälligkeiten in der Ausprägung der Kriteriumsvariablen, Meßfehler bei der Messung des Kriteriums (nicht jedoch der Prädiktoren, siehe unten). Das Modell zerfällt also in eine deterministische, systematische Komponente $\tilde{y}_j = \beta_0 + \beta_1 x_{j1} + \ldots + \beta_{p-1} x_{jp-1}$ und in eine nicht-deterministische, sogenannte Zufallskomponente („random component" bei Hanushek & Jackson 1977) ϵ_j. Man kann diesen Umstand auch so ausdrücken: Durch das Modell der multiplen Regression wird die Gesamtvarianz der Kriteriumsvariablen in einen Anteil (durch die Prädiktoren) „aufgeklärter" Varianz und einen Anteil „unaufgeklärter" Fehlervarianz zerlegt. Die Werte $\tilde{y}_j$ der systematischen Komponente sind die Vorhersagen der Kriteriumswerte aus dem Modell, es gilt: $\epsilon_j = y_j - \tilde{y}_j$.

Die üblichere und zweckmäßigere Darstellung des Regressionsmodells erfolgt in *Matrizenschreibweise*:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.2.37}$$

wobei:

$\mathbf{y}$	=	der $1 \times n$-Vektor der Werte der Kriteriumsvariablen.
$\mathbf{X}$	=	die $n \times p$-Prädiktorenmatrix.
$\boldsymbol{\beta}$	=	der $p \times 1$-Vektor der β-Gewichte.
$\boldsymbol{\epsilon}$	=	der $1 \times n$-Vektor der Fehlerwerte.

Die Prädiktorenmatrix $\mathbf{X}$ enthält in der ersten Spalte nur Einsen (für die Regressionskonstante) und besteht ansonsten aus den Spaltenvektoren der Werte

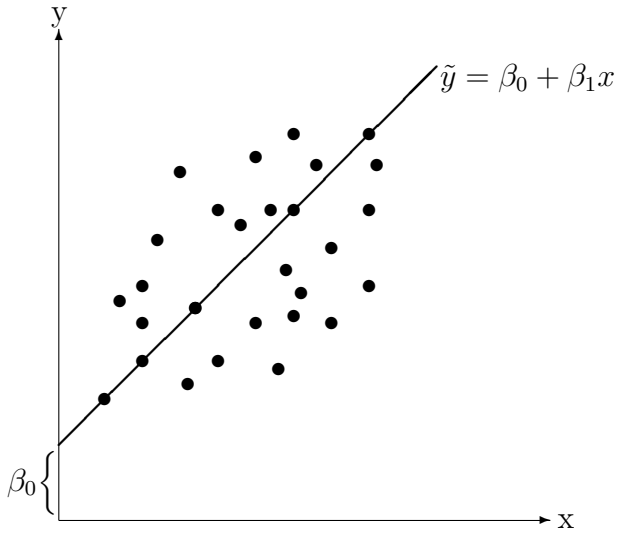

Abbildung 1.4: Geometrische Darstellung der Regressionsgeraden bei nur einem Prädiktor x.

der $p-1$ Prädiktorvariablen.

Die Bedeutung der β-Gewichte läßt sich geometrisch wie folgt darstellen: Die Werte $y_j, x_{j1}, \ldots, x_{jp-1}$ der Beobachtung j definieren einen Punkt im p-dimensionalen Koordinatensystem mit den Achsen $y, x_1, \ldots, x_{p-1}$. Durch den OLS-Algorithmus wird nun versucht, durch den Schwarm der n Punkte eine (Hyper-)Ebene so zu legen, daß die Summe der quadrierten Abstände der Punkte zu der Ebene entlang der y-Dimension minimiert wird. Abbildung 1.4 veranschaulicht dieses für die einfache Regression mit nur einer Prädiktorvariablen: Durch den Punkteschwarm der Beobachtungen im zweidimensionalen Koordinatensystem wird die Regressionsgerade $\tilde{y} = \beta_0 + \beta_1 x$ gelegt, der Schnittpunkt der Geraden mit der y-Achse bildet die Regressionskonstante, die Steigung der Geraden den Wert β_1 und die Werte ϵ_j ergeben sich als vertikale Abstände der Punkte von der Geraden.

Das *Ziel der OLS-Schätzung* der β-Gewichte ist es also, die Summe der quadrierten Differenzen zwischen den tatsächlichen und den aus der systematischen Komponente vorhergesagten Kriteriumswerten — die Summe der quadrierten Fehler — zu minimieren:

$$\text{Minimiere} \quad \sum_{j=1}^{n} (y_j - \tilde{y}_j) = \boldsymbol{\epsilon'\epsilon}$$

Für $\boldsymbol{\epsilon'\epsilon}$ gilt (vgl. (1.2.37)):

$$\boldsymbol{\epsilon'\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta'}\mathbf{X}'\mathbf{y} + \boldsymbol{\beta'}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \tag{1.2.38}$$

Zur Minimierung von $\boldsymbol{\epsilon'\epsilon}$ muß (1.2.38) partiell nach $\boldsymbol{\beta}$ abgeleitet werden:

$$\frac{\partial(\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta'}\mathbf{X}'\mathbf{y} + \boldsymbol{\beta'}\mathbf{X}'\mathbf{X}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

und es muß die Schätzung $\mathbf{b}$ für $\boldsymbol{\beta}$ gesucht werden, durch die die erste Ableitung gleich Null wird:

$$-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = 0$$

Dies führt zu den **OLS-Schätzungen** der Parameter des Modells[26]:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{1.2.39}$$

Die besondere Bedeutung der OLS-Schätzer ergibt sich daraus, daß sie, sofern die *Voraussetzungen der OLS-Regression* erfüllt sind, BLUE („*b*est *l*inear *u*nbiased *e*stimator“) sind, d.h sie sind erwartungstreu und haben unter allen erwartungstreuen linearen Schätzern die kleinste Varianz. Die beiden wichtigsten Voraussetzungen für die OLS-Schätzung wurden bereits auf S. 19 genannt — es handelt sich um Annahmen über Mittelwert und Streuung des Fehlers. Eine ausfühliche Diskussion der Voraussetzungen liefern Hanushek & Jackson (1977, S. 47ff und S. 115ff), die folgende Zusammenstellung folgt weitgehend dieser Darstellung:

V1: $r(\mathbf{X}) = p$, d.h. die *Prädiktorenmatrix hat vollen Spaltenrang*. Diese Voraussetzung ist notwendig, damit $\mathbf{X}'\mathbf{X}$ nichtsingulär ist und die Inverse $(\mathbf{X}'\mathbf{X})^{-1}$) existiert, so daß Gleichung (1.2.39) schätzbar ist[27]. Diese Forderung bedeutet, daß für jedes Paar von Prädiktoren $x_c, x_d (c \neq d)$ gilt: $|r_{cd}| < 1$, d.h. es sollten keine Paare von Prädiktorvariablen in Gleichung (1.2.36) aufgenommen werden, die miteinander vollständig korrelieren bzw. jeder Prädiktor muß einen gewissen Anteil an (von den anderen) unabhängiger Variation aufweisen. Letztere Formulierung enthält auch die Forderung, daß jeder Prädiktor überhaupt Varianz besitzt — eine Voraussetzung, die wegen ihrer Selbstverständlichkeit in Darstellungen der OLS-Schätzung normalerweise nicht eigens postuliert wird.

[26] Es ist allerdings noch nicht der Beweis erbracht, daß mit (1.2.39) tatsächlich ein Minimum und nicht etwa ein Maximum erreicht ist. Die von mir gelesenen Autoren übergehen diesen Beweis (z.B. Hanushek & Jackson 1977, S. 43: „*More complicated differential calculus is required to determine whether this point is a maximum or a minimum ... but that need not concern us here since it will be a minimum for the sum of squared residuals function*“). Man kann leicht zeigen, daß die zweite Ableitung von Gleichung (1.2.38) $2\mathbf{w}$ wird, wenn $\mathbf{w}$ ein $p\times 1$-Vektor ist, der die Diagonalelemente von $\mathbf{X}'\mathbf{X}$ enthält — letztere sind nie negativ, sodaß die 2. Ableitung stets ≥ 0 wird, wodurch der Beweis für ein Minimum erbracht wäre.

[27] Die Möglichkeiten, ein Modell mit singulärer Prädiktorenmatrix zu lösen, übergehe ich hier (vgl. dazu Werner 1993, S. 354ff: Kap. 4.4).

V2: Für die Prädiktoren gilt die *Annahme von „fixen x"* („fixed X" in englischsprachigen Darstellungen): Diese Annahme setzt voraus, daß die Werte der Prädiktoren konstant sind, d.h. bei Replikation der Datenerhebung unter konstanten Bedingungen verändern sich lediglich die Kriteriumswerte. Es sind — mit anderen Worten — nichtstochastische Prädiktoren gefordert, so daß lediglich der Fehlerterm als Quelle zufälliger Variation im Modell enthalten ist. Nicht gefordert ist tatsächliche, praktische Replizierbarkeit der Prädiktorwerte: *„This assumption ... is analogous to the procedure of a physical scientist who repeats a controlled experiment under laboratory conditions several or many times. ... Although in a nonexperimental situation such exact replications of an 'experiment' are impossible, we want to act as if these replications are possible for the sake of discussing the distributions of possible coefficients"* (Hanushek & Jackson 1977, S. 47f). Die Annahme nichtstochastischer Prädiktoren mag allerdings für Anwendungen des OLS-Algorithmus in psychologischen Untersuchungen etwas abenteuerlich anmuten, denn welche psychologische Variable kann schon fehlerfrei gemessen werden? Glücklicherweise zieht der Wegfall dieser Voraussetzung nur geringe Konsequenzen nach sich, insbesondere bleibt die Erwartungstreue des OLS-Schätzers (1.2.39) auch für stochastische Prädiktoren erhalten (vgl. Werner 1993, S. 184: Kap. 3.25.2).

V3: $E(\boldsymbol{\epsilon}) = \mathbf{0}$, d.h. für jede Beobachtung j ist $\bar{\epsilon}_j = 0$, bei den (theoretisch unendlich vielen) Replikationen einer Beobachtung j verteilen sich die Fehlerwerte um den Mittelwert 0 und dieses gilt für jede Beobachtung gleichermaßen.

V4: $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I}$, d.h. bei den (theoretisch unendlich vielen) Replikationen der Beobachtungen haben die Fehler jeder Beobachtung dieselbe Varianz σ^2 (Homoskedastizität) und sind die Fehler der verschiedenen Beobachtungen nicht miteinander korreliert. Mit anderen Worten: Bei jeder Beobachtung wird der Fehler aus derselben Verteilung mit Mittelwert 0 und Varianz σ^2 gezogen und diese Ziehungen sind unabhängig voneinander.

V5: $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, d.h. die Fehler sind normalverteilt.

In diesen Voraussetzungen ist eine weitere wichtige enthalten, auf die häufig zurückgegriffen wird: $\sigma_{\tilde{y}\epsilon} = 0$, d.h. die Kovarianz zwischen Fehler- und systematischer Komponente ist Null bzw. die beiden Komponenten sind unkorreliert.

Die Voraussetzungen **V2** und **V3** garantieren die *Erwartungstreue der OLS-Schätzer* d.h. für den Erwartungswert von $\mathbf{b}$ gilt: $E(\mathbf{b}) = \boldsymbol{\beta}$. Zunächst ist:

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \quad [\text{(1.2.37) in (1.2.39}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \end{aligned}$$

$$= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \tag{1.2.40}$$

Mit (1.2.40) läßt sich nun der Erwartungswert $E(\mathbf{b})$ des Schätzers berechnen:

$$\begin{aligned} E(\mathbf{b}) &= E(\boldsymbol{\beta}) + E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\epsilon}) \quad \left[\text{aus } \mathbf{V2} \text{ folgt: } E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] \\ &= \boldsymbol{\beta} \end{aligned}$$

Zusammen mit **V2** und **V3** garantiert die Voraussetzung **V4** die *„best"-Qualität der OLS-Schätzer*, d.h. ihre minimale Varianz (unter allen linearen, erwartungstreuen Schätzern). Für die Varianz-Kovarianz-Matrix $\boldsymbol{\Sigma}_b$ der Schätzungen erhält man unter diesen Voraussetzungen (vgl. Hanushek & Jackson 1977, S. 119: Formel (5.16) oder Werner 1993, S. 97: Formel (3.13.3)):

$$\boldsymbol{\Sigma}_b = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{1.2.41}$$

Man kann zeigen (vgl. Hanushek & Jackson ebd., S. 120), daß die Varianz-Kovarianz-Matrix eines weiteren erwartungstreuen linearen Schätzers die Form $\sigma^2((\mathbf{X}'\mathbf{X})^{-1} + \mathbf{A})$ haben muß, wobei $\mathbf{A}$ eine $p\times p$-Matrix ist, die nur positive Diagonalelemente enthält — folglich hat $\mathbf{b}$ die geringeste Varianz unter allen Konkurrenten.

Voraussetzung **V5** wird zum Nachweis der BLUE-Eigenschaften nicht benötigt, ist jedoch zur Durchführung von *Signifikanztests* für die geschätzten Koeffizienten nötig, denn aus ihr kann gefolgert werden, daß auch $\mathbf{b}$ normalverteilt ist. Dies ist aus Gleichung (1.2.40) ersichtlich: Dort sind $\boldsymbol{\beta}$ und $\mathbf{X}$ (über alle Replikationen hinweg) feste Größen, die Verteilung von $\mathbf{b}$ hängt von der von $\boldsymbol{\epsilon}$ ab. Folglich gilt: $\mathbf{b} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X}^{-1})$. Zur Prüfung eines Koeffizienten $b_q (q = 1, 2, \ldots, p)$ kann allerdings nicht direkt auf die Normalverteilung zurückgegriffen werden, da ja auch σ^2 nicht direkt beobachtet werden kann und geschätzt werden muß (s. u.). Stattdessen bietet sich folgender t-Test an: $t_{(n-p)} = (b_q - \beta_q)/s\sqrt{a_{qq}}$, wobei s die Schätzung von σ (Standardabweichung der Fehler) und a_{qq} das q-te Diagonalelement von $(\mathbf{X}'\mathbf{X})^{-1}$ ist (zum Nachweis der t-Verteilung des Ausdrucks siehe Hanushek & Jackson 1977, S. 123). Eine F-Statistik zum simultanen Test mehrerer Koeffizienten zeigen Hanushek & Jackson (ebd., S. 124ff) und Werner (1993, S. 99: Kap. 3.15).

Zur *Schätzung der Fehlervarianz* σ^2 werden die Residuen $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ — wobei $\hat{\mathbf{y}} = \mathbf{Xb}$ — herangezogen[28]. Deren um die $n - p$ Freiheitsgrade des Modells (n

[28] ($\mathbf{e}$ ist von $\boldsymbol{\epsilon}$ zu unterscheiden: Es handelt sich bei ersteren um die Differenzen zwischen Kriteriumswerten y_j und deren Vorhersagen $\hat{y}_j$, die sich aus den geschätzten b-Koeffizienten ergeben — weshalb diese Differenzen e_j selbst als Schätzungen der „wahren" Fehler ϵ_j aufzufassen sind, welche nach Gleichung (1.2.37) die Differenzen zwischen Kriteriumswerten und deren Vorhersagen aus der systematischen Komponente mit den „wahren" β-Koeffizienten bilden.

Beobachtungen und p Parameter) „berichtigte“ Varianz gibt eine erwartungstreue Schätzung der Fehlervarianz:

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{p-n} \quad \text{und} \quad E(s^2) = \sigma^2 \tag{1.2.42}$$

Den Beweis der Erwartungstreue von s^2 liefern Hanushek & Jackson (1977, S. 138).

Eine weitere wichtige Statistik ist R^2, der *Determinationskoeffizient* der multiplen Korrelation. Die multiple Korrelation R ist die Korrelation der tatsächlichen Werte y_j mit den aus der Schätzung vorhergesagten $\hat{y}_j$: $R = r_{y\hat{y}}$ und wie die Bezeichnung R^2 zeigt, kann der Determinationskoeffizient einfach als Quadrat der multiplen Korrelation definiert werden (vgl. z.B. die Definition bei Bortz 1979, S. 257). Man könnte aber, zieht man verschiedene Autoren zu Rate, den Eindruck gewinnen, daß hinsichtlich Definition und Interpretation des Determinationskoeffizienten (der OLS-Schätzung) gewisse Unsicherheiten bestehen. So definieren Hanushek & Jackson (1977, S. 212):

$$R^2 = 1 - \frac{\sum e_j^2}{\sum (y_j - \bar{y})^2} = \frac{\sum (\hat{y}_j - \bar{y})^2}{\sum (y_j - \bar{y})^2} \tag{1.2.43}$$

Danach wird $R^2 = 1$, wenn alle Beobachtungen der y-Werte exakt auf der geschätzten Regressionsebene liegen ($\sum e_j^2 = 0$), und $R^2 = 0$, wenn Kriterium und Prädiktoren keine gemeinsame Varianz aufweisen (dann werden $b_1 \ldots b_{p-1} = 0$ und $b_0 = \bar{y}$, folglich $e_j = y_j - \bar{y}$ bzw. $\sum e_j^2 = \sum (y_j - \bar{y})^2$). R^2 kann also Werte zwischen 0 und 1 annehmen und ist dergestalt ein Maß für die Modellanpassung, daß sein Wert in dem Maße abnimmt, wie die Beträge der Residuen zunehmen. Wenn man in Gleichung (1.2.43) Zähler und Nenner der Brüche jeweils durch n teilt, resultiert

$$R^2 = 1 - \frac{s_e^2}{s_y^2} = \frac{s_{\hat{y}}^2}{s_y^2}$$

wobei $s_e^2, s_y^2, s_{\hat{y}}^2$ die Varianzen der jeweiligen Werte in der Stichprobe sind. Hanushek & Jackson interpretieren darum R^2 als den Anteil an der Varianz des Kriteriums, der durch die Regression erklärt wird. Sie leiten diese Interpretation aus folgender Varianzzerlegung her: $s_y^2 = s_{\hat{y}}^2 + s_e^2$. Es irritiert, daß hier die aus der Schätzung resultierenden Stichprobenwerte $\hat{y}_j$ und e_j verwendet werden, denn die Zerlegung der Gesamtvarianz des Kriteriums in „systematische“ und Fehlervarianz setzt die (aus den Voraussetzungen des Modells für die Populationsparameter ableitbare) Unkorreliertheit von und systematischer und Zufallskomponente voraus[29]. Die Frage ist, ob man die vorausgesetzte Unkorreliertheit der beiden Komponenten in der Population einfach auf die aus der

[29] $\sigma_y^2 = \frac{1}{n}\sum (y_j - \bar{y})^2 = \frac{1}{n}\sum (\tilde{y}_j + \epsilon_j - \bar{y})^2 = \frac{1}{n}\sum \left((\tilde{y}_j - \bar{y})^2 + 2(\tilde{y}_j - \bar{y})\epsilon_j + \epsilon_j^2\right) = \sigma_{\tilde{y}}^2 + \sigma_\epsilon^2$, da der Mittelwert $\bar{\epsilon} = 0$ und ebenso die Kovarianz $(\sum (\tilde{y}_j - \bar{y})\epsilon_j)/n$.

Stichprobenschätzung erzielten Werte $\hat{y}_j$ und e_j übertragen kann! Das Problem resultiert wohl daraus, daß Hanushek & Jackson nicht auf den Umstand eingehen, daß auch R^2 als Schätzung eines Populationsparameters aufzufassen ist. Die Definition dieses Parameters ergäbe sich, wenn man in Gleichung (1.2.43) die Stichprobenwerte $\hat{y}_j$ und e_j durch die Populationswerte $\tilde{y}_j$ und ϵ_j ersetzen würde, und für diesen Populationswert wäre dann die auf der Varianzzerlegung beruhende Interpretation unproblematisch. Auf die Qualität von R^2 als Schätzer geht Werner (1993, S. 82f: Kap. 3.8) ein: R^2 ist nicht erwartungstreu, es werden dort Formeln für erwartungstreue Schätzungen angegeben.

Jedenfalls wird der Determinationskoeffizient üblicherweise als Maß für den „Wert“ eines Modells aufgefäßt — z.B. mit den Worten von Hanushek & Jackson (1977, S. 121): *“First, if we consider a naive estimate of any behavioral variable to be its mean, which is our best guess at the population mean, then R^2 gives a comparison of how well our ‘sophisticated’ estimate relying on the values of a set of different variables X_k* [$x_1 \ldots x_{p-1}$ in meiner Notation] *does in comparison to the mean“.*

In Anbetracht der oben versprochenen Kürze der Darstellung der OLS-Regression beende ich damit dieselbe — sie ist noch lange nicht vollständig, es sei hier auf die angegebene Literatur verwiesen — und wende den Blick auf Besonderheiten, die sich bei der **Anwendung des Algorithmus für die CA** ergeben. Das Grundprinzip dieser Anwendung wurde bereits in Abschnitt 1.1.2 mit Gleichung (1.1.2) beschrieben (S. 18): Als Kriteriumswerte y_j fungieren hier die Gesamtnutzenwerte der Stimuli $1, 2, \ldots, n$, als Prädiktorvariablen dichotome Kodiervariablen für die Stufen der Attribute, die lediglich die Zuordnung der Stufen zu den Stimuli anzeigen. Es sind hierbei 2 besondere Umstände zu berücksichtigen:

(1) In der vorne schon mehrmals als „typisch“ bezeichneten CA-Untersuchung werden nicht solche Gesamtnutzenurteile über die multiattributiven Stimuli erhoben, die als intervallskaliert zu interpretieren sind, sondern ordinale Rangwerte über die Stimuli. Genaugenommen dürften solche Werte nicht als Kriteriumswerte einer OLS-Regression eingesetzt werden, denn der Algorithmus erfordert metrische y_j. Dieses geht schon aus der Interpretation der β-Gewichte als Partialregressionskoeffizienten hervor: Die Rede davon, daß ein Koeffizient angibt, um wieviel Einheiten sich der Kriteriumswert verändert, wenn die Prädiktorvariable um eine Einheit nach oben fortgeschaltet wird, macht keinen Sinn für Kriteriumswerte mit lediglich ordinalem Skalenniveau. Es wurde bereits darauf hingewiesen (S. 23), daß es in praktischen Anwendungen der CA durchaus üblich ist, trotzdem OLS-Analysen auf der Basis von Rangwerten durchzuführen. Dies geschieht gewissermaßen im Vertrauen darauf, daß die Rangwerte annähernd metrische Qualitäten besitzen, daß also die

numerischen Abstände zwischen den Rangwerten auch annähernd die „subjektiven Abstände“ in der Präferenz wiedergeben. Es wurde auch schon erwähnt, daß die Rechtfertigung für dieses Vorgehen aus Simulationsstudien — diese werden in Abschnitt 2.1.1 dargestellt werden — gezogen wird, und daß es sich bei der Untersuchung, die im folgenden präsentiert werden soll, um eine solche handelt.

(2) Bei der Verwendung von Dummy-Variablen als Prädiktoren sind die geschätzten Koeffizienten nicht einfach als Schätzungen der Teilnutzenwerte des zugehörigen Attributlevels (inhaltlich) interpretierbar: Wird eine Dummy-Kodierung im engeren Sinne verwendet[30], so werden für ein Attribut l mit s Stufen nur $s-1$ Dummy-Variablen gebildet. Hat z.B. ein Attribut 3 Stufen, so erhielte man, wenn man für jede Ausprägung eine Kodiervariable verwendete:

	x_{j1}	x_{j2}	x_{j3}
Ausprägung 1:	1	0	0
Ausprägung 2:	0	1	0
Ausprägung 3:	0	0	1

Damit aber wäre die Voraussetzung **V1** des OLS-Modells verletzt, denn es bestünde eine perfekte lineare Abhängigkeit zwischen den Prädiktoren (nimmt man noch $x_0 = 1$ hinzu, so ist z.B. $\mathrm{x}_{j3} = x_0 - x_{j1} - x_{j2}$), der Wert jeweils einer Spalte ist aus den beiden anderen vorhersagbar bzw. eine Spalte ist redundant. Man wäre also gezwungen z.B. x_3 wegzulassen[31] und damit wäre Ausprägung 3 in der Prädiktorenmatrix immer noch eindeutig identifiziert durch diejenige Zeile, in der beide verbliebenen Dummy-Variablen den Wert 0 annehmen. Durch das zu diesem Beispiel gehörige OLS-Modell (sozusagen eine 3^1-Struktur „multi“-attributiver Stimuli mit nur einem Attribut) — $\hat{y}_j = b_0 + b_1 x_{j1} + b_2 x_{j2}$ — erhält der Stimulus, bei dem das Attribut in Ausprägung 3 vorliegt, als geschätzten („Gesamt“-)Nutzenwert b_0. Liegt Ausprägung 2 vor, dann wird $\hat{y}_j = b_0 + b_2$, und bei Ausprägung 1 wird $\hat{y}_j = b_0 + b_1$. Ausprägung 3 fungiert hier also als eine Art Referenzstufe, die Koeffizienten b_1 und b_2 schätzen, wie sich der („Gesamt“-)Nutzenwert gegenüber dem Stimulus mit Ausprägung 3 verändert, wenn Ausprägung 1 bzw. 2 vorliegt: Die Koeffizienten sind nicht ohne weiteres als Teilnutzenwerte im Sinne des CA-Modells (vgl. Gleichung (1.1.1), S. 18) zu interpretieren, denn diese Teilnutzenwerte enthalten den gesamten (additiven) Nutzenbeitrag einer

[30] vgl. Werner (1993, S. 195: Kap. 4.3.1), der auf die uneinheitliche Verwendung des Begriffs der Dummy-Kodierung hinweist: Einmal in einem allgemeineren Sinne für verschiedene Formen der Kodierung kategorialer Prädiktoren, ein andermal im engeren Sinne für diejenige dieser Kodierformen, bei der die Variable nur in 2 Zuständen (0 und 1) auftritt.

[31] Ich übergehe wieder die Möglichkeiten der Lösung singulärer Prädiktorenmatrizen, vgl. Fußnote 27.

Attributstufe zum Gesamtnutzenwert. Im Beispiel mit nur einem Attribut könnte b_0 noch als Schätzung des entsprechenden β in Gleichung (1.1.1) aufgefaßt werden, bei mehreren Attributen müßte aber für jedes Attribut eine Referenzstufe festgelegt werden und b_0 enthielte den kumulierten Nutzenbeitrag all dieser Referenzstufen — der Nutzenbeitrag jedes einzelnen der Referenzlevels wäre auch indirekt (über die geschätzten Koeffizienten) nicht mehr zu ermitteln.
Einen einfachen Weg um zu solchen Koeffizienten zu gelangen, aus denen sich alle Teilnutzenwerte zumindest indirekt ableiten lassen, bietet die Effektkodierung: Hier werden Koeffizienten als Effekte im Sinne der ANOVA geschätzt, d.h. der Koeffizient (Effekt) jeder Stufe einer kategorialen Prädiktorvariablen gibt die Abweichung des Mittelwerts der Kriteriumswerte auf dieser Stufe vom Gesamtmittelwert aller Kriteriumswerte wieder. Es muß über alle Stufen einer Prädiktorvariablen hinweg die Summe dieser Abweichungen — d.h. die Summe aller Koeffizienten dieser Stufe — Null werden (dies ist die sogenannte Σ-Restriktion). In der Kodierung führt dies dazu, daß in allen Spalten der Prädiktorenmatrix (außer der ersten für die Konstante) die Spaltensumme Null sein muß. Am einfachsten ist dies dadurch zu erreichen, daß, ausgehend von der Dummy-Kodiermatrix, für jede katogoriale Variable (für jedes Attribut im Falle der CA) in den Zeilen, die die Referenzstufe repräsentieren, die Nullen (in den zur kategorialen Variablen gehörenden Dummy-Variablen) durch -1 ersetzt werden. Im obigen Beispiel eines dreistufigen Attributs erhielte man so die folgende Kodierung (Ausprägung 3 wieder als Referenzstufe):

	x_{j1}	x_{j2}
Ausprägung 1:	1	0
Ausprägung 2:	0	1
Ausprägung 3:	-1	-1

Aufgrund der Σ-Restriktion kann der Wert des Koeffizienten der Referenzstufe jetzt indirekt aus den Koeffizienten der anderen Stufen des Attributs berechnet werden: $b_3 = -b_1 - b_2$ — bei mehreren Attributen ebenso die Werte der Koeffizienten von deren Referenzstufen. Die so ermittelten Koeffizienten enthalten insofern den gesamten additiven Nutzenbeitrag einer Attributstufe, als sie die Abweichung vom Mittelwert der Gesamtnutzenwerte — d. i. b_0 bei der Effektkodierung — über alle Stimuli des Modells, der aus der jeweiligen Stufe resultiert, angeben (solange man nicht annimmt, daß der „Nutzen“ einen echten Nullpunkt besitzt, ist die Bedeutung der Konstanten nebensächlich, da jede Gesamtnutzenskala so transformiert werden kann, daß $b_0 = 0$) [32].

[32]Merkwürdigerweise habe ich in der Literatur zur CA keine Erwähnung dieser mit der Kodierung verbundenen Probleme der Ableitung von Teilnutzenwerten aus den Koeffizien-

Die Betrachtung der theoretischen Grundlagen der nachfolgend beschriebenen Untersuchung kann damit wohl abgeschlossen werden. Es sollten sozusagen die beiden Teile, aus denen der Begriff CA zusammengesetzt ist, näher beleuchtet werden: Was hat es mit dem „Conjoint“ auf sich und von welcher „Analyse“ ist die Rede? Angesichts der Vielzahl theoretischer Probleme und methodologischer Bereiche, die solch ein Überblick streifen muß, kann er natürlich nicht an jeder Stelle den Gegenstand erschöpfend behandeln und muß oft eine gewisse Oberflächlichkeit bewahren — zur Klärung offener Fragen sollte aber die jeweils angegebene Literatur ausreichen.

ten gefunden, weshalb ich sie hier ausdrücklich behandelt habe. Dies mag daran liegen, daß das Interesse meist weniger an den Koeffizienten und ganz an der Vorhersage der Gesamtnutzenwerte liegt, weshalb die Interpretation der Koeffizienten im Grunde keine Rolle spielt, solange nur die Präferenzen für bestimmte Kombinationen von Ausprägungen zuverlässig vorhergesagt werden können.

Teil 2: Untersuchung

2.1 Planung und Durchführung

Im Teil 1 (Abschnitt 1.1.2) wurde eine Auswahl von Verfahren zur CA präsentiert und es wurde grundsätzlich zwischen metrischen und nichtmetrischen Verfahren unterschieden (S. 23). Einem Untersucher, der eine CA durchführen will, stellt sich natürlich die Frage, welches dieser Verfahren er anwenden soll, welches unter welchen Bedingungen die zuverlässigsten Ergebnisse liefert. Insbesondere wird er zu erwägen haben, ob er die metrische oder die nichtmetrische Analyse wählen soll.

Die metrische Analyse wird sich vielleicht oft schon deswegen anbieten, weil die metrischen Verfahren im allgemeinen (einem sozialwissenschaftlichen Experimentator) schon bekannt — und damit schon verstanden — sind. Hinzu kommt der geringere Rechenaufwand dieser Verfahren gegenüber den nichtmetrischen, bei denen es sich, wie gezeigt, um iterative Prozeduren handelt[33]. Und v.a.: Es existieren für die nichtmetrischen Verfahren keine Fehlertheorien! Dieses geht aus den in den Abschnitten 1.1.2 und 1.2.1 wiedergegeben Modellen hervor, deren Formulierungen keine Fehlerterme enthalten — es sind folglich weder für geschätzten Parameter (Teilnutzenwerte), noch für die verschiedenen Maße für die Modellanpassung (Streß, C^*) Signifikanzaussagen möglich[34].

Dagegen liegt der Nachteil der metrischen Verfahren gegenüber den nichtmetrischen darin, daß in der typischen CA ordinale Präferenzurteile der Versuchsperson(en) die Datenbasis bilden, während die metrischen Verfahren das Inter-

[33]Der Umstand kann schwerer wiegen, als es auf den ersten Blick erscheint: LINMAP wurde im Verlauf der Untersuchung auf mehreren (IBM-kompatiblen) Microcomputern gerechnet — mit erheblichen Unterschieden im Zeitaufwand je nach Ausstattung des Computers. Ein AT mit 80286 Prozessor und 12MHz Taktfrequenz benötigte bisweilen zur Analyse eines vollen Designs einer 3^4-Struktur (mit 81 Rangwerten als abhängiger Variable) über 2 Stunden, während diesselbe Analyse bei einem 80486-Prozessor mit Math-Coprozessor ungefähr eine Minute dauerte.

[34]Es existiert ein Ansatz zur Ermöglichung solcher Signifikanzaussagen: Mullet & Karson (1986) simulieren die Häufigkeitsverteilung des LINMAP-Index C^*, die sich aus den LINMAP-Analysen von jeweils 2000 zufällig generierten Rangordnungen ergibt, sie tun dieses für 16 verschiedene multiattributive Designs. Erzielt man also mit einem dieser Designs ein bestimmtes C^*, so könnte man es unter Zugrundelegung dieser Verteilungen gegen die Nullhypothese einer „zufälligen“ Präferenzordnung testen. Umesh & Mishra (1990) simulieren solche Verteilungen für C^*, für den MONANOVA-Streß und für R^2 der OLS-Regression. Sie simulieren darüber hinaus die Häufigkeitsverteilungen dieser Werte auch für nichtzufällige Daten mit verschieden großen Fehlerkomponenten, es sollen auf der Basis von Vergleichen eines in der Analyse erzielten Fit-Wertes mit den Prozenträngen, die dieser Wert in diesen verschiedenen Verteilungen innehat, Aussagen über die Güte der Modellanpassung, die er signalisiert, möglich werden (über bloßen Signifikanztest gegen das Zufallsmodell hinaus).

vallskalenniveau der abhängigen Variablen voraussetzen: Werden die ordinalen Rangdaten einer metrischen Analyse unterzogen, so wird von ihnen möglicherweise ein falscher Gebrauch gemacht (vgl. S. 23f). Die Frage ist, wie schwer dieser Nachteil wiegt! Man könnte ja argumentieren, daß empirisch erhobene Rangwerte ungefähr auch den „subjektiven Abstand“ in den Präferenzen des Individuums wiederspiegeln: Hat z.B. eine Versuchsperson 5 Stimuli in aufsteigender Folge nach ihrer Präferenz geordnet, so könnte man theoretisch diese Folge auf einer Ordinalskala durch die numerischen Rangwerte 1, 6, 34, 100, 20000 erfassen (bzw. jede andere monoton steigende Zahlenfolge mittels der für Ordinalskalen zulässigen Transformationen in diese Werte überführen), aber die Rangwerte, von denen hier die ganze Zeit die Rede ist, meinen natürlich die Zahlenfolge 1, 2, 3, 4, 5 — und für diese Folge mag das Argument gelten, daß sie sozusagen quasi-metrische Qualitäten besitze, daß also der zweite Stimulus gegenüber dem ersten ungefähr genauso „viel“ bevorzugt werde, wie der dritte gegenüber dem zweiten usw.. Die Stichhaltigkeit dieses Arguments kann nicht prinzipiell beurteilt werden, es sind inter- und intraindividuelle Unterschiede in subjektiven Präfernzstrukturen denkbar, durch die es im einen Falle zutrifft und im anderen die Realität grob verfälscht. Jedenfalls sollte, wer ordinale Rangwerte einer metrischen Analyse unterzieht, zeigen können, daß das Argument der quasi-metrischen Qualität der Rangdaten unter Umständen zutrifft und daß es im konkreten Fall plausibel ist, diese besonderen Umstände für gegeben zu halten.

Überhaupt muß man, wenn mehrere Verfahren zur Auswahl stehen, fragen, welches dieser Verfahren unter welchen Umständen die „besseren“ Ergebnisse bringt, bzw. muß, wenn man eines dieser Verfahren verwendet, zeigen können, daß dieses — im Großen und Ganzen und im Vergleich zu den konkurrierenden Methoden — unter den gegebenen Umständen zufriedenstellende Ergebnisse liefern kann. Die Umstände, an die hierbei zu denken ist, betreffen psychologische Aspekte der Urteilsbildung der Versuchsperson. Entschließt man sich zur Datenerhebung mittels der Profilmethode und zur Rangreihung der Stimulusprofile, so sind u.a. folgende Fragen zu stellen:

- Wie „gut“ schätzt ein Verfahren bei wachsendem Fehler der Beurteilungen, bzw. wie weit hängt die Güte der Ergebnisse eines Verfahrens vom Ausmaß des Fehlers ab?

- Wenn die Versuchsperson zuviele Stimuli rangreihen muß, so ist eine gewisse Überforderung zu befürchten, durch welche die Fehlerhaftigkeit der Urteile erhöht würde. Greift man deswegen auf reduzierte faktorielle Designs zurück: Leidet die Güte der Ergebnisse darunter, d.h. unter der reduzierten Anzahl von Beobachtungen (und damit unter weniger Freiheitsgraden)?

- Wie sensibel reagiert ein Verfahren auf Verletzungen der additiven Kompositionsregel? In der CA wird, wie schon gezeigt wurde und wie es in der

Modellgleichung (1.1.1) zum Ausdruck kommt, normalerweise ein rein additiver Zusammenhang der Teilnutzenwerte vorausgesetzt. In vielen praktischen Anwendungen, oft in der Marktforschung, verbieten schon ökonomische Gründe die Testung der Kompositionsregel, so daß meist darauf vertraut wird, daß auch bei leichten Verletzungen der Additivität, die bei der einen und anderen Versuchsperson denkbar sind — v.a. durch Interaktionen —, noch eine gute Vorhersage der tatsächlichen Präferenzen mittels der Ergebnisse eines „additiven" Verfahrens möglich ist.

Zur Klärung dieser Fragen bieten sich Simulationsstudien an, bei denen die genannten Umstände bzw. Bedingungen bei der Generierung von Rangwerten simuliert werden: Es kann so auf quasi empirischem Wege erforscht werden, wie verläßlich die Ergebnisse sind, die durch die verschiedenen Verfahren unter diesen Bedingungen erzielt werden.

2.1.1 Simulationstudien

Im folgenden werden 3 Untersuchungen dargestellt, die die zuletzt genannte Zielsetzung verfolgten. Sowohl hinsichtlich der Vorgehensweise bei der Generierung multiattributiver Präferenzurteile, als auch hinsichtlich der Auswahl experimentell variierter Bedingungen der CA, als auch hinsichtlich der Auswertung der erzielten Ergebnisse waren diese Arbeiten richtungsweisend für die eigene Untersuchungsplanung, Aus der — u.a. auch kritischen — Betrachtung dieser Studien resultieren bis zu einem gewissen Maße Vorgehensweise und Hypothesen meiner Untersuchung, so daß eine genauere Betrachtung dieser Arbeiten hier unumgänglich ist.

Die Untersuchung von Carmone, Green & Jain (1978):

Die Autoren untersuchen eine 3^5-Struktur — also 5 Attribute mit jeweils 3 Stufen — , sie verwenden dazu einen Satz von Teilnutzenwerten, den sie aus einer anderen Studie mit empirischen Rangdaten gewonnen haben. Auf der Basis dieser Werte wurden neue „Präferenzurteile" generiert, und zwar unter Variation folgender experimenteller Bedingungen (unabhängiger Variablen):

A: Typus der „wahren" Kompositionsregel: Rein additives Modell vs. interaktives Modell (additive Haupteffekte der Attribute plus einem Interaktionsterm).

B: Fehler: deterministisches Modell (kein Fehler) vs. stochastisches Modell (moderater Fehler).

C: Anzahl der Beobachtungen für die Parameterschätzung: 243 Beobachtungen (volles faktorielles Design) vs. 54 Beobachtungen vs. 27 Beobachtungen vs. 18 Beobachtungen (reduzierte faktorielle Designs).

D: Typus der Präferenzurteile: Rangreihung vs. Ratingskala vs. „Rohdaten“ (s.u.).

E: Typus der Analyse: Metrisch (ANOVA) vs. nichtmetrisch (MONANOVA).

Aus A×B×C×D×E resultieren also $2 \times 2 \times 4 \times 3 \times 2 = 96$ Zellen des experimentellen Designs, sie wurden in den folgenden Schritten realisiert:

- Zunächst wurde für das additive Modell der fehlerfreie Gesamtnutzenwert $\tilde{y}_j$ eines jeden Stimulus $j (j = 1, \ldots, 243)$ durch simple Addition der 5 Teilnutzenwerte der ihm zugehörigen Attributstufen berechnet: $\tilde{y} = \sum_l \sum_k \beta_{lk} x_{lk}$ (vgl. Gleichung (1.1.1), S. 18). Für das interaktive Modell wurde dabei zusätzlich zu $\sum_l \sum_k \beta_{lk} x_{lk}$ noch $\frac{1}{5}$ des Produkts der jeweiligen 5 Teilnutzenwerte addiert. Es wurde mit dem interaktiven Modell also eine Modellmisspezifikation erzeugt, da der Interaktionsterm ja nicht im additiven Schätzmodell der CA enthalten ist.

- Zu den resultierenden 2 Sätzen von jeweils 243 $(= 3^5)$ „deterministischen“ Gesamtnutzenwerten wurde für die Bedingung ohne Fehler nichts hinzuaddiert $(\tilde{y}_j = y_j)$, für die Fehlerbedingung wurde (bei jedem der beiden Datensätze) zunächst die Standardabweichung $\sigma_{\tilde{y}}$ aller 243 Werte $\tilde{y}_j$ berechnet, aus der Normalverteilung mit einer Standardabweichung von jeweils $\frac{1}{5}\sigma_{\tilde{y}}$ und dem Mittelwert Null wurden dann für jeden Stimulus Fehlerwerte ϵ_j gezogen und zu den fehlerfreien Gesamtnutzenwerten addiert $(y_j = \tilde{y}_j + \epsilon_j)$.

- Aus den so gewonnen 4 Sätzen von Gesamtnutzenwerten wurden im nächsten Schritt 16 neue Datensätze gebildet, indem unter Rückgriff auf reduzierte faktorielle Designs jeweils aus der Gesamtmenge der 243 Stimuli Teilmengen ausgewählt und deren Gesamtnutzenwerte y_j in die neuen Datensätze aufgenommen wurden (für die vollen faktoriellen Designs wurden natürlich die 4 zuvor gewonnenen Datensätze ganz übernommen).

- Die Werte in den daraus resultierenden 16 Datensätzen wurden dann einmal in Rangwerte umgewandelt, einmal in Werte auf einer 6-stufigen Ratingskala transformiert und einmal überhaupt nicht verändert („Rohdaten“).

- Das Ergebnis sind 48 Datensätze, die jeweils einmal mit MONANOVA und einmal mit ANOVA analysiert wurden — es wurden also 96 Analysen durchgeführt.

Zur Messung der „Güte“ der Analyseergebnisse berechnen die Autoren die Tau-Korrelationen zwischen den „wahren“ Teilnutzenwerten, die die Basis der Datengenerierung bildeten, und den jeweiligen Schätzungen dieser Parameter,

wie sie sich aus den 96 Analysen ergaben. Die Korrelationen wurden in Fisher-Z-Werte transformiert (vgl. Bortz 1979, S. 260) und als abhängige Variable einer Varianzanalyse mit A bis E als unabhängigen Variablen unterzogen. Die (von den Autoren mitgeteilten) relevanten Ergebnisse dieser Varianzanalyse: Alle 3 Interaktionen ersten Grades zwischen den Variablen A, B und D waren signifikant auf dem 1-Prozent-Level und weder der Haupteffekt der Variablen C, noch der der Variablen E war signifikant.

Für die signifikanten Interaktionen gilt im einzelnen:
1. A×D: Unter dem interaktiven Modell gab es zwischen den 3 Typen von Präferenzdaten keine nennenswerten Differenzen zwischen den Korralationen, während unter dem additiven Modell diese Differenzen bedeutsam waren, und zwar erzielten hier die „Rohdaten“ die höchste Korrelation, gefolgt von den Rangwerten, während die Ratings am schlechtesten abschnitten.
2. A×B: Auch hier ergaben sich unter dem interaktiven Modell keine bedeutsamen Differenzen zwischen der fehlerfreien und der Fehlerbedingung, während unter dem additiven Modell eine moderate Verschlechterung der Korrelation mit dem Fehler einhergeht.
3. B×D: In der Fehlerbedingung gab es keine Differenzen zwischen den 3 Typen von Präferenzdaten, dagegen waren solche Differenzen bei den fehlerfreien Daten bedeutsam (mit demselben Trend wie bei der A×D-Interaktion).

Insgesamt jedoch lassen in dieser Veröffentlichung die sehr knappen Mitteilungen der Resultate einiges zu wünschen übrig. So heißt es von den Haupteffekten der Variablen A, B und D nur: *„Not relevant as interaction term has already been found significant“* (ebd., S. 302). Nun beträgt aber der mittlere Tau-Koeffizient aller Zellen mit additivem Modell .945, dagegen der aller Zellen mit interaktivem Modell .639 — eine so beträchtlichen Differenz, daß es ein gewisses Unbehagen hinterläßt, wenn sie von den Autoren einfach nicht weiter berücksichtigt wird. Berechnet man mithilfe der (immerhin mitgeteilten) Quadratsummen und Freiheitsgrade die F-Statistiken der Haupteffekte A, B und D, so erweisen sich diese Effekte als signifikant auf dem 1-Prozent-Niveau[35] — wobei allerdings die tatsächlichen Differenzen der mittleren Tau-Koeffizienten zwischen den Stufen von B bzw. D allesamt sehr gering ($> .15$) ausfallen.

Man kann aus den Ergebnissen dieser Untersuchung 3 Schlußfolgerungen ableiten:

(**1**) ANOVA und MONANOVA produzieren unter verschiedenen Umständen Ergebnisse von derselben Güte (weder signifikanter Haupteffekt von E,

[35] Wenn man die mitgeteilte Residualquadratsumme als Prüfgröße verwendet, was aber — da es sich hier um eine ANOVA mit nur einer Untersuchungseinheit (Tau-Korrelation) pro Variablenstufenkombination handelt, bei der in der Residualquadratsumme sowohl Varianzanteile der Fehlerkomponente, als auch der Interaktion 4. Grades enthalten sein können — genaugenommen ohne weitere Information auch nicht ganz korrekt ist.

noch irgendeine Interaktion mit den anderen Variablen).

(**2**) Auch der Rückgriff auf reduzierte faktorielle Designs mit sehr wenigen Beobachtungen vermindert nicht die Güte der Analyseergebnisse in nennenswerter Weise (kein signifikanter Haupteffekt von C, keine Interaktion mit den anderen Variablen).

(**3**) Bei substantiellem Fehler des Schätzmodells — sei es durch Modellmisspezifikation (interaktives Modell), sei es durch einen „echten“ Zufallsfehler, der zu den Urteilen beiträgt, — liefern durch Ratingskalen erhobene Präferenzurteile Ergebnisse von derselben Güte, wie Rangwerte (Interaktionen A×D und B×D).

Bleibt die Frage nach dem Stellenwert dieser Ergebnisse. Denn die Untersuchung weißt einen grundsätzlichen Mangel auf, der Zweifel an der Bedeutung der Resultate nahelegt: Es wurde nur ein einziger Fall generiert, d.h. es wurde nur ein Satz von Teilnutzenwerten „gezogen“ und entsprechend der experimentellen Bedingungen zu den verschiedenen Sätzen von Präferenzdaten verrechnet — es befindet sich in jeder Zelle des Untersuchungsdesigns nur ein Fall, sozusagen nur eine einzige simulierte Versuchsperson mit Präferenzurteilen, und folglich wird pro Zelle nur eine einzige Analyse gerechnet und geht nur eine einzige Tau-Korrelation in die abschließende Varianzanalyse ein. Man darf sich also nicht von den aus dem 3^5-Design resultierenen 243 Beobachtungen, welche (unter der Bedingung eines vollen faktoriellen Designs) für die einzelne CA zur Verfügung stehen, täuschen lassen: Die tatsächliche Stichprobengröße in diesem Untersuchungsdesign beträgt 1. Es ist darum sehr fraglich, ob aus den Ergebnissen, die auf der Basis dieses einen experimentell variierten Falles gewonnen wurden, überhaupt allgemeine Aussagen über die Güte verschiedener CA-Methoden unter den getesteten Bedingungen abgeleitet werden können. Als repräsentative Stichprobe (für die Grundgesamtheit möglicher Sätze von Teilnutzenwerten) wird man den einen Satz von 15 Teilnutzenwerten, der die Datenbasis der Untersuchung bildet, jedenfalls kaum akzeptieren können — zumal er noch nicht einmal zufällig gezogen wurde —, so daß die Ergebnisse dieser Untersuchung wohl allenfalls explorativen oder demonstrativen Wert besitzen.

Ein weiterer Kritikpunkt wird angesichts des grundsätzlicheren ersten fast nebensächlich: Die von den Autoren generierte Interaktionsbedingung ist von zweifelhaftem Wert für die Praxis der CA. Ein gleichzeitiges interagieren der Attribute in der höchsten Ordnung wird man wohl in den seltensten Fällen erwarten, „realistischer“ wären da einzelne Interaktionen 1. Ordnung. Die Interaktion 4. Ordnung hat — wie gezeigt — die Güte der Ergebnisse deutlich reduziert, die interessantere Frage wäre, ob und in welchem Ausmaß selbiges auch durch Interaktionen niederer Ordnung bewirkt wird.

Die Untersuchung von Cattin & Bliemel (1978):

Die Autoren vergleichen OLS und MONANOVA. Sie gehen dabei von der Hypothese aus, daß MONANOVA bei fehlerfreien Daten (Präferenzurteilen) OLS übertrifft, und zwar sowohl bei Rangdaten, als auch bei Ratings, während OLS sowohl bei fehlerbehafteten Ratings MONANOVA übertrifft, als auch bei Rangwerten dann, wenn das datengenerierende Modell kompensatorisch ist[36]. Es werden in der Simulation die folgenden Bedingungen experimentell variiert:

A: Anzahl der Attribute: 4 vs. 9.

B: Fehler: deterministisches Modell vs. stochastisches Modell.

C: Typus der Präferenzurteile: Rangreihung vs. Ratingskalen.

D: Tendenz der Beurteilung auf der Ratingskala: „fair judge“ vs. „exaggerating judge“ vs. „indecisive judge“ (s. u.).

Es wurden für jede Zelle des Untersuchungsdesigns 50 Fälle generiert. Die Datengenerierung wird um der besseren Überschaubarkeit willen im folgender Abbildung dargestellt:

<table>
<tr><td>1</td><td colspan="3">Setze $\beta_{l1} = 0$ und ziehe β_{l2} aus $N(0,1)$
für jedes Attribut $l = 1, \ldots, 4$ | für jedes Attribut $l = 1, \ldots, 9$</td></tr>
<tr><td>2</td><td colspan="3">Ermittle $\tilde{y}_j$ nach Gleichung (1.1.1), S. 18: $\tilde{y}_j = \sum_{l=1}^{t} \sum_{k=1}^{2} \beta_{lk} x_{lk}$
($t = 16$, reduziertes faktorielles Design für die 2^9-Struktur)</td></tr>
<tr><td>3</td><td colspan="2">Ziehe für jedes $\tilde{y}_j$ einen Fehlerwert e_j aus $N(0, 0.75)$</td><td>Setze Fehlerwerte $e_j = 0$</td></tr>
<tr><td colspan="4">Berechne Gesamtnutzenwert $y_j = \tilde{y}_j + e_j$</td></tr>
<tr><td>4</td><td>Wandle die y_j in Rangwerte von 1 bis 16</td><td colspan="2">Transformiere die y_j in $a_j = 7\frac{y_j - y_{min}}{y_{max} - y_{min}}$ (wobei y_{min} und y_{max} Minimum und Maximum der 16 Werte y_j sind).
Bilde dann aus den a_j die (Rating-)Skalenwerte r_j vermittels der Transformation $f(b)$ = der nächsthöhere ganzzahlige Wert ($\neq 0$) zu b:</td></tr>
<tr><td></td><td>$r_j = f(a_j)$
(fair judge)</td><td>$r_j = f(-\frac{2}{49}a_j^3 + \frac{3}{7}a_j^2)$
(exaggerating judge)</td><td>$r_j = f(-\frac{4}{49}a_j^3 + \frac{6}{7}a_j^2 + 3a_j)$
(indecisive judge)</td></tr>
</table>

[36] „*For a compensatory model, the influence of an attribute with an unfavorable level tends to be offset by the influence of another attribute with a favourable level*“ (Wittink & Cattin 1981, S. 102). Kompensatorisch ist ein Modell der Datengenerierung praktisch dann, wenn es additiv (oder multiplikativ) ist und alle Attribute ungefähr dieselbe Wichtigkeit besitzen.

Der Kasten 1 zeigt die Ziehung von Teilnutzenwerten: Es wurden durchweg zweistufige Attribute verwendet, und zwar gemäß Variable A einmal 4 (2^4-Struktur) und einmal 9 (2^9-Struktur) Attribute. Für jedes Attribut wurden die Teilnutzenwerte der beiden Stufen ermittelt, indem jeweils ein Wert gleich Null gesetzt und der andere aus der Normalverteilung mit Mittelwert 0 und Standardabweichung 1 gezogen wurde — es mußten also in der 2^4-Bedingung 4, in der 2^9-Bedingung 9 Werte (pro Fall) gezogen werden. Der Kasten 2 enthält die Errechnung von „deterministischen“ Gesamtnutzenwerten ($\tilde{y}_j$ — die Benennung folgt derjenigen in Abschnitt 1.2.2) aus den Teilnutzenwerten nach dem additiven Modell. Es wurden sowohl für die 2^4-, als auch für die 2^9-Struktur jeweils nur 16 Gesamtnutzenwerte berechnet, bei ersterer besteht das volle faktorielle Design aus 16 Stimuli, bei letzterer wurden die Werte für ausgewählte Stimuli eines reduzierten faktoriellen Designs ermittelt. Kasten 3 gibt die Erzeugung der Fehler (Variable B) wieder: Die Fehlerkomponenten für jeden der 16 Stimuli in jedem der beiden faktoriellen Designs wurden aus der Normalverteilung mit Mittelwert 0 und Standardabweichung 0.75 gezogen, dies entspricht hier bei 4 Attributen 25%, bei 9 Attributen 11% Fehlervarianz. Schließlich werden — siehe Kasten 4 — gemäß Variable C die Präferenzwerte y_j einmal in Rangwerte von 1 bis 16, ein andermal in Werte auf der Ratingskala umgewandelt — letzteres gemäß D auf dreifach verschiedene Weise: Zunächst werden die Präferenzwerte y_j linear in den Wertebereich von 0 bis 7 transformiert (so daß $y_{min} = 0$ und $y_{max} = 7$). Dann werden diese transformierten Werte in der Bedingung „fair judge“ in die nächsthöheren ganzzahligen Werte der 7-Punkte-Skala umgewandelt, was einer unverrzerrten Umwandlung der metrischen Präferenzwerte in Ratings entspricht. In der Bedingung „exaggerating judge“ werden die transformierten Werte vor der Umwandlung in ganzzahlige Skalenwerte so weitertransformiert, daß sich eine Tendenz zu Extremantworten ergibt, d.h. die Ratings für die 16 Stimuli versammeln sich mehrheitlich an den Rändern der 7-Punkte-Skala. In der Bedingung „indecisive judge“ schließlich wurde eine Antworttendenz zur Mitte erzeugt, hier fallen die meisten Ratings in den mittleren Bereich der 7-Punkte-Skala.

Das experimentelle Design hat also 16 Zellen ($\mathsf{A}\times\mathsf{B}\times\mathsf{C},\mathsf{D}= 2 \times 2 \times 4$) und jede Zelle enthält 50 Fälle[37]. Jeder Fall in jeder Zelle wurde einmal mit OLS und einmal mit MONANOVA analysiert, so daß insgesamt 1600 Analysen resultieren.

Zur Auswertung definieren die Autoren einen Wert, den sie „$SAVE$“ — für „*s*um of the *a*bsolute *v*alues of *e*rrors“ — nennen. Dazu müssen zunächst die „wahren“ Teilnutzenwerte β_{l2} und deren Schätzungen aus der OLS-Regression

[37]Es geht aus dem Text nicht genau hervor, ob für jede Zelle des Designs jeweils 50 Fälle neu generiert wurden, oder ob nur nur einmal 50 Sätze von Teilnutzenwerten gezogen wurden, aus denen dann für jede Zelle entsprechend der Kombination experimenteller Bedingungen die 50 Sätze von Präferenzurteilen ermittelt wurden.

b_{l2}^O und aus MONANOVA b_{l2}^M standardisiert werden[38]. Dies geschieht unter der Bedingung:

$$\sum_{l=1}^{t} |\beta_{l2}| = 1 \; , \; \sum_{l=1}^{t} |b_{l2}^O| = 1 \; , \; \sum_{l=1}^{t} |b_{l2}^M| = 1$$

d.h. für jeden einzelnen Fall werden die Gewichte so transformiert, daß die Summe ihrer Beträge über die t Attribute jeweils 1 ergibt. Dann kann $SAVE$ einmal für OLS und einmal für MONANOVA wie folgt berechnet werden:

$$\begin{aligned} SAVE_M &= \sum_{l=1}^{t} |b_{l2}^M - \beta_{l2}| \\ SAVE_O &= \sum_{l=1}^{t} |b_{l2}^O - \beta_{l2}| \end{aligned}$$

bzw., für jede Zelle des experimentellen Designs über die 50 Fälle hinweg:

$$\begin{aligned} \overline{SAVE}_M &= \sum_{i=1}^{50} \sum_{l=1}^{t} |b_{il2}^M - \beta_{il2}| \\ \overline{SAVE}_O &= \sum_{i=1}^{50} \sum_{l=1}^{t} |b_{il2}^O - \beta_{il2}| \end{aligned}$$

(Die Koeffizienten in diesen Formeln sind die standardisierten, auf eine zusätzliche Bezeichnung eigens für diese wird verzichtet.)

Für jede Zelle wurden ermittelt: $\overline{SAVE}_O$, $\overline{SAVE}_M$, die Zahl der Fälle mit $SAVE_M < SAVE_O$, $SAVE_M = SAVE_O$ und $SAVE_M > SAVE_O$, außerdem wurde jeweils ein t-Test der Differenzen zwischen beiden $\overline{SAVE}$-Werten berechnet[39].

In 5 der 8 Zellen mit fehlerfreien Präferenzurteilen war $\overline{SAVE}_M$ niedriger als $\overline{SAVE}_O$, wobei die Differenz einmal signifikant auf dem 5%-Level (bei 9 Attributen und „indecisive judge“) und einmal signifikant auf dem 1%-Level (bei 9 Attributen und Rangwerten) war, in den restlichen 3 „deterministischen“ Zellen war $\overline{SAVE}_M = \overline{SAVE}_O$. Nur einmal, in der hochsignifikantem Zelle,

[38] $\beta_{l1} = 0$, damit haben die Autoren einen Referenzstimulus ($\tilde{y}_1 = \sum \beta_{l1} = 0$) generiert, entsprechend können bei Dummy-Kodierung *im engeren Sinne* die Gewichte b_{l1}^O und b_{l1}^M als Referenzstufen gesetzt werden, so daß nur b_{l2}^O bzw. b_{l2}^M geschätzt werden und die Abweichung im Gesamtnutzen vom Referenzstimulus, die aus der 2. Stufe des jeweiligen Attributs resultiert, wiedergeben — vgl. S. 50: Die dort angesprochene Problematik solcher Kodierung ist hier irrelevant, da es sich eben um ein simuliertes Modell handelt, bei dem durch die Generierung „natürliche“ Referenzstufen gegeben sind, deren absolute Teilnutzenbeiträge (als Abweichung vom mittleren Gesamtnutzenwert aller Stimuli) nicht interssieren.

[39] Die Autoren geben hier keine weiteren Erläuterungen, es ist anzunehmen, daß nach dem t-Test für abhängige Stichproben die Zellenmittelwerte der $SAVE_O$ und $SAVE_M$ verglichen wurden. Nach dem zentralen Grenzwerttheorem (vgl. Bortz 1979, S. 118) sind diese Mittelwerte bei 50 Fällen pro Zelle in jedem Fall normalverteilt.

übertraf hier die Anzahl der Fälle mit $SAVE_M < SAVE_O$ die der Fälle mit $SAVE_M = SAVE_O$, ansonsten war letztere Anzahl immer > 45. Ganz anders dagegen die Ergebnisse in den Zellen mit fehlerbehafteten Daten: Hier war in allen 8 Zellen $\overline{SAVE}_O$ niedriger als $\overline{SAVE}_M$, in 4 Zellen (4 Attr., „indecisive"; 9 Attr., Rangwerte; 9 Attr., „fair"; 9 Attr., „exaggerating") war diese Differenz hochsignifikant auf dem 1%-Niveau, 3 mal war sie signifikant auf dem 5%-Level und einmal (9 Attribute, „indecisive judge") nicht signifikant. Allerdings zeigten alle diese Zellen auch einen nicht geringen (zwischen 10 und 20) Anteil von Fällen, in denen $SAVE_M < SAVE_O$ war.

Die Ergebnisse der Untersuchung lassen sich also in 3 Punkten zusammenfassen:

(**1**) Bei fehlerfreien Präferenzurteilen führt MONANOVA zu Ergebnissen, die entweder genauso gut oder besser als diejenigen sind, die durch die OLS-Regression erzielt werden. Dieses ist dadurch zu erklären, daß MONANOVA mit einer ANOVA-Lösung (welche dasselbe Ergebnis wie OLS produziert) startet: Erreicht diese bereits einen hinreichend kleinen Streßwert, dann bricht die Prozedur ab, ansonsten sucht sie nach besseren Lösungen, durch die der Streß hinreichend klein wird (vgl. Abschnitt 1.1.2, 19f) — MONANOVA kann hier also nur die gleichen oder bessere Ergebnisse produzieren. Dieser Effekt zeichnet sich mit wachsender Zahl von Attributen deutlicher ab und tritt v.a. dann auf, wenn die Präferenzdaten als Rangwerte vorliegen (vgl. die 1%-Signifikanz in der entsprechenden Zelle).

(**2**) Bei stochastischen Präferenzurteilen produziert OLS häufig deutlich bessere Koeffizientenschätzungen, als MONANOVA. Dieses Ergebnis zeigt sich auch dann, wenn die Kriteriumsvariable in Rangwerten vorliegt. Auch dieser Effekt ist bei größerer Anzahl von Attributen ausgeprägter (1%-Signifikanzen gegenüber 5%-Signifikanzen in der 4-Attribut-Bedingung).

(**3**) Hinsichtlich der Variablen C,D zeigt der Blick auf die Werte sowohl von $\overline{SAVE}_O$, als auch von $\overline{SAVE}_M$, daß immer die Bedingung „fair judge" die niedrigsten $\overline{SAVE}$-Werte erzielt, meist gefolgt der Rangwerte-Bedingung, unter der die $\overline{SAVE}$-Werte ungefähr auf demselben Niveau liegen, wie bei „exaggerating judge". Mit einer Ausnahme produziert „indecisive judge" die mit deutlichem Abstand höchsten $\overline{SAVE}$-Werte.

Wie ist es möglich, daß MONANOVA in der Fehlerbedingung durch OLS übertroffen wird, wo doch MONANOVA mit der metrischen Lösung beginnt und bei weiteren Iterationen Lösungen gesucht werden, die das Streß-Kriterium weiter verbessern? Es hat offensichtlich damit zu tun, daß das Modell von MONANOVA ein deterministisches ist und folglich eine stochastische Kompo-

nente nur unzureichend berücksichtigen kann. Wenn der Streß über die Iterationen reduziert wird, während gleichzeitig die Güte der Ergebnisse — ausgedrückt in den $SAVE$-Werten — sinkt, dann taugt der Streßwert bei Vorliegen einer stochastischen Komponente nicht als Indikator dafür, wie genau eine bestimmte Lösung die „wahren“ Parameter schätzt. Wie aber ist dieses zu erklären? Schließlich zeigt der Blick auf die zu minimierende Streßfunktion (Gleichung (1.1.4), S. 20), daß es sich hier um eine Kleinst-Quadrate-Schätzung handelt, die dem OLS-Schätzer (Gleichung 1.2.39, S. 45) sehr ähnlich ist: Wird im einen Fall die Summe der quadrierten Diffferenzen zwischen monoton transformierten Rangwerten z_j und den Schätzungen $\hat{y}_j$ minimiert, so erfolgt die OLS-Minimierung im Fall der CA mit Rangwerten für die Quadratsumme zwischen den untransformierten Rangwerten p_j und den $\hat{y}_j$. Ich habe in der Literatur keine theoretische Klärung der Eigenschaften von MONANOVA-Schätzungen gefunden, stattdessen schreiben Cattin & Bliemel (S. 472): „*... the properties of the estimates produced by a conjoint measurement algorithm like MONANOVA are unknown*“. Vielleicht kann die folgende Überlegung die obigen Fragen erklären:

Die Vorstellung, die der CA mit Rangwerten als abhängiger Variable zugrundeliegt, ist die, daß die Versuchsperson aus „wahren“ Teilnutzenwerten β_{lk} und eventuell aus einem Fehlerwert ϵ_j metrische Gesamtnutzenwerte y_j zusammensetzt (vgl. Gleichung (1.1.1)) und dann, bei der Rangreihung, diese metrischen Werte in ordinale Rangwerte transformiert. Man kann sich diese Transformation so vorstellen, daß dabei zu jedem y_j ein gewisser Betrag δ_j addiert wird, so daß die empirisch erhobenen Kriteriumswerte eine weitere additive Komponente enthalten — in Matrizendarstellung:

$$\mathbf{p} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\epsilon} \tag{2.1.1}$$

wobei $\mathbf{p}$ der Vektor der Rangwerte p_j und $\boldsymbol{\delta}$ der Vektor der Beträge δ_j ist (ansonsten wie wie vorne in Teil 1). Dabei wüchse die Varianz der Werte δ_j mit der Varianz der Abstände zwischen (im Rang) aufeinanderfolgenden metrischen Gesamtnutzenwerten y_j [40].

[40] Für die Varianz der δ-Beträge gilt: $\sigma_\delta^2 = E(\delta_j - E(\delta_j))^2 = E(p_j - \tilde{y}_j - \epsilon_j - (\bar{p} - \bar{y} - \bar{\epsilon}))^2 = E((p_j - \bar{p})^2 + (\tilde{y}_j - \bar{y})^2 + (\epsilon_j - \bar{\epsilon})^2 - 2(p_j - \bar{p})(\tilde{y}_j - \bar{y}) - 2(p_j - \bar{p})(\epsilon_j - \bar{\epsilon}) + 2(\tilde{y}_j - \bar{y})(\epsilon_j - \bar{\epsilon}) = \sigma_p^2 + \sigma_{\tilde{y}}^2 + \sigma_\epsilon^2 - 2\sigma_{p\tilde{y}} - 2\sigma_{p\epsilon}^2 = \sigma_p^2 + \sigma_y^2 - 2\sigma_{py}$.
Für ein gegebenes faktorielles Design mit n Stimuli sind Mittelwert und Varianz der p-Werte feste Größen ($\bar{p} = (n+1)/2$, $\sigma_p^2 = (1/n)\sum_{j=1}^{n}(j - \frac{n}{2})^2$), man kann der Einfachheit halber (und ohne Verlust an Allgemeingültigkeit) annehmen, daß die Werte y_j vor ihrer Umwandlung in Rangwerte so linear transformiert sind, daß $\bar{y} = \bar{p}$ und $\sigma_y = \sigma_p$. Es gilt dann $\sigma_\delta^2 = 2\sigma_p^2 - 2\sigma_{py}$.
Da σ_p^2 eine feste Größe ist, wird σ_δ^2 umso kleiner, je größer σ_{py} wird, ihren maximalen Wert erreicht diese Kovarianz mit $\sigma_{py} = \sigma_p^2$, dieses wird genau dann der Fall, wenn $y_j = p_j$ — wodurch alle $\delta_j = 0$ werden. Allgemein — ohne vorherige Annahme einer Angleichung der y-Skala an die Rangskala hinsichtlich Mittelwert und Standardabweichung — kann man dar-

MONANOVA, so könnte man sagen, ist daraufhin entwickelt, die Komponente $\boldsymbol{\delta}$ zu minimieren, diese ist sozusagen der „Fehler“, den MONANOVA kennt, während es die Komponente $\boldsymbol{\epsilon}$ nicht erkennt und als Bestandteil von $\boldsymbol{\delta}$ auffaßt[41]. Umgekehrt läßt sich so auch das Problem der OLS-Schätzung bei ordinalen Rangwerten darstellen: OLS kennt die Komponente $\boldsymbol{\delta}$ nicht, so daß die Beträge δ_j bei OLS in die Fehlerbeträge eingehen — wodurch diejenigen OLS-Voraussetzungen, die die Fehler betreffen, verletzt werden. Es sind aber für MONANOVA keine vergleichbaren Annahmen über die δ-Beträge formuliert, so daß es hier nicht so einfach ist, eine Vorstellung davon zu gewinnen, wie sich die Subsumierung der Komponente $\boldsymbol{\epsilon}$ unter $\boldsymbol{\delta}$ auswirkt.

Unter Annahme des obigen Modells (2.1.1) erhält man die folgende Aufgliederung der Gesamtvarianz der Rangwerte p_j:

$$\begin{aligned}
\sigma_p^2 &= E(p_j - E(p))^2 = E(\tilde{y}_j + \epsilon_j + \delta_j - (\bar{y} - \bar{\epsilon} - \bar{\delta}))^2 \\
&= E((\tilde{y}_j - \bar{y}) + (\epsilon_j - \bar{\epsilon}) + (\delta_j - \bar{\delta}))^2 \\
&= E((\tilde{y}_j - \bar{y})^2 + (\epsilon_j - \bar{\epsilon})^2(\delta_j - \bar{\delta})^2 + 2(\tilde{y}_j - \bar{y})(\epsilon_j - \bar{\epsilon}) \\
&\quad +2(\tilde{y}_j - \bar{y})(\delta_j - \bar{\delta}) + 2(\epsilon_j - \bar{\epsilon})(\delta_j - \bar{\delta})) \\
&= \sigma_{\tilde{y}}^2 + \sigma_\epsilon^2 + \sigma_\delta^2 + \sigma_{\tilde{y}\delta} + \sigma_{\epsilon\delta} \qquad (2.1.2)
\end{aligned}$$

MONANOVA beginnt — wie gesagt — mit einer ANOVA, also praktisch mit der OLS-Lösung. Sind deren Voraussetzungen für die Fehlerwerte erfüllt, so sind die gewonnen Schätzungen BLUE — von weiteren Schätzungen ist keine Verbesserung zu erwarten. Die tatsächliche Fehlervarianz des Modells bei der OLS-Schätzung ergibt sich sozusagen aus alledem, was in Gleichung

---

aus die Aussage ableiten, daß σ_δ^2 dann einen minimalen Wert annimmt, wenn die Abstände zwischen aufeinanderfolgenden Werten y_j jeweils gleich sind.

Da y_j sich additiv aus systematischer und stochastischer Komponente zusammensetzt, wird der Fall $y_j = p_j$ empirisch kaum auftreten, man kann aber nach möglichen Umständen fragen, durch die dieses näherungsweise erreicht wird. Dies könnte v.a. dann der Fall sein, wenn $\tilde{y}_j \approx p_j$ und σ_ϵ^2 sehr klein im Verhältnis zu $\sigma_{\tilde{y}}^2$, so daß durch die Fehlerbeträge nicht nur die Rangfolge der y_j gegenüber den $\tilde{y}_j$ nicht verändert wird (was durch $\tilde{y}_j \approx p_j$ bereits vorausgesetzt ist), sondern auch die Abstände zwischen aufeinanderfolgenden Werten ungefähr gleich bleiben und $p_j \approx \tilde{y}_j \approx y_j$.

Stellt man sich weitere Fälle vor, bei denen $y_j \approx p_j$, dann läuft es eigentlich immer darauf hinaus, daß $\tilde{y}_j \not\approx p_j$ — sei es, daß die Rangfolge der $\tilde{y}$-Werte nicht den p_j entspricht, sei es, daß die Abstände stark variieren — und die näherungsweise Übereinstimmung von y- und p-Werten durch die Fehlerwerte zufällig erzielt wird.

[41] $\boldsymbol{\delta}$ ist natürlich keine Fehlerkomponente in dem Sinne, daß sie tatsächlich unsystematische, von den Prädiktoren unabhängige Variation der Kriteriumswerte enthält. Trotzdem ist diese Komponente für das deterministische Modell von MONANOVA, da ja die metrischen Gesamtnutzenwerte y_j, aus denen die δ_j berechenbar wären, nicht direkt beobachtet werden, sozusagen der Umstand, der die „wahren“ Parameter verdunkelt und die ganze iterative Prozedur erst nötig macht: Gäbe es $\boldsymbol{\delta}$ nicht, d.h. lägen die metrischen Gesamtnutzenwerte untransformiert vor, dann müßten die Parameter nicht geschätzt, sondern könnten exakt ermittelt werden.

(2.1.2) rechts von $\sigma^2_{\tilde{y}}$ steht: Durch die additive Komponente $\boldsymbol{\delta}$ sind die OLS-Voraussetzungen nicht erfüllt, jedoch kann man vermuten, daß dann, wenn das datengenerierende Modell kompensatorisch — also rein additiv mit Attributen von annähernd gleicher Wichtigkeit — und die Varianz der Fehlerkomponente sehr klein im Vergleich zu der systematischen Komponente ist, die Abstände zwischen aufeinanderfolgenden Gesamtnutzenwerten y_j nicht sehr stark variieren (vgl. S. 74), so daß σ^2_δ sehr gering wird (vgl. Fußnote 40) und die OLS-Voraussetzungen nur geringfügig verletzt sind.

Dies wird deutlich, wenn man denjenigen Ausdruck in Gleichung (2.1.2) betrachtet, der besonders problematisch für die OLS-Schätzung erscheint, nämlich die Kovarianz zwischen systematischer Komponente und δ-Werten $\sigma^2_{\tilde{y}\delta}$: Dadurch ist direkt die grundsätzliche Annahme der Unkorreliertheit von systematischer und unsystematischer Komponente verletzt. Wäre $\tilde{y}_j = p_j$ — wie in Fußnote 40 wieder unter der Annahme, daß Mittelwert und Standardabweichung der y-Werte den Rangwerten angepaßt sind —, so würde $p_j = \tilde{y}_j + \epsilon_j + \delta_j = p_j + \epsilon_j + \delta_j$ und $-\epsilon_j = \delta_j$, d.h. die δ-Beträge dienten sozusagen der Korrektur der durch den Fehler verursachten Verzerrung der $\tilde{y}$-Werte, wären mit den Fehlerwerten perfekt negativ korreliert und somit: $\sigma^2_{\tilde{y}\delta} = \sigma^2_{p\delta} = E((p_j - \bar{p})(\delta_j - \bar{\delta})) = E((\tilde{y}_j + \epsilon_j + \delta_j - (\bar{y} + \bar{\epsilon} + \bar{\delta}))\delta_j) = E(-\epsilon_j(\tilde{y}_j - \bar{y})) = -\sigma^2_{\tilde{y}\epsilon} = 0$. Außerdem wäre natürlich $\sigma^2_\delta = \sigma^2_\epsilon$ und $\sigma_{\delta\epsilon} = -1$, so daß die tatsächliche Fehlervarianz bei OLS-Schätzung gleich $2\sigma^2_\epsilon - 1$ würde. Damit aber wäre zwar die Fehlervarianz vergrößert, aber nur eine grundsätzliche Annahme verletzt: $E(\boldsymbol{\epsilon}) = -1$! Diese Verletzung führt aber lediglich zu einem Bias bei der Schätzung der Konstanten β_0, die OLS-Schätzungen der restlichen Parameter wären unbiased und „best“. Die Überlegung setzt allerdings wieder, wie schon in Fußnote 40 angemerkt, voraus, daß σ^2_ϵ so klein ist, daß die Rangfolge der y_j gegenüber den $\tilde{y}_j$ nicht verändert wird. Im Fall $\tilde{y}_j \approx p_j$, so kann man vermuten, treffen diese Überlegungen näherungsweise zu und wird also v.a. die für die OLS-Schätzung problematische Kovarianz $\sigma_{\tilde{y}\delta}$ gegen Null gehen.

Falls $\tilde{y}_j \not\approx p_j$ kann aufgrund der Gleichung (2.1.2) ganz allgemein vermuten, daß die Qualität der OLS-Schätzungen davon abhängt, wie groß σ^2_δ im Verhältnis zu σ^2_ϵ ist. Eine im Vergleich zu σ^2_ϵ sehr geringe δ-Varianz läßt erwarten, daß unter den in (2.1.2) die tatsächliche Fehlervarianz bildenden Summanden alle drei zu σ^2_ϵ hinzukommenden Beträge relativ klein und unwichtig werden. In diesem Falle resultieren möglicherweise wiederum nur geringfügige Verletzungen der OLS-Fehlerannahmen.

Aufgrund dieser Überlegung kann man eine allgemeine Hypothese formulieren: Dann, wenn das datengenerierende Modell kompensatorisch ist, so daß $\tilde{y}_j \approx p_j$ (und wenn die Fehlerkomponente im Verhältnis zur systematischen Komponente hinreichend klein ist, daß durch sie keine größeren Unterschiede zwischen

der Rangfolge der $\tilde{y}$- und der y-Werte resultieren), greift die BLUE-Qualität der OLS-Lösung auch gegenüber der MONANOVA-Lösung. Andernfalls kann man auch dann gute Koeffizientenschätzungen durch den OLS-Algorithmus erwarten, wenn die Streuung der Komponente $\boldsymbol{\epsilon}$ sehr groß ist im Vergleich zu der der Komponente $\boldsymbol{\delta}$. Dagegen müßte in dem Maße, in dem die Varianz der δ-Beträge gegenüber der Fehlervarianz wächst, die nichtmetrische Methode im Vergleich mit der metrischen bessere Ergebnisse produzieren.

Die voranstehenden Überlegungen können hoffentlich einigermaßen erklären, warum in der Untersuchung von Cattin & Bliemel MONANOVA bei fehlerbehafteten Rangwerten (als abhängiger Variable) durch OLS übertroffen wurde, denn die Datengenerierung in dieser Studie erzeugt annähernd kompensatorische Modelle (die Teilnutzenwerte für alle Attribute wurden jeweils aus der gleichen Normalverteilung gezogen). Auch kann so erklärt werden, warum dieser Effekt bei 9 Attributen deutlich stärker ist, als bei 4: Die Fehlerwerte wurden jeweils aus derselben Verteilung gezogen, wodurch in den 2^9-Strukturen σ^2_ϵ relativ zur $\sigma^2_{\tilde{y}}$ kleiner wurde. Daß derselbe Effekt unter der Bedingung „fair judge“ erzielt wurde, kann nicht erstaunen, da bei dieser Transformation der y-Werte in Ratingskalenwerte die metrischen Eigenschaften der ersteren weitgehend erhalten bleiben. Die Erklärung, warum der Effekt bei „exaggerating judge“ auftritt, jedoch bei „indecisive judge“ nicht, fiele wieder schwerer, könnte aber anhand ähnlicher Überlegungen erfolgen — ich übergehe sie hier, da diese Bedingungen für die eigene Untersuchung nicht weiter von Interesse sind.

Die Untersuchung von Wittink & Cattin (1981):

Die Autoren vergleichen 4 Methoden: ANOVA, MONANOVA, LINMAP und LOGIT. Bei letzterem Verfahren handelt es sich um eines aus der Gruppe derer, die in der Übersicht in Abschnitt 1.1.2 nur kurz erwähnt und nicht weiter erläutert wurden: Methoden, die Regressionsmodelle mit qualitativ abgestufter (diskreter) abhängiger Variablen vermittels Maximum-Likelihood-Schätzung analysieren — es werden also diejenigen Parameterschätzungen gesucht, für die die empirisch ermittelten Antwortraten (der Stufen der abhängigen Variablen) am wahrscheinlichsten sind. In der CA bilden hierbei die Paarvergleiche zwischen den Stimuli eines faktoriellen Designs die empirische Grundlage. LOGIT wird hier nicht weiter erläutert, denn die Maximum-Likelihood-Verfahren sind für das Thema der eigenen Untersuchung nicht weiter von Bedeutung, und scheinen auch in der CA-Literatur eine eher untergeordnete Rolle zu spielen — es sei darum auf die Darstellung von McFadden (1976) verwiesen.

Die grundlegende Hypothese lautet auch hier, daß bei kompensatorischem Modell der Datengenerierung Präferenzurteile in Form von Rangwerten annähernd die metrischen Qualitäten der metrischen Gesamtnutzenwerte erhalten und somit bei solcher Datengenerierung die metrische Methode zumindest nicht schlechter abschneidet, als die nichtmetrischen. Aus diesem Grund wurde ei-

ne Bedingung nichtkompensatorischer Datengenerierung durch ein dominantes Attribut — d.h. ein solches, dessen Beitrag zum Gesamtnutzen die der anderen Attribute deutlich übertrifft — erzeugt. Es wurden die folgenden Bedingungen variiert:

A: Typus des (datengenerierenden) Modells: kompensatorisch vs. dominantes Attribut.

B: Fehlervarianz: 5% vs 20% vs. 35%.

C: Verteilung der Fehler: normal vs. uniform vs. Weibull.

Die letzte Variable C trägt dem Umstand Rechnung, daß die verschiedenen Verfahren unterschiedliche Annahmen über die Verteilung der Fehlerwerte enthalten: ANOVA setzt deren Normalverteilung voraus, für MONANOVA und LINMAP sind keine solchen Voraussetzungen formuliert und LOGIT nimmt die Weibull-Verteilung des Fehlers an. Die Ergebnisse der Simulation zeigten aber keinerlei Einfluß dieser Variablen auf die Güte der Analyseergebnisse, so daß sie bei der Darstellung dieser Ergebnisse von den Autoren nicht mehr berücksichtigt wurde.

Der Untersuchung wurde eine 3^6-Struktur zugrundegelegt und es wurden wie bei Cattin & Bliemel (1978) pro Zelle des Untersuchungsdesigns jeweils 50 Fälle generiert:

- Zunächst wurden wieder die Teilnutzenwerte ermittelt. Für die Bedingungen mit kompensatorischem Modell der Datengenerierung wurden alle $3 \times 6 = 18$ Werte aus der Normalverteilung mit Mittelwert 0 und Standardabweichung 1 gezogen, für die Bedingung mit dominantem Attribut ging man so vor:

  1. Man zog für 5 (nicht dominante) Attribute 15 Teilnutzenwerte β_{lk} aus NV(0,1) ($l = 1, \ldots, 5$).
  2. Man berechnete für jedes dieser 5 Attribute den Abstand $S_l = |\beta_{l(max)} - \beta_{l(min)}|$ zwischen den beiden extremen Teilnutzenwerten und bildete die Summe dieser Abstände $S = \sum_l S_l$.
  3. Die Teilnutzenwerte des sechsten Attributs wurden so ermittelt, daß hier der Abstand zwischen den Extremwerten $\beta_{6(max)} - \beta_{6(min)} = 9S$ und der mittlere Wert in der Mitte zwischen den Extremwerten liegt[42].

- Für ein reduziertes faktorielles Design mit 27 Stimuli wurden die 27 „wahren“ Gesamtnutzenwerte $\tilde{y}_j$ nach der üblichen additiven Formel berechnet und zu den $\tilde{y}_j$ dann Fehlerwerte ϵ_j addiert.

[42] Dieses berichten die Autoren, es geht daraus nicht hervor, wie die Werte *gezogen* wurden — man kann vermuten, daß der mittlere Wert aus NV(0,1) gezogen wurde und zu diesem $4.5S$ addiert bzw. subtrahiert wurde.

- Die ϵ_j wurden gezogen einmal aus einer Normalverteilung, einmal aus einer uniformen Verteilung und einmal aus einer Weibull-Verteilung (vgl. dazu z.B. Hartung et al. 1986, S. 230ff) jeweils um Mittelwert Null mit einer Standardabweichung, die ...

- ... einmal einem Anteil von 5%, einmal von 20% und einmal von 35% Fehlervarianz an der Gesamtvarianz der $y_j = \tilde{y}_j + \epsilon_j$ entsprach: Wenn E der gewünschte Anteil der Fehlervarianz an der Varianz σ_y^2 ist, dann ergibt sich die gewünschte Fehlervarianz σ_ϵ^2 aus $E = \sigma_\epsilon^2/(\sigma_{\tilde{y}}^2 + \sigma_\epsilon^2)$.

- Die resultierenden Werte y_j wurden in Rangwerte transformiert.

Aus A×B×C resultieren $2 \times 3 \times 3 = 18$ Zellen mit jeweils 50 Fällen — da sich aber die Variable C als unwichtig erwies, ergeben sich für die Ergebnisse, die die Autoren mitteilen, nur noch 6 (A×B) Zellen mit jeweils 150 Fällen. Auch hier geht aus der Beschreibung der Autoren nicht hervor, ob die Ziehung der „wahren" Teilnutzenwerte für jeden Fall in jeder Zelle neu erfolgte, oder ob nur ein Satz von $50 \times 6 \times 3$ solcher Werte gezogen (bzw. eventuell 2 Sätze für kompensatorisches Modell bzw. Modell mit dominantem Attribut) und zur Generierung der verschiedenen Bedingungen (Zellen) dann entsprechend verarbeitet wurde.

Zur Auswertung der Ergebnisse erzeugten die Autoren zu jedem Fall in jeder Zelle ein „validation sample", nämlich die fehlerfreien Gesamtnutzenwerte $\tilde{y}'_j$ der 27 Stimuli eines anderen reduzierten faktoriellen Designs (mit anderen Stimuli als dasjenige, welches die zu analysierenden Daten lieferte) und sie berechneten die prädiktive Validität („predictive validity") mittels 3 verschiedener Kriterien: **1.** Die mittlere Produkt-Moment-Korrelation zwischen „wahren" Werten $\tilde{y}'_j$ und vorhergesagten Werten $\hat{y}'_j$ im „validation sample" für jede der 6 Zellen. **2.** Den mittleren Streß zwischen den $\tilde{y}'_j$ und den $\hat{y}'_j$ für jede der 6 Zellen[43]. **3.** Die mittlere Spearman-Rangkorrelation zwischen „wahren" Werten $\tilde{y}'_j$ und vorhergesagten Werten $\hat{y}'_j$ für jede der 6 Zellen.

Die Verwendung der beiden verschiedenen Korrelationsmaße ergibt sich für die Autoren aufgrund einer gewissen Vorsicht bezüglich der Skalenqualität der geschätzten $\hat{y}$-Werte: *„The estimated part worths for a conjoint analysis model with a dependent variable measured on an ordinal scale have metric properties only in the limit"* (ebd., S. 104). Es wurde schon vorne, in Fußnote 12 (S. 17), auf die Vereinfachung hingewiesen, die man macht, wenn man die geschätzten Teilnutzenwerte als metrische Werte interpretiert. Wenn man die-

[43] Die Autoren geben nicht Berechnungsformel für diesen Streßwert an und verweisen auf Kruskal (1965). Dort ist das Maß lediglich so definiert, wie in Gleichung (1.1.4), S. 20, wiedergegeben — also mit der Summe der Abweichungsquadrate zwischen vorhergesagten Werte $\hat{y}_j$ und monoton transformierten Rangwerten z_j im Zähler. Gemeint ist hier wohl dieser Wert: $S = \sqrt{(\tilde{y}_j - \hat{y}_j)^2/(\hat{y}_j - \bar{y})^2}$.

se Vereinfachung nicht einfach mitmachen will, dann muß man natürlich auch die metrischen Qualitäten der aus der Addition der geschätzten Teilnutzenwerte sich ergebenden $\hat{y}$-Werten in Zweifel ziehen. Leider liefern die Autoren zu diesem Punkt weder genauere Darlegungen, noch Literaturhinweise auf solche Darlegungen. Jedenfalls resultiert aus der Vorsicht gegenüber den metrischen Qualitäten der geschätzten Gesamtnutzenwerte die Berechnung der Spearman-Korrelationen zusätzlich zu den Produkt-Moment-Korrelationen, da letztere ja die Intervallskalierung der korrelierten Variablenwerte voraussetzen. Die Verwendung von Rangkorrelationen macht darüberhinaus auch inhaltlich Sinn: Bei der abhängigen Variablen der typischen CA handelt es sich ja um Rangwerte, die Rangkorrelation zwischen vorhergesagten und „wahren" Gesamtnutzenwerten ist ein Maß genau dafür, wie gut die empirisch erhobenen Werte durch ein Schätzmodell „erklärt" werden, d.h. wie gut der „wahre Rang" eines Stimulus vorhergesagt werden kann.

Die Berechnung von mittleren Streßwerten als Ausdruck prädiktiver Validität begründen die Autoren damit, daß in der von ihnen verwendeten Programmversion von LINMAP — LINMAP II — die Berechnung von Produkt-Moment-Korrelationen für die LINMAP-Resultate nicht möglich sei. Da mir die Eigenarten dieser LINMAP-Version nicht bekannt sind — für die eigene Untersuchung wurde die Version LINMAP IV angeschafft, für die die Schwierigkeit nicht besteht —, gebe ich den Grund für diesen Umstand mit den Worten der Autoren wieder:„ *. . . some estimation procedures may require more than one stage for the estimation of part worths. This is particularly true for LINMAP when the data are generated by the dominant attribute model. Under multistage estimation the part worths obtained in one stage cannot be compared directly with part worths obtained in a later stage. For this reason Pearson correlation coefficients were not computed for the LINMAP results*" (ebd., S. 103).

Die Ergebnisse ihrer Untersuchung listen die Autoren in 3 Tabellen auf:

(**1**) Bei den mittleren Produkt-Moment-Korrelationen zeigen sich zwischen ANOVA und MONANOVA keinerlei erwähnenswerte Differenzen, während die LOGIT-Resultate in allen Zellen des experimentellen Designs niedrigere mittlere Korrelationen erreichen — mit einer Ausnahme sind die Differenzen zwischen ANOVA bzw. MONANOVA einerseits und LOGIT andererseits signifikant ($p < .05$).

(**2**) Bei den mittleren Streßwerten zeigt sich folgendes Bild: Beim kompensatorischen Modell der Datengenerierung erreicht ANOVA die niedrigsten Werte, dicht gefolgt (mit nicht signifikanten Differenzen) von MONANOVA. Bei LOGIT sind die Werte signifikant ($p < .05$) erhöht, am schlechtesten schneidet LINMAP ab, mit einer sehr deutlichen, signifikanten Erhöhung der Werte gegenüber LOGIT. Diese Verhältnisse kehren sich beim Modell mit dominantem Attribut genau um: Hier sind die

LINMAP-Werte deutlich die niedrigsten, mit signifikanten Abständen gefolgt von LOGIT, während ANOVA und MONANOVA — wiederum mit signifikanten Differenzen zu LOGIT — am schlechtesten abschneiden (mit leichten, nicht signifikanten Vorteilen jetzt für MONANOVA). Merkwürdigerweise sind dabei die mittleren Streßwerte unter der Bedingung mit dominantem Attribut insgesamt im Niveau niedriger, als bei kompensatorischem Modell (z.B. beträgt der niedrigste Wert bei kompensatorischem Modell — bei ANOVA und 5% Fehler — .122, der bei Modell mit dominantem Attribut — LINMAP-Analyse und 5% Fehler — .004).

(**3**) Die mittleren Rangkorrelationen berechnen die Autoren nur noch für LINMAP- und ANOVA-Ergebnisse, da diese beiden Methoden sich in den vorhergegangenen Auswertungen als die unter den beiden Typen des datengenerierenden Modells überlegenen herauskristallisiert haben. Hier nun zeigt sich, daß bei kompensatorischem Modell ANOVA immer höhere mittlere Korrelationen erzielt, als LINMAP, mit Differenzen, die jeweils auf dem 1%-Niveau signifikant sind. Bei Modell mit dominantem Attribut erzielt LINMAP bei 5% Fehler einen ebenfalls auf dem 1%-Niveau sigifikant höherern Wert als ANOVA, während bei 20% Fehler zwischen den beiden Methoden nur noch eine unwesentliche Differenz zugunsten von LINMAP besteht und sich das Verhältnis bei 35% Fehler umkehrt, so daß jetzt ANOVA den signifikant besseren Wert erzielt.

(**4**) Wie zu erwarten zeigt sich mit steigendem Fehler für alle Verfahren und für beide Modelltypen (A) eine Verschlechterung der prädiktiven Validität (Rückgang der mittleren Korrelationen bzw. Anstieg des mittleren Streß).

Man kann also die Ergebnisse dieser Studie auf den folgenden Punkt bringen: Bei kompensatorischem Modell der Datengenerierung zeigt sich die metrische Analyse als optimale Methode, während bei Vorliegen eines dominanten Attributs LINMAP die beste Schätzung bringt, zumindest solange die Fehlervarianz ein moderates Ausmaß annimmt. Damit kann man direkt an die Überlegungen anknüpfen, die oben im Zusammenhang mit den Gleichungen (2.1.1) und (2.1.2) angestellt wurden: Bei kompensatorischem Modell ist zu erwarten, daß die metrische Methode sehr gute Ergebnisse (im Vergleich zu jedem anderen linearen Schätzer) produziert, bei nicht-kompensatorischer Datengenerierung dagegen müßte aufgrund jener Überlegungen ANOVA (oder OLS) schlechter abschneiden, als nichtmetrische Methoden. Auch wurde oben angenommen, daß (bei nicht-kompensatorischem Modell) die vergleichsweise Güte der Ergebnisse aus metrischer Analyse mit der relativen Größe der Fehlervarianz wächst, was die unter (**3**) beschriebene Umkehrung der Differenzen der mittleren Spearman-Korrelationen bei dominantem Attribut mit steigendem Fehler

erklären könnte. Die Tendenz, daß die metrische Methode bei kompensatorischem Modell besser und beim nicht-kompensatorischen schlechter abschneidet als eine nichtmetrische, zeigt sich (in mittleren Pearson-Korrelationen und mittleren Streßwerten) auch gegenüber MONANOVA, jedoch sind hier, wie gesagt, die Differenzen sehr gering. Offensichtlich bleibt bei MONANOVA durch weitere Iterationen erzielte Lösung in jedem Fall nahe bei der anfänglichen ANOVA-Lösung — berücksichtigt man den größeren Rechenaufwand und die geringere Verfügbarkeit des Verfahrens gegenüber ANOVA oder OLS, so spricht dieses gegen die praktische Anwendung von MONANOVA, welches durch die metrischen Verfahren bequem und ohne Verlust an Qualität der Ergebnisse ersetzt werden kann.

Die Autoren unternehmen zu letzterem Punkt allerdings noch eine Analyse ihrer Ergebnisse, indem sie die Differenzen im Streß zwischen MONANOVA und LOGIT beim „estimation sample“, also bei den Stimuli des zur Parameterschätzung verwendeten reduzierten faktoriellen Designs, berechnen und argumentieren, daß immer dann, wenn diese Differenz sehr deutlich zugunsten von LOGIT ausfalle, dies ein Indikator dafür sei, daß MONANOVA nur ein lokales Optimum erreicht habe. Sie finden diesen Fall mit großer Häufigkeit beim Modell mit dominantem Attribut. Die Folgerung daraus ist, daß MONANOVA hier möglicherweise deutlich besser als ANOVA abschneiden könnte, wenn es mit einer anderen als der ANOVA-Lösung gestartet würde.

Die Auswertung der Ergebnisse dieser Untersuchung läßt allerdings einige Fragen offen, sie betreffen v.a. die Mittelwertsbildungen und den Streßwert. Grundsätzlich gilt: *„Da Korrelationswerte keine Maßzahlen auf einer Intervallskala darstellen, sind auch Mittelwerte und Varianzen von mehreren Korrelationen nicht interpretierbar“*(Bortz 1979, S. 261). Damit Mittelwerte von Korrelationen sinnvolle Werte werden, sind die Korrelationen zunächst in Fishers-Z-Werte zu transformieren (vgl. Bortz 1979, S. 260), wodurch sie annäherungsweise normalverteilt und intervallskaliert werden. Hinsichtlich der interessierenden Signifikanzen bei Mittelwertsunterschieden kann man sich, wenn man es nicht so genau nimmt, auf das zentrale Grenzwerttheorem berufen: Unabhängig von der Verteilungsform der Korrelationen (in der Grundgesamtheit, d.h. hier über unendlich viele generierbare Fälle hinweg) ist dadurch bei einer Stichprobengröße von 50 Fällen pro Zelle des experimentellen Designs gewährleistet, daß sich die Mittelwerte pro Zelle (in unendlich vielen Replikationen der Stichprobengenerierung) um den wahren Wert normalverteilen (vgl. Bortz 1979, S. 118f). Genaugenommen aber verbietet die fehlende Intervallskalenqualität der Korrelationen die Mittelwertsbildung! Den Streßwerten schließlich haftet eine gewisse Undurchsichtigkeit an: Zumindest für die nichtmetrischen Formen ist aufgrund des Fehlens einer Fehlertheorie die Verteilungsform der Streßwerte weder bekannt, noch ableitbar, und auch die Intervallskalenqualität dieser Werte kann zumindest dann, wenn man, wie oben dargelegt, die metri-

schen Qualitäten der $\hat{y}$-Werte in Zweifel zieht, als fraglich betrachtet werden. Darüber hinaus ist wenig über die Eigenschaften des Streß bei den verschiedenen verwendeten Verfahren bekannt, ganz abgesehen davon, daß aus den Darlegungen der Autoren gar nicht genau hervorgeht, wie sie diesen Wert berechnen (vgl. Fußnote 43). Wie ist es z.B. zu erklären, daß bei dominantem Attribut LOGIT signifikant schlechtere mittlere Produkt-Moment-Korrelationen erzielt, als ANOVA bzw. MONANOVA, jedoch signifikant bessere mittlere Streßwerte? Warum zeigt sich der Effekt bei den mittleren Rangkorrelationen, daß bei großem Fehler auch in der Bedingung mit dominantem Attribut ANOVA signifikant besser als LINMAP abschneidet, nicht beim mittleren Streß? Und was hat es zu bedeuten, daß die mittleren Streßwerte ausgerechnet bei dominantem Attribut insgesamt auf deutlich niedrigerem Niveau liegen, als bei kompensatorischer Datengenerierung? Die Beantwortung solcher Fragen scheitert daran, daß über die Streßwerte zu wenig bekannt ist.

2.1.2 Experimentelle Faktoren und Hypothesen

In allen der zuletzt dargestellten Untersuchungen zeigt sich, daß MONANOVA — unter welchen Umständen auch immer — eigentlich nie wesentlich bessere Ergebnisse produziert, als die metrischen Methoden, wenn überhaupt, dann kann man bei wachsender Fehlerbehaftetheit der Daten von einer Überlegenheit von ANOVA bzw. OLS gegenüber MONANOVA reden. Es ist mit Blick auf die berichteten Ergebnisse zu erwarten, daß auch bei weiterer Simulation experimenteller Bedingungen keine Umstände gefunden würden, unter denen sich MONANOVA als zu bevorzugende Methode erwiese — zumindest solange man hierbei im Rahmen solcher Bedingungen bleibt, die praktischen Wert insofern besitzen, als sie Umstände eines individuelles Urteilsprozesses simulieren, welche in der Realität zu erwarten sind. Die Resultate der Wittink-Cattin-Studie zeigen, daß dieses nicht auf alle nonmetrischen Verfahren generalisierbar ist: Es scheint so, daß hinsichtlich der „Güte“ der Ergebnisse unter solchen Bedingungen, bei denen man überhaupt einen Unterschied zwischen metrischen und nichtmetrischen Verfahren erwartet, diese Unterschiede am ehesten in der Gegenüberstellung von metrischen Verfahren und LINMAP auftreten. Tatsächlich unterscheidet sich LINMAP prinzipiell von anderen nichtmetrischen Verfahren dadurch, daß es sich hier eben nicht um eine iteratives Gradientenverfahren handelt, das mit einer „metrischen“ Lösung einsetzt. Interessant also erscheint die Gegenüberstellung von LINMAP und metrischen Verfahren: Das grundlegende Anliegen der vorliegenden Arbeit war der Vergleich von LINMAP und OLS hinsichtlich der „Robustheit“ der Analyseergebnisse.

Der Begriff Robustheit, der ja auch im Titel der vorliegenden Arbeit gebraucht wurde, wurde von Carmone, Green & Jain (1978) übernommen und meint

hier das, was in den oben besprochenen Simulationsstudien als abhängige Variable der experimentell variierten Bedingungen untersucht und bisher meist mit „Güte“ der Analyseergebnisse umschrieben wurde. „Robustheit“ erscheint mir hierfür besonders geeignet, weil dieser Begriff nicht nur in seiner umgangssprachlichen Bedeutung genau das trifft, was gemeint ist, sondern auch in seiner in der Statistik gebräuchlichen Verwendung: *„Als robust bezeichnen wir ... einen Test, wenn Entscheidungen, die auf Grund des Tests getroffen werden, weitgehend davon unabhängig sind, ob die Voraussetzungen erfüllt sind oder nicht“* (Bortz 1979, S. 166). Es geht hier um Robustheit in diesem Sinne, daß untersucht wird, wie bestimmte Verletzungen von Modellannahmen (fehlerbehaftete Daten bei nichtmetrischen Verfahren, ordinalskalierte Daten bei metrischen Verfahren, Interaktionen zwischen Attributen etc.) die „Güte“ der Ergebnisse beeinflussen. Die verschiedenen besprochenen Untersuchungen verwenden unterschiedliche Maße, durch die diese Robustheit operationalisiert wird — über die Operationalisierung in der vorliegenden Untersuchung werden weiter hinten noch Erörterungen angestellt.

In den besprochenen Untersuchungen bestätigte sich zum Teil die Vermutung, daß dann, wenn das datengenerierende Modell kompensatorisch ist, metrische Methoden ungeachtet der Verletzung ihrer Voraussetzung bezüglich der Skalenqualität der abhängigen Variablen sehr robust sind: Kompensatorisch ist ein Modell dann, wenn bei rein additiver Verknüpfung der Attribute dieselben jeweils dieselbe Wichtigkeit besitzen, sodaß eine „hohe“ Stufe in einem Attribut eine „niedrige“ in einem anderen kompensieren kann — man könnte von einem „streng kompensatorischen“ Modell dann sprechen, wenn der Range und die Abstände der metrischen Teilnutzenwerte eines Attributs für jedes Attribut gleich sind (z.B. in dem einfachen Fall, daß jedes Attribut dieselben Teilnutzenwerte besitzt). Wenn nun die Teilnutzenwerte eines jeden Attributs aus derselben Verteilung mit derselben Standardabweichung gezogen werden, dann ist zu erwarten, daß das datengenerierende Modell aufgrund der gezogenen Werte annähernd kompensatorisch wird, weil diese eben demselben Wertebereich entstammen.

Wenn dagegen der Range der Teilnutzenwerte eines bestimmten Attributs sehr groß im Vergleich zu allen anderen Attributen ist, so daß in der Addition der Teilnutzenwerte zu Gesamtnutzenwerten diejenigen dieses Attributs sehr viel größere additive Beiträge liefern, als alle anderen, dann ist dieses Attribut dominant und das datengenerierende Modell nicht mehr auch nur annähernd kompensatorisch: Für diesen Fall ist zu erwarten, daß sich die „wahren“ Gesamtnutzenwerte $\tilde{y}_j$ sozusagen in Gruppen um die Teilnutzenwerte des dominaten Attributs verteilen — mit kleinen Abständen innerhalb dieser Gruppen und großen zwischen ihnen, so daß den gleichen Abständen, die zwischen aufeinanderfolgenden Rangwerten p_j bestehen sehr große Unterschiede in den Abständen entsprechender aufeinanderfolgender $\tilde{y}$-Werte gegenüberstünden.

Damit aber gäben die Rangwerte nicht mehr die metrischen Qualitäten der „wahren“ Werte wieder. Bei kompensatorischem Modell ergeben sich zwar auch unterschiedliche Abstände zwischen im Rang aufeinanderfolgenden $\tilde{y}_j$, jedoch dürften diese Unterschiede nicht so stark variieren, wie im Falle eines dominanten Attributs, so daß die zugehörigen Rangwerte hier „stärker metrisch“ sein sollten und folglich die Voraussetzungen metrischer Verfahren weniger verletzen. In den Überlegungen, die vorne rund um die Gleichungen (2.1.1) und (2.1.2) angestellt wurden, wurde versucht, diesen Gedanken zu präzisieren.

Das in Tabelle 2.1 wiedergegebene Zahlenbeispiel soll diese Zusammenhänge veranschaulichen: Beispiel 1 ist im obigen Wortsinn „streng kompensatorisch“ (vgl. die β-Werte), während für Beispiel 2 für 3 Attribute jeweils 3 Teilnutzenwerte aus NV(0,1) gezogen wurden. Zur Erzeugung eines dominanten Attributs wurde dann in Beispiel 3 zum höchsten der 3 Teilnutzenwerte des ersten Attributs 10 addiert und vom niedrigsten dieser 3 Werte 10 subtrahiert, im Beispiel 4 wurde dementsprechend 20 addiert bzw. subtrahiert. Die hier nur zu Demonstrationszwecken aufgeführten Beispiele enthalten keinen Fehler. Aus der 3^3-Struktur ergeben sich durch Addition der Teilnutzenwerte 27 Gesamtnutzenwerte $\tilde{y}_j$, welche in Tabelle 2.1 jeweils in der Reihenfolge ihrer Rangwerte wiedergegeben werden — nachdem sie so transformiert wurden, daß sie hinsichtlich Mittelwert und Standardabweichungen den Rangwerten p_j angepaßt sind. Die aufgeführten δ-Werte geben die Abweichugen dieser transformierten $\tilde{y}_j$ von den p_j wieder.

Es ist deutlich zu erkennen, wie bei den Beispielen mit dominantem Attribut die $\tilde{y}$-Werte sich auf 3 Gruppen verteilen, innerhalb derer sehr kleine Abstände zwischen diesen Werten bestehen, zwischen denen jedoch die Abstände sehr groß sind — die 3 Gruppen werden durch die 3 Stufen des dominanten Attributs gebildet. Auch sieht man ohne weitere Berechnungen, daß die $\tilde{y}_j$ des Beispiels 2 (fast) immer relativ nahe bei den ihnen zugehörigen Rangwerten liegen — zumindest gilt dies ab dem neunten Wert. In den Beispielen 3 und 4 dagegen weichen die Gesamtnutzenwerte stärker von den Rangwerten ab. Entsprechend sind die Beträge der δ_j in letzteren beiden Beispielen höher, als im Beispiel 2 — was sich wiederum deutlich auf die Varianzen σ_δ^2 auswirkt. Merkwürdigerweise ist σ_δ^2 im Beispiel 2 niedriger, als im „streng kompensatorischen“ Beispiel 1. Dieser Umstand braucht hier aber nicht weiter zu verwirren, da der für Beispiel 1 ermittelte Werte nicht unbedingt mit den Werten aus den anderen Beispielen vergleichbar ist: Das Beispiel enthält nur 7 verschiedene $\tilde{y}$-Werte und die δ_j geben hier die Abweichung eines jeden $\tilde{y}_j$ vom mittleren Rangwert der Gruppe, zu der er gehört, wieder (z.B. haben die 6 Stimuli mit $\tilde{y}_j = 8.492$ den mittleren Rangwert 7.5). Unter der Voraussetzung, daß die Rangwertbildung forciert wird, d.h. daß keine gleichen Ränge auftreten dürfen, gäbe für die Rangfolge innerhalb der Gruppen gleich präferierter Stimuli der Fehler den Ausschlag, welcher in den Beispielen nicht enthalten

ist. Würde man die δ_j des Beispiels 1 als tatschliche Abweichungen von den in der ersten Spalte angegebenen p_j berechnen, so bliebe zwar deren Summe gleich, die Varianz aber würde sich (wegen der darin enthaltenen Quadrierung der Abstände) verändern. Trotzdem veranschaulicht die Gegenüberstellung der Beispiele 1 und 2 die kompensatorischen Qualitäten von Modellen, die dadurch generiert wurden, daß die Teilnutzenwerte aller Attribute aus jeweils derselben Normalverteilung gezogen wurden.

So, wie man vermuten kann, daß dominante Attribute bei der Datengenerierung die „pseudometrischen“ Qualitäten der Rangwerte mindern, könnte man ähnliches auch für den Fall erwarten, daß im Schätzmodell nicht spezifizierte Interaktionen zwischen den Attributen bei der Bildung multiattributiver Präferenzurteile wirksam werden: Ist eine Struktur von Teilnutzenwerten zweier Attribute so beschaffen, daß bei rein additiver Verknüpfung ein hoher Wert einer Stufe eines Attributs einen niedrigen Wert einer Stufe des anderen Attributs ausgleicht, dann wird dies nicht mehr der Fall sein, wenn zur additiven Verknüpfung noch eine Interaktion hinzutritt — es ist nicht mehr zu erwarten, daß Modelle der Datengenerierung, bei denen Interaktionen eine Rolle spielen, annähernd kompensatorischen Charakter haben. Angesichts der Vielzahl von Interaktionen unterschiedlichen Grades und Ausmaßes, welche auch in Abhängigkeit von der Größe einer gegebenen multiattributiven Struktur vorstellbar sind, fällt es hier schwerer, anhand anschaulicher Beispiele die erwartete Auswirkung von Interaktionen auf die Präferenzdaten zu präzisieren, jedoch kann ganz allgemein vermutet werden, daß die Abnahme der kompensatorischen Qualitäten des Modells auch eine Abnahme der metrischen Qualitäten der Rangwerte nach sich zieht — was zu der Hypothese führt, daß im Falle solcher (nicht spezifizierter) Interaktionen die nichtmetrischen Verfahren robuster sind, als die metrischen.

Unter den Studien, die oben besprochen wurden, enthält nur die von Carmone et al. die experimentelle Bedingung einer Interaktion, die bei der Datengenerierung wirksam wird — allerdings sind, wie schon besprochen, die Ergebnisse dieser Untersuchung nur von begrenztem Wert. Es ergab sich dort, abgesehen von der allgemeinen Verschlechterung der Analyseergebnisse (welche auf die Modellmisspezifikation, die durch die Interaktion gegeben ist, zurückzuführen ist) kein Unterschied in der Robustheit von MONANOVA und ANOVA bei Vorliegen der Interaktion. Es wurde bei der Besprechung dieser Studie auch schon darauf hingewiesen, daß die untersuchte Interaktion 4. Ordnung nur von begrenztem praktischen Wert ist: Interessanter scheint die Frage, wie Interaktionen niederer Ordnung, die empirisch viel eher zu erwarten sind, die Genauigkeit der Schätzungen beeinflussen. Aufgrund der besprochenen Unzulänglichkeiten der Arbeit von Carmone et al. darf man wohl sagen, daß der Einfluß nicht spezifizierter Interaktionen auf die Robustheit verschiedener Verfahren der CA noch kaum untersucht ist.

	Beispiel 1		Beispiel 2		Beispiel 3		Beispiel 4	
β_{11}	1		0.3946		10.3946		20.3946	
β_{12}	0		0.0815		0.0815		0.0815	
β_{13}	-1		-0.1088		-10.1088		-20.1088	
β_{21}	1		0.98383		0.98383		0.98383	
β_{22}	0		-0.0122		-0.0122		-0.0122	
β_{23}	-1		-1.2715		-1.2715		-1.2715	
β_{31}	1		-1.2865		-1.2865		-1.2865	
β_{32}	0		0.32034		0.32034		0.32034	
β_{33}	-1		-0.08986		-0.08986		-0.08986	
p_j	$\tilde{y}_j$	δ_j	$\tilde{y}_j$	δ_j	$\tilde{y}_j$	δ_j	$\tilde{y}_j$	δ_j
1	-2.523	3.523	-1.614	2.614	2.626	-1.63	3.503	-2.50
2	2.985	0.015	-0.343	2.343	3.730	-1.73	4.066	-2.07
3	2.985	0.015	1.749	1.251	3.787	-.787	4.095	-1.10
4	2.985	0.015	6.380	-2.38	4.108	-.108	4.259	-.259
5	8.492	-.992	6.799	-1.80	4.706	0.294	4.563	0.437
6	8.492	-.992	7.651	-1.65	4.890	1.110	4.658	1.342
7	8.492	-.992	8.070	-1.07	5.269	1.731	4.850	2.150
8	8.492	-.992	9.120	-1.12	5.809	2.191	5.126	2.874
9	8.492	-.992	9.743	-.743	6.187	2.813	5.318	3.682
10	8.492	-.992	10.162	-.162	12.021	-2.02	12.991	-2.99
11	14.000	0.000	10.392	0.608	13.124	-2.12	13.553	-2.55
12	14.000	0.000	12.483	-.483	13.182	-1.18	13.583	-1.58
13	14.000	0.000	13.453	-.453	13.502	-.502	13.746	-.746
14	14.000	0.000	14.724	-.724	14.100	-.100	14.051	-.051
15	14.000	0.000	14.793	0.207	14.285	0.715	14.145	0.855
16	14.000	0.000	16.064	-.064	14.663	1.337	14.338	1.662
17	14.000	0.000	16.816	0.184	15.203	1.797	14.613	2.387
18	19.508	0.992	17.533	0.467	15.581	2.419	14.806	3.194
19	19.508	0.992	18.156	0.844	21.528	-2.53	22.536	-3.54
20	19.508	0.992	18.804	1.196	22.631	-2.63	23.099	-3.10
21	19.508	0.992	20.896	0.104	22.689	-1.69	23.128	-2.13
22	19.508	0.992	21.447	0.553	23.010	-1.01	23.292	-1.29
23	19.508	0.992	22.718	0.282	23.607	-.607	23.596	-.596
24	25.015	-.015	24.187	-.187	23.792	0.208	23.691	0.309
25	25.015	-.015	24.810	0.190	24.171	0.829	23.883	1.117
26	25.015	-.015	25.458	0.542	24.711	1.289	24.159	1.841
27	30.523	-3.52	27.550	-.550	25.089	1.911	24.351	2.649
	$\sigma_\delta^2 = 1.409$		$\sigma_\delta^2 = 1.282$		$\sigma_\delta^2 = 2.636$		$\sigma_\delta^2 = 4.560$	

Tabelle 2.1: Zahlenbeispiele für kompensatorische und nichtkompensatorische Modelle der Datengenerierung einer 3^3-Struktur.

Ausgehend von den vorgestellten früheren Ergebnissen und den oben angestellten Überlegungen sollte also in der vorliegenden Untersuchung die Robustheit metrischer vs. nichtmetrischer CA unter verschiedenen Bedingungen des datengenerierenden Modells untersucht werden: Ein kompensatorisches Modell sollte Modellen mit dominantem Attribut und Modellen mit nichtspezifizierten Interaktionen gegenübergestellt werden, wobei die „Stärke“ der Dominanz bzw. der Interaktion variiert werden sollte. Dabei wird von der allgemeinen Hypothese ausgegangen, daß mit wachsender Dominanz eines Attributs einerseits und mit wachsender Stärke einer Interaktion andererseits die relative Güte der Analyseergebnisse von LINMAP im Vergleich zu OLS zunimmt. Es kann außerdem ganz allgemein vermutet werden, daß die Güte von Analyseergebnisse — mit welcher Methode auch immer erzielt — sowohl mit ansteigender Dominanz eines Attributs, als auch mit wachsender nichtspezifizierter Interaktion abnimmt: Es wird jeweils unsystematische Varianz vergrößert, im einen Fall durch das Anwachsen von σ_δ^2 bei der Transformation in Rangwerte, im anderen durch den Beitrag der unberücksichtigten Interaktionskomponente.

In den Überlegungen zu Gleichung (2.1.2) wurde auch erwogen, daß bei nichtkompensatorischen Modellen der Datengenerierung die relative Genauigkeit der Ergebnisse aus metrischer Analyse dann sehr hoch sein dürfte, wenn die Fehlervarianz im Vergleich zur Varianz der δ-Werte sehr hoch ist: Die Fehlerkomponente bei der OLS-Schätzung dürfte dann näherungsweise die in Abschnitt 1.2.2 dargelegten Voraussetzungen erfüllen. Umgekehrt könnte man vermuten, daß in diesem Fall die nichtmetrischen Verfahren, die ja eigentlich deterministische Modelle voraussetzen, relativ schlecht abschneiden. In den Ergebnissen der Studie von Wittink & Cattin scheint sich diese Vermutung zu bestätigen: Bei dominatem Attribut und niedrigem Fehler schneidet ANOVA relativ schlecht im Vergleich zu LINMAP ab, bei hohem Fehler kehrt sich dieses Verhältnis um. In der vorliegenden Arbeit wurde also auch die Größe der Fehlerkomponente variiert, mit der Erwartung, daß bei nicht-kompensatorischen Modellen OLS sich unter hoher Fehlervarianz als robuster erweist, als LINMAP und also mit wachsendem Fehler den erwarteten Vorteil des nichtmetrischen Verfahrens bei diesen Modellen ausgleicht.

Eine in den dargestellten Studien kaum untersuchte Bedingung ist die Größe des faktoriellen Designs. Lediglich Carmone et al. (vgl. S. 55ff) variieren dieselbe mit dem Ergebnis, daß die Verfahren gegenüber diesem Faktor sehr robust sind — bei wiederum fraglicher Generalisierbarkeit der Ergebnisse dieser Untersuchung. Ansonsten wird es in Veröffentlichungen zur CA im Allgemeinen die Robustheit aller Verfahren gegenüber der Größe des faktoriellen Designs vorausgesetzt, z.B. schreiben Cattin & Bliemel (1978, S. 473) nur: „*An orthogonal array can be used since it is sufficient to estimate the main effects of each attribute*“. Dieses hat eine gewisse Berechtigung darin, daß es sich bei den in der CA verwendeten reduzierten faktoriellen Designs praktisch im-

mer um „orthogonal arrays“ handelt, also um Versuchspläne, die prinzipiell eine Schätzung der Haupteffekte der unabhängigen Variablen ohne jede Einschränkung erlauben bzw. so konstruiert sind, daß die Reduktion der Beobachtungen (im Fall der CA Stimuli) sich lediglich auf die Schätzbarkeit von Interaktionseffekten auswirkt. Abgesehen von der prinzipiellen Schätzbarkeit der Haupteffekte durch reduzierte orthogonale Designs aber kann in der CA die Robustheit dieser Schätzungen aufgrund des Skalenniveaus der abhängigen Variablen hinterfragt werden: Angenommen die Rangwerte p_j eines vollständigen faktoriellen Designs hätten annähernd metrische Qualitäten, d.h. die Abstände zwischen den (im Rang) aufeinanderfolgenden metrischen Gesamtnutzenwerten y_j wären ungefähr gleich, dann stellt sich die Frage, was passiert, wenn nun einige der y_j ausgewählt und in Rangwerte umgewandelt werden. Hätte man z.B. die 5 Gesamtnutzewerte $(1.1, 1.9, 3, 4.1, 4.9)$ mit den „pseudometrischen“ Rangwerten $(1, 2, 3, 4, 5)$, dann hätten bei Auswahl der Werte $(1.1, 1.9, 4.9)$ die zugehörigen Rangwerte $(1, 2, 3)$ keineswegs mehr annähernd metrische Qualität.

Es kann aufgrund der Orthogonalitätsbedingung, welche reduzierte orthogonale Designs erfüllen, für den hier interessierenden Fall, daß bei kompensatorischem Modell der Datengenerierung die Rangwerte (des vollen faktoriellen Designs) annähernd metrisch sind, gefolgert werden, daß dieses auch für die Rangwerte eines reduzierten Designs gelten muß. Nach Addelman (1992a, S. 23) ist die Orthogonalität zweier Attribute in einem reduzierten faktoriellen Design denn gegeben, wenn

$$\nu_{ab} = \frac{\nu_{a\cdot}\nu_{\cdot b}}{n} \tag{2.1.3}$$

wobei:

$\nu_{a\cdot}$	=	die Häufigkeit, mit der die Stufe a des Attributs A im Design vorkommt,
$\nu_{\cdot b}$	=	die Häufigkeit, mit der die Stufe b des Attributs B im Design vorkommt,
ν_{ab}	=	die Häufigkeit, mit der Stufe a des Attributs A zusammen mit Stufe b des Attributs B im Design vorkommt,
n	=	Anzahl der Stimuli, die im Design erhalten sind.

Diese Bedingung ist auf den ersten Blick wenig anschaulich, bedeutet aber praktisch, daß in einem orthogonalen Design die Stufen eines jeden Attributs mit proportionalen Häufigkeiten zusammen mit den Stufen eines jeden anderen Attributs vorkommen. Kommt also z.B. in einem 3^2-Design die 1. Stufe des 1. Attributs einmal zusammen mit der 1. Stufe des 2. Attributs, einmal zusammen mit der 2. Stufe des 2. Attributs und zweimal zusammen mit der 3. Stufe des 2. Attributs vor, so stehen die Häufigkeiten, in denen die 2. bzw. die 3. Stufe des 1. Attributs jeweils zusammen mit den Stufen 1 bis 3 des 2. Attributs vorkommt, im Verhältnis $1 : 1 : 2$. In den gebräuchlichen orthogonalen Designs

für symmetrische Strukturen (d.h. solche mit gleicher Anzahl von Stufen pro Attribut) ist diese Bedingung dadurch erfüllt, daß praktisch jede Kombination der Stufen zweier Attribute für jede Kombination von 2 Attributen gleich häufig vorkommt (vgl. z.B. Abbildung 1.3, S. 16: Jede Kombination der Stufen zweier Attribute kommt genau einmal vor). Aufgrund dieses Umstands kann man genauso, wie man bei kompensatorischem Modell der Datengenerierung für das volle Design erwartet, daß die Abstände aufeinanderfolgender Gesamtnutzenwerte einander ähnlich sind, dieses auch für die Abstände der aufeinanderfolgenden y-Werte des reduzierten Designs erwarten: Es ist gewährleistet, daß z.B. die Kombinationen „hoher“ Levels der Attribute nicht überproportional häufig vorkommen.

Die erschöpfende Darlegung bzw. Veranschaulichung dieses Zusammenhangs würde allerdings hier wohl zuviel Aufwand beanspruchen — ich begnüge mich deshalb mit der Feststellung, daß zwar einerseits die annähernd metrischen Qualitäten der Rangwerte aus kompensatorischen Modellen auch bei reduzierten orthogonalen Designs (bei symmetrischen Strukturen) zu erwarten sind, daß aber insgesamt unklar bleibt, inwieweit die bisher in diesem Kapitel angestellten Überlegungen für den Fall reduzierter orthogonaler Designs plausibel bleiben. Welche Veränderung ist z.B. für das Verhältnis der in Gleichung 2.1.2 enthaltenen Varianzen σ_δ^2 und σ_ϵ^2 zu erwarten, wenn die Anzahl der Stimuli reduziert wird? Es erschien angesichts solcher Unklarheiten in jedem Falle lohnend, in der vorliegenden Untersuchung auch die Größe des faktoriellen Designs als experimentellem Faktor zu behandeln und entsprechend zu variieren. Eine sehr vage Hypothese könnte lauten, daß eventuell bei dominantem Attribut der Wegfall einzelner Stimuli bewirken könnte, daß für die reduzierte Designgröße Unterschiede in den Abständen von im Rang aufeinanderfolgenden Stimuli sozusagen stärker ins Gewicht fallen, so daß, um im Rahmen der angestellten Überlegungen zu bleiben, σ_δ^2 in Relation zu σ_ϵ^2 wächst. Es könnte also sein, daß bei nicht-kompensatorischer Datengenerierung — v.a. dei Dominanz eines Attributs — die sinkende Größe des faktoriellen Designs sich zugunsten der LINMAP-Resultate und zuungunsten OLS-Ergebnisse auswirkt. Bei kompensatorischer Datengenerierung ist solches nicht zu erwarten.

Nach dem allgemeinen Tenor in der CA-Literatur ist zwar keine generelle Verschlechterung der Analyseergebnisse durch Reduktion des faktoriellen Designs zu erwarten, jedoch erscheint dieses fraglich: Immerhin bedeutet diese Reduktion ein Herabsetzen der Freiheitsgrade, die zur Schätzung der Parameter zur Verfügung stehen. Es wird aus diesem Grund die obige Erwartung bezweifelt und angenommen, daß eine solche generelle Verschlechterung stattfinden wird.

Es ergeben sich somit für das Design der vorliegenden Untersuchung die in Tabelle 2.2 in tabellarischer Form wiedergegebenen experimentellen Faktoren bzw. abhängige Variablen. Dort sind auch die Bedingungen angegeben, die als

experimentelle Faktoren		Anz. d.	Bedingungen	
Bez.	Inhalt	Stufen	Bez.	Inhalt
A	Typus des datengenerierenden Modells	5	A1	kompensatorisch
			A2	schwach dominantes Attribut
			A3	stark dominantes Attribut
			A4	schwache Interaktion
			A5	starke Interaktion
B	Ausmaß des Fehlers	3	B1	10% Fehlervarianz
			B2	20% Fehlervarianz
			B3	33% Fehlervarianz
C	Größe des faktoriellen Designs	3	C1	27 Stimuli
			C2	18 Stimuli
			C3	9 Stimuli
D	Typus der Analyse	2	D1	OLS (metrisch)
			D2	LINMAP (nichtmetrisch)

Tabelle 2.2: Experimentelle Faktoren und Bedingungen.

Stufen dieser Faktoren gebildet wurden — bei der Beschreibung der Datengenerierung in Abschnitt 2.1.3 werden diese Bedingungen genauer dargestellt und eventuell mit ihrer Auswahl verbundene Probleme diskutiert. Auf die in der Tabelle angegebenen Bezeichnung (Spalten „Bez.“) von Faktoren und Bedingungen (Stufen der Faktoren) wird in der weiteren Darstellung zurückgegriffen.

Ähnlich wie in der Untersuchung von Carmone et al. sollten die generierten Daten varianzanalytisch ausgewertet werden (vgl. Abschnitt 2.2.1). Die oben besprochenen Hypothesen lauten für diese Varianzanalyse:

- Signifikanz der Haupteffekte A (generelle Verschlechterung der Analyseergebnisse bei nicht-kompensatorischer Datengenerierung), B (generelle Verschlechterung der Analyseergebnisse mit wachsendem Fehler) und C (generelle Verschlechterung bei Reduktion des faktoriellen Designs), keine Signifikanz von D.

- Signifikanz der Interaktionen erster Ordnung A×D (Vorteil der metrischen Methode bei kompensatorischem Modell, Vorteil der nichtmetrischen bei nicht-kompensatorischem), B×D (mit wachsendem Fehler Vorteil der metrischen Methode).

- Signifikanz der Interaktion 2. Ordnung A×B×D (Umkehrung des erwarteten Vorteils von LINMAP bei nicht-kompensatorischen Modellen mit wachsendem Fehler). Eventuell Signifikanz der Interaktion

A×C×D (Verstärkung des erwarteten Vorteils von LINMAP bei nichtkompensatorischen Modellen bei Reduktion des faktoriellen Designs).

Es sei aber betont, daß diese Hypothesen den Charakter vager Erwartungen haben. Dieses liegt in der Natur einer solchen Simulationsstudie, die ja eben darum angestellt wird, weil der Einfluß verschiedener Bedingungen auf die Robustheit der Methoden theoretisch nicht geklärt ist und folglich nur unklar vorausgesagt werden kann.

2.1.3 Datengenerierung

In Abbildung 2.5 sind — ähnlich wie bei der Darstellung der Studie von Cattin & Bliemel (1978) — alle Schritte zur Bildung der Daten für die 90 Zellen des experimentellen Designs zusammengefaßt, sie werden im folgenden besprochen.

Der vorliegenden Untersuchung wurde eine 3^3-Struktur — 3 Attribute mit jeweils 3 Stufen — zugrundegelegt. Es gibt eigentlich keine besonderen Gründe, die für oder gegen gerade diese Struktur zur Klärung der in Abschnitt 2.1.2 entwickelten Fragestellungen sprechen: Man hätte auch eine andere aus der Vielzahl möglicher Produktstrukturen auswählen können, die Frage, wie die Robustheit verschiedener Verfahren der CA bzw. wie die Ergebnisse von Studien wie der vorliegenden von der Anzahl der Attribute oder der Stufen einer Struktur abhängen, stellt ein weiteres mögliches Thema für Simulationsstudien dar und wäre unabhängig von den hier behandelten Fragestellungen zu untersuchen. Cattin & Bliemel (1978) haben in ihrer Studie die Anzahl der Attribute als einen experimentellen Faktor variiert (vgl. S. 59ff), jedoch sind die für diesen Faktor gewonnenen Ergebnisse kaum interpretierbar, da mit der größeren Anzahl der Attribute eine Verkleinerung der Fehlervarianz und ein Rückgang vom vollen faktoriellen Design auf ein reduziertes einherging, so daß Unterschiede zwischen den Abstufungen dieses Faktors ebenso auf die Veränderung dieser Bedingungen zurückführbar sind.

Es soll allerdings nicht unerwähnt bleiben, daß ursprünglich geplant war, eine 3^4-Struktur zu verwenden. Nachdem zum Zwecke dieser Untersuchung die zur Zeit erhältliche Programmversion von LINMAP — LINMAP IV wird unter der Bezeichnung Conjoint-LINMAP von der Firma Bretton-Clark Software vertrieben[44] — angeschafft war, stellte sich heraus, daß das Programm Probleme bei der Berechnung des vollen faktoriellen Designs mit 81 Stimuli hatte: In vielen Fällen war es nicht in der Lage, die Analyse zu beenden, sondern brach diese mit einer Meldung ab, welche auf einen Fehler bei der Programmierung schließen läßt („illegal negative value“). Auch nach mehreren Rückfragen war

[44] Anschrift: Bretton-Clark Software, 89 Headquarter Plaza, North Tower 14^{th} Floor, Morristown, NJ 07960, USA.

die Firma nicht in der Lage, das Problem zu beheben. Offensichtlich ist diese Version nicht in der Lage, eine größere Anzahl von Rangwerten zu verarbeiten, die Beteuerung der Firma, daß das Problem noch nie aufgetreten sei, ist vielleicht dadurch zu erklären, daß in praktischen Anwendungen fast immer reduzierte faktorielle Designs bearbeitet werden. Da das Programm bei der 3^3-Struktur fehlerfrei arbeitete und die begonnene Untersuchung fortgesetzt werden mußte (und zur Abfassung einer Diplomarbeit auch nicht endlos Zeit vorhanden ist), ergab sich der Rückgriff auf diese Struktur, gegen die es ja auch keine prinzipiellen theoretischen Einwände gibt.

Es wurden für jede Zelle des experimentellen Designs 100 Fälle generiert, d.h. in jeder der 90 Zellen, die sich aus den möglichen Kombinationen $\mathsf{A}\times\mathsf{B}\times\mathsf{C}\times\mathsf{D}$ ergeben, befinden die Präferenzdaten von 100 „Versuchspersonen“. Die Basis dieser Generierung bildet für alle Zellen *ein* Satz von 100×9 Teilnutzenwerten $\beta_{lk}(l, k = 1, 2, 3)$, welche jeweils aus einer Normalverteilung mit Mittelwert 0 und Standardabweichung 1 zufällig gezogen wurden (Kasten $\boxed{1}$ in Abbildung 2.5). Man kann darüber diskutieren, ob es nicht notwendig gewesen wäre, diese Ziehung für jede Zelle des Designs neu vorzunehmen. Mir erschien das gewählte Vorgehen nicht nur ökonomischer, sondern v.a. auch sinnvoller: Es ist so gewährleistet, daß durch die wiederholten Ziehungen keine zusätzliche zufällige Varianz zwischen den Zellen erzeugt wird, die durch das experimentelle Design nicht bewältigt bzw. fälschlicherweise auf die spezifizierten unabhängigen Variablen zurückgeführt wird. Es könnte z.B. sein, daß bei separater Ziehung für jede Zelle zufällig in einer Zelle überproportional viele „Versuchspersonen“ mit einem sehr hohen Range der Teilnutzenwerte auf einem Attribut enthalten sind — auf diese Weise enthielte diese Zelle dann sozusagen eine klammheimliche, wenn auch sehr schwache, Dominant-Attribut-Bedingung, welche eventuell Einfluß auf die durchschnittliche Güte der Analyseergebnisse in dieser Zelle hätte. Man muß allerdings einräumen, daß bei 100 generierten Fällen — also 100×9 Zufallsziehungen — pro Zelle ein solcher zufällig entstandener systematischer Zelleneffekt sehr unwahrscheinlich ist. Dennoch folgt die einmalige Ziehung der β-Werte dem Grundsatz, beim Experiment möglichst alle Bedingungen außer den experimentell variierten konstant zu halten. Bei der vorgesehenen varianzanalytischen Auswertung der Ergebnisse (vgl. Abschnitt 2.2.1) ist dann zu berücksichtigen, daß man es in den einzelnen Zellen mit abhängigen Messungen zu tun hat, daß es sozusagen dieselben Versuchspersonen sind, die unter den verschiedenen Bedingungen Daten generiert haben und getestet wurden. Folglich ist eine Varianzanalyse mit Meßwiederholung zu rechnen: Eine solche Analyse hat den Vorteil, daß sie durch Eliminierung der Varianz „zwischen den Versuchspersonen“ — also hier zwischen den Fällen — die Fehlervarianz verkleinert, so daß dieses Verfahren besonders sensibel gegenüber auch sehr kleinen Mittelwertsunterschieden ist und damit signifikante Effekte eher aufdeckt, als eine Varianzanalyse ohne Meßwiederholung (vgl. Bortz 1979,

1		Ziehe β_{lk} aus NV(0,1) für $l,k = 1,2,3$
2	A1	Bilde: $\tilde{y}_j = \sum_{l=1}^{3}\sum_{k=1}^{3} x_{lk}\beta_{lk}$
	A2	Setze: $\beta'_{lk} = \beta_{lk}$ für $l = 2,3$ $\beta'_{1k} = \beta_{1k} + 5$ wenn $\beta_{1k} = \max(\beta_{1k})$ $\beta'_{1k} = \beta_{1k} - 5$ wenn $\beta_{1k} = \min(\beta_{1k})$ $\beta'_{1k} = \beta_{1k}$ wenn $\beta_{1k} \neq \min(\beta_{1k})$ und $\beta_{1k} \neq \max(\beta_{1k})$ Bilde: $\tilde{y}_j = \sum_{l=1}^{3}\sum_{k=1}^{3} x_{lk}\beta'_{lk}$
	A3	Setze: $\beta''_{lk} = \beta_{lk}$ für $l = 2,3$ $\beta''_{1k} = \beta_{1k} + 15$ wenn $\beta_{1k} = \max(\beta_{1k})$ $\beta''_{1k} = \beta_{1k} - 15$ wenn $\beta_{1k} = \min(\beta_{1k})$ $\beta''_{1k} = \beta_{1k}$ wenn $\beta_{1k} \neq \min(\beta_{1k})$ und $\beta_{1k} \neq \max(\beta_{1k})$ Bilde: $\tilde{y}_j = \sum_{l=1}^{3}\sum_{k=1}^{3} x_{lk}\beta''_{lk}$
	A4	Bilde: $\tilde{y}_j = \sum_{l=1}^{3}\sum_{k=1}^{3} x_{lk}\beta_{lk} + (\sum_{k=1}^{3} x_{1k}\beta_{1k})(\sum_{k=1}^{3} x_{2k}\beta_{2k})$
	A5	Bilde: $\tilde{y}_j = \sum_{l=1}^{3}\sum_{k=1}^{3} x_{lk}\beta_{lk} + 2(\sum_{k=1}^{3} x_{1k}\beta_{1k})(\sum_{k=1}^{3} x_{2k}\beta_{2k})$
3		Berechne $\sigma^2_{\tilde{y}}$
	B1:	Ziehe ϵ_j aus NV$(0, \sqrt{0.11\sigma^2_{\tilde{y}}})$, bilde $y_j = \tilde{y}_j + \epsilon_j$
	B2:	Ziehe ϵ_j aus NV$(0, \sqrt{0.25\sigma^2_{\tilde{y}}})$, bilde $y_j = \tilde{y}_j + \epsilon_j$
	B3:	Ziehe ϵ_j aus NV$(0, \sqrt{0.5\sigma^2_{\tilde{y}}})$, bilde $y_j = \tilde{y}_j + \epsilon_j$
4	C1:	Transformiere die 27 y_j in Rangwerte p_j
	C2:	Wähle 18 y_j nach Spalte C2 in Tab. 2.3 aus und transformiere sie in Rangwerte p_j
	C3:	Wähle 9 y_j nach Spalte C3 in Tab. 2.3 aus und transformiere sie in Rangwerte p_j
5	D1:	Analysiere die p_j mit OLS
	D2:	Analysiere die p_j mit LINMAP

Abbildung 2.5: Datengenerierung

S. 407f). Dieser prinzipielle theoretische Vorteil der Meßwiederholungsanalyse gab letztlich den Ausschlag zum Aufbau eines Meßwiederholungsdesigns[45].

Die Zufallsziehung der β-Werte — wie auch der Fehlerwerte ϵ_j (s.u. bzw. $\boxed{3}$ in Abb. 2.5) erfolgte mithilfe der SAS-Funktion RANNOR (SAS Institute 1990, S. 589f). Im Anhang A wird der SAS-Job zur Generierung der β_j abgedruckt.

Der nächste Schritt nach der Ziehung von jeweils 9 β-Werten für 100 Fälle ist die Generierung von „wahren“ Teilnutzenwerten $\tilde{y}_j$ ($\boxed{2}$ in Abb. 2.5). Entsprechend Faktor A sollte dabei das Modell der Datengenerierung fünffach variiert werden. Das kompensatorische Modell — A1 — wurde in gewohnter Weise durch $\tilde{y}_j = \sum_l \sum_k x_{lk}\beta_{lk}$ gebildet. Zur Bildung der Bedingungen mit dominantem Attribut — A2, A3 — wurden vor dieser Addition die Teilnutzenwerte des 1. Attributs β_{1k} modifiziert: Es wurde bei jedem Fall zum größten dieser 3 Werte ein bestimmter Betrag addiert, vom kleinsten dieser Werte derselbe Betrag subtrahiert. Dieser Betrag war einmal, für „schwache“ Dominanz des ersten Attributs, 5 und einmal, für „starke“ Dominanz, 15. Die Beträge wurden mehr oder weniger willkürlich festgelegt, sie gewährleisten, daß der Range der Teilnutzenwerte des 1. Attributs deutlich gegenüber dem des 2. bzw. 3. Attributs vergößert wird, und sie stellen wohl realistische Größenordnungen von Dominanzen dar. Man hätte die Dominanz mit noch höheren Beträgen auf die Spitze treiben können, nähert sich aber damit dem Fall, daß die nichtdominanten Attribute schlicht irrelevant werden. Zur Bildung der Bedingungen mit nichtspezifizierten Interaktionen — A4, A5 — wurde der Gesamtnutzenwert eines jeden Stimulus durch die übliche Addition der zugehörigen Teilnutzenwerte und zusätzlich des Produkts des Teilnutzenwerts des 1. mit dem des 2. Attributs gebildet. Es wurde also nur eine Interaktion 1. Ordnung erzeugt. Selbstverständlich wären hier weitere Bedingungen mit weiteren Interaktionen denkbar, die Beschränkung auf die eine Interaktion erfolgte aufgrund der Notwendigkeit, den Untersuchungsgegenstand bzw. die Anzahl der variierten Bedingungen einzuschränken, und der Erwartung, daß eine solche Interaktion 1. Ordnung in realen Conjoint-Strukturen am ehesten eine Rolle spielen könnte (vgl. auch die in Abschnitt 2.1.1 geübte Kritik an der Untersuchung von Carmone, Green & Jain 1978). Zur Erzeugung einer „starken“ Interaktion wurde

[45]Dem theoretischen Vorteil stehen allerdings, wie noch zu zeigen sein wird, eine Fülle praktischer Nachteile gegenüber, welche sich dem Anwender allerdings erst im Verlauf der Anwendung erschließen. Diese Nachteile beruhen alle auf einer gewissen mangelhaften Eignung der verfügbaren ANOVA-Computerprogramme zur Berechnung von Meßwiederholungsdesigns. Bei der Durchführung der Untersuchung wurden — wie gesagt — die daraus resultierenden Schwierigkeiten erst offenbar, als die Analyse mit diesen Programmen durchgeführt werden sollte — um die Fertigstellung der Diplomarbeit nicht endlos zu verzögern, wurde zu diesem Zeitpunkt nicht nocheinmal von vorne begonnen. Ich will aber gerne gestehen, daß bei nochmaliger Durchführung einer solchen Untersuchung die praktischen Nachteile der Meßwiederholungsanalyse gegenüber den theoretischen Vorteilen den Ausschlag gäben.

der Interaktionsterm mit 2 multipliziert. Am Ende von Schritt [2] stehen also 5 Datensätze, von denen jeder 100×27 „wahre “ Gesamtnutzenwerte $\tilde{y}_j$ enthält.

Man könnte an diesem experimentellen Faktor A einen Schönheitsfehler darin erkennen, daß in ihm gewissermaßen 2 inhaltlich zu unterscheidende unabhängige Variablen vermengt sind, nämlich einmal Dominanz vs. Nicht-Dominanz eines Attributs und zum anderen korrekte (additive) Spezifikation des Modells vs. Modellmisspezifikation (Interaktion). Diese Vermengung hätte durch die Einführung von 2 getrennten experimentellen Faktoren — z.B. „Ausmaß der Dominanz“ und „Größe der Interaktion“ — vermieden werden können, was aber im Detail einige Tücken nach sich zöge: Wie hätte man dann die Bedingungen dieser beiden Faktoren praktisch kombinieren sollen? Hätte z.B. bei dominantem Attribut dieses an der Interaktion 1. Ordnung beteiligt oder hätte diese zwischen den restlichen beiden Attributen stattfinden sollen bzw. müßte man dann nicht alle diese beiden Möglichkeiten untersuchen? Und welche Erwartungen bestünden für jede dieser Möglichkeiten bzw. welche theoretischen Vorstellungen kann man sich für den Fall, daß in einer Präferenzstruktur sowohl ein dominantes Attribut, als auch eine Interaktion wirksam werden, bilden? Kurz: Das Zusammenspiel von dominantem Attribut und Interaktion erfordert einigen theoretischen und untersuchungstechnischen Mehraufwand, der angesichts der im Abschnitt 2.1.2 diskutierten Fragestellungen eigentlich überflüssig erscheint. Bereits dann, wenn man den 5-stufigen Faktor A nur in 2 3-stufige Faktoren aufteilt, ergeben sich 162 Zellen des experimentellen Designs (anstatt der ohnehin schon großen Zahl von 90 Zellen)! Und hinsichtlich der Fragestellung, die zur Bildung des Faktors A geführt hat — nämlich die nach der vergleichsweisen Robustheit von OLS vs. LINMAP unter kompensatorischen vs. nicht-kompensatorischen Bedingungen der Datengenerierung —, ist das Zusammenspiel von dominantem Attribut und Interaktion wenig interessant, es interessiert hier eigentlich nur der Vergleich von kompensatorischer Bedingung und Bedingungen mit dominantem Attribut einerseits und der von kompensatorischer Bedingung und Interaktionsbedingungen andererseits. Die angestrebte varianzanalytische Auswertung (vgl. Abschn. 2.2.1) erfordert lediglich kategoriale Abstufungen der unabhängigen Variablen, im Rahmen dieser Auswertung sind dann die interessierenden Einzelvergleiche innerhalb der Stufen des Faktors A anzustellen. Problematisch wird dieses Vorgehen dann, wenn sich der Faktor A wider Erwarten weder in seinem Haupteffekt, noch in Interaktionen als signifikant erweisen sollte: Es wäre dann denkbar, daß durch Wegfall der Interaktions- oder der Dominant-Attribut-Bedingungen der jeweils verbleibende — dann 3. stufige — Faktor Signifikanzen produzierte, daß also z.B. Unterschiede zwischen kompensatorischem Modell und Modell mit dominantem Attribut innerhalb des 5-stufigen Faktors dadurch sozusagen verwischt würden, daß 2 (Interaktions-)Bedingungen hinzugezogen werden, welche sich vom kompensatorischen Modell überhaupt nicht unterscheiden. Sollte also der

Attribut						Attribut					
A_1	A_2	A_3	C1	C2	C3	A_1	A_2	A_3	C1	C2	C3
1	1	1	•	•	•	2	2	3	•		•
1	1	2	•			2	3	1	•	•	•
1	1	3	•	•		2	3	2	•		
1	2	1	•			2	3	3	•	•	
1	2	2	•	•	•	3	1	1	•	•	
1	2	3	•	•		3	1	2	•	•	
1	3	1	•	•		3	1	3	•		•
1	3	2	•	•		3	2	1	•	•	•
1	3	3	•		•	3	2	2	•		
2	1	1	•			3	2	3	•	•	
2	1	2	•	•	•	3	3	1	•		
2	1	3	•	•		3	3	2	•	•	•
2	2	1	•	•		3	3	3	•	•	
2	2	2	•	•							

Tabelle 2.3: Faktorielle Designs: 1,2,3 repräsentiert die jeweilige Stufe des Attributs, • zeigt die Zugehörigkeit des jeweiligen Stimulus zum faktoriellen Design an.

Fall völliger Nichtsignifikanz des Faktors A eintreten, so könnten — gewissermaßen als Kompromißlösung aller genannten Schwierigkeiten — 2 getrennte Varianzanalysen mit einem jeweils 3-stufigen Faktor A gerechnet werden, indem einmal alle Zellen mit den Interaktionswerten, also die Bedingungen A4, A5, und einmal die Bedingungen A2, A3 aus dem varianzanalytischen Design entfernt werden.

Aus den 5 Datensätzen „wahrer“ Gesamtnutzenwerte werden nun (Schritt 3 in Abbildung 2.5) entsprechend Faktor B 15 Sätze von Gesamtnutzenwerten y_j gebildet: Dazu muß zunächst für jeden einzelnen Fall in jedem der 5 $\tilde{y}$-Datensätze die Varianz der $\tilde{y}$-Werte $\sigma^2_{\tilde{y}}$ berechnet werden. Die 3 Abstufungen in der Größe der Fehlervarianz ergeben sich dann dadurch, daß man (für jeden einzelnen Fall) die Fehlerwerte ϵ_j aus einer Normalverteilung um Null mit einer Varianz, die einem zuvor spezifizierten Anteil der Fehlervarianz σ^2_ϵ an der Gesamtvarianz der y-Werte — $\sigma^2_y = \sigma^2_{\tilde{y}} + \sigma^2_\epsilon$ — entspricht, zieht: Soll dieser Anteil z.B. 10% betragen, so ergibt sich die Fehlervarianz aus $0.\bar{1}\sigma^2_{\tilde{y}}$. Für die 3 Bedingungen B1, B2, B3 wurden sukzessive die Fehlervarianzen $0.11\sigma^2_{\tilde{y}}, 0.25\sigma^2_{\tilde{y}}$ und $0.25\sigma^2_{\tilde{y}}$ bei der Ziehung der Fehlerwerte verwendet, diese entsprechen (ungefähr) Varianzanteilen von 10%, 20% und 33% an der Gesamtvarianz σ^2_y.

Aus den 15 Datensätzen mit den y-Werten von jeweils 100 Fällen wurden dann

(Schritt [4] in Abb. 2.5) 45 Datensätze gebildet, die Rangwerte p_j enthalten: Es wurden einmal gemäß Bedingung C1 die 27 Gesamtnutzenwerte y_j eines jeden Falles in ganzzahlige Rangwerte von 1 bis 27 transformiert (volles faktorielles Design), dann wurde gemäß Bedingungen C2 und C3 reduzierte orthogonale faktorielle Designs mit 18 bzw. 9 Stimuli ausgewählt und bei jedem Fall die y_j der in diesem Designs enthaltenen Stimuli in ganzzahlige Rangwerte von 1 bis 18 bzw. 1 bis 9 transformiert. Tabelle 2.3 zeigt die in den 3 faktoriellen Designs enthaltenen Stimulusprofile in tabellarischer Form. Im Anhang A ist einer von vielen SAS-Jobs, mit dem die Schritte [2] bis [4] zur Erzeugung eines der 45 Datensätze durchgeführt wurden, abgedruckt.

Bei der Auswahl der reduzierten Designs wurde auf die schon erwähnten „basic plans“ von Addelman (1964a) zurückgegriffen (vgl. Abb. 1.3, S. 16). Dort allerdings findet sich kein solcher Plan für die 3^3-Struktur, die hier verwendeten reduzierten Designs wurden aus den Spalten 1,2,3 des „basic plan 2“ für die 3^4-Struktur (ebd., S. 36) und aus den Spalten 5,6,7 des „basic plan 4“ für die 3^7-Struktur (ebd., S.37) gebildet. Wie man Tabelle 2.3 entnehmen kann, gilt die mit Gleichung (2.1.3) (S. 78) angegebene Orthogonalitätsbedingung für jedes Paar von Attributen bei beiden reduzierten faktoriellen Designs: Im Design C2 kommt in jeder Spalte (bei jedem Attribut) jede Stufe genau 6 mal vor, bei jedem Spaltenpaar jede mögliche Stufenkombination genau 2 mal — $(6 \times 6)/18 = 2$. Im Design C3 kommt in jeder Spalte jede Stufe genau 3 mal vor, bei jedem Spaltenpaar jede Stufenkombination genau einmal — $(3 \times 3)/9 = 1$.

Schließlich wurde (Schritt [5] in Abbildung 2.5) jeder einzelne Fall in jedem der 45 Sätze von Rangwerten einmal mit OLS und einmal mit LINMAP analysiert, es wurden also insgesamt 4500 OLS-Regressionen und 4500 LINMAP-Analysen gerechnet. Ersteres geschah unter Verwendung der SAS-Prozedur REG (SAS Institute 1989, S. 1351ff), der Regression wurde die Effektkodierung der Stimulusprofile zugrundegelegt — die Form dieser Kodierung und der Grund für ihre Verwendung wurden bereits im Punkt 2, S. 50f besprochen. Für jedes 3-stufige Attribut ergeben sich demnach 2 Kodiervariablen, die die Werte 1, 0 und -1 annehmen können. Für das verwendete reduzierte 9-Stimulus-Design (vgl. C3 in Tabelle 2.3) ergibt sich die folgende Kodiermatrix $\mathbf{X}$ des OLS-Schätzmodells

$\mathbf{p} = \mathbf{Xb} + \mathbf{e}$:

$$\begin{array}{c} \begin{array}{ccccc} A_1 & \vdots & A_2 & \vdots & A_3 \end{array} \\ \left[\begin{array}{ccccc} \beta_{11} & + & \beta_{21} & + & \beta_{31} \\ \beta_{11} & + & \beta_{22} & + & \beta_{32} \\ \beta_{11} & + & \beta_{23} & + & \beta_{33} \\ \beta_{12} & + & \beta_{21} & + & \beta_{32} \\ \beta_{12} & + & \beta_{22} & + & \beta_{33} \\ \beta_{12} & + & \beta_{23} & + & \beta_{31} \\ \beta_{13} & + & \beta_{21} & + & \beta_{33} \\ \beta_{13} & + & \beta_{22} & + & \beta_{31} \\ \beta_{13} & + & \beta_{23} & + & \beta_{32} \end{array}\right] \end{array} \quad \rightarrow \quad \mathbf{X} = \begin{array}{c} \begin{array}{cccccc} A_1 & & \vdots & A_2 & \vdots & A_3 \end{array} \\ \left[\begin{array}{cccccc} 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & -1 & -1 & -1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & -1 & -1 \\ 0 & 1 & -1 & -1 & 1 & 0 \\ -1 & -1 & 1 & 0 & -1 & -1 \\ -1 & -1 & 0 & 1 & 1 & 0 \\ -1 & -1 & -1 & -1 & 0 & 1 \end{array}\right] \end{array}$$

Für das 18-Stimuli-Design (C2 in Tabelle 2.3)) erhält man so:

$$\begin{array}{c} \begin{array}{ccccc} A_1 & \vdots & A_2 & \vdots & A_3 \end{array} \\ \left[\begin{array}{ccccc} \beta_{11} & + & \beta_{21} & + & \beta_{31} \\ \beta_{11} & + & \beta_{21} & + & \beta_{33} \\ \beta_{11} & + & \beta_{22} & + & \beta_{32} \\ \beta_{11} & + & \beta_{22} & + & \beta_{33} \\ \beta_{11} & + & \beta_{23} & + & \beta_{31} \\ \beta_{11} & + & \beta_{23} & + & \beta_{32} \\ \beta_{12} & + & \beta_{21} & + & \beta_{32} \\ \beta_{12} & + & \beta_{21} & + & \beta_{33} \\ \beta_{12} & + & \beta_{22} & + & \beta_{31} \\ \beta_{12} & + & \beta_{22} & + & \beta_{32} \\ \beta_{12} & + & \beta_{23} & + & \beta_{31} \\ \beta_{12} & + & \beta_{23} & + & \beta_{33} \\ \beta_{13} & + & \beta_{21} & + & \beta_{31} \\ \beta_{13} & + & \beta_{21} & + & \beta_{32} \\ \beta_{13} & + & \beta_{22} & + & \beta_{31} \\ \beta_{13} & + & \beta_{22} & + & \beta_{33} \\ \beta_{13} & + & \beta_{23} & + & \beta_{32} \\ \beta_{13} & + & \beta_{23} & + & \beta_{33} \end{array}\right] \end{array} \quad \rightarrow \quad \mathbf{X} = \begin{array}{c} \begin{array}{cccccc} A_1 & & \vdots & A_2 & \vdots & A_3 \end{array} \\ \left[\begin{array}{cccccc} 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & -1 & -1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & -1 & -1 \\ 1 & 0 & -1 & -1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & -1 & -1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & -1 & -1 & 1 & 0 \\ 0 & 1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 1 & 0 & 1 & 0 \\ -1 & -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 0 & 1 & 1 & 0 \\ -1 & -1 & 0 & 1 & -1 & -1 \\ -1 & -1 & -1 & -1 & 0 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 \end{array}\right] \end{array}$$

Bei dieser Kodierung erhält man für jedes Attribut l die beiden OLS-Schätzungen b_{l1} und b_{l2}, die als direkte Schätzungen der Teilnutzenwerte β_{l1} und β_{l2} der generierten Produktstruktur aufzufassen sind. Die Schätzung für β_{l3} erhält man durch $b_{l3} = -b_{l1} - b_{l2}$ (vgl. wieder Punkt 2, S. 50).

Zur LINMAP-Analyse wurde, wie bereits erwähnt, die Programmversion LINMAP IV verwendet. Es müssen hier keine Kodierungen spezifiziert werden, das Programm erfordert eigentlich nur die Eingabe der Werte für die abhängige Variable und einige Angaben zum Typus dieser Werte (Rangwerte oder Ratings etc.) und zu den Charakteristika der zugrundeliegenden Produktstruktur

und liefert als Ergebnis der Analyse die Schätzungen der Teilnutzenwerte. Diese sind so skaliert, daß pro Attribut ihre Summe Null ergibt und daß sich die relative prozentuale Wichtigkeit eines Attributs direkt aus dem Range seiner Teilnutzenwerte ablesen läßt (d.h. die Ranges aller Attribute summieren sich zu 100, ein Range von 10,5 bedeutet dann, daß die relative Wichtigkeit dieses Attributs 10,5% beträgt).

2.2 Auswertung

Das experimentelle Design der vorliegenden Untersuchung legt eine varianzanalytische Auswertung nahe. Dabei allerdings steckt sozusagen der Teufel im Detail — und dieses besonders hinsichtlich der vorhandenen Computerprogramme, die hierfür verwendbar sind. Es ist deshalb notwendig, vor der eigentlichen Darstellung und Diskussion erzielter Ergebnisse der Auswertungsmethodik einen eigenen Abschnitt zu widmen.

2.2.1 Auswertungsmethodik

Bei der Auswertung der gewonnen CA-Ergebnisse stellt sich zunächst die Frage, wie die zu untersuchende Robustheit der Analysemethoden erfaßt werden soll. Es muß also ein Maß für die Qualität dieser Analyseergebnisse gefunden werden, welches den Vergleich der durchschnittlichen Qualität unter den verschiedenen Bedingungskombinationen (experimentellen Zellen) erlaubt. Wie die in Abschnitt 2.1.1 besprochenen Untersuchungen zeigen, können hier unterschiedliche Wege eingeschlagen werden — es finden sich dort 4 verschiedene Maße: Produkt-Moment-Korrelationen zwischen vorhergesagten Werten $\hat{y}_j$ und „wahren“ Werten $\tilde{y}_j$, Rangkorrelationsmaße zwischen $\hat{y}$- und $\tilde{y}$-Werten, Streßwerte und die bei Cattin & Bliemel (1978) definierten $SAVE$-Werte. Es wurden bei der Besprechung der Untersuchung von Wittink & Cattin (vgl. S. 66ff) bereits kritische Anmerkungen zu den Produkt-Moment-Korrelationen und Streßwerten gemacht: Bei ersteren besteht ein Vorbehalt hinsichtlich der Skalenqualität der $\hat{y}$-Werte, bei letzteren ist die Verteilungsform zumindest bei nichtmetrischer Analyse unklar. Wenn man es ganz genau nimmt, dann muß man ähnliche Vorbehalte auch gegenüber den $SAVE$-Werten haben: In der Berechnungsformel (S. 61) sind auch hier die Koeffizientenschätzungen als Summanden enthalten, deren Intervallskalenqualität bei einer CA mit Rangwerten als abhängiger Variable nicht sicher ist, so daß diese Unsicherheit auch für die $SAVE$-Werte gelten muß. Die Vorsicht bezüglich der Skalenqualität der Koeffizientenschätzungen wiegt einerseits zwar nicht besonders schwer, es ist durchaus zu rechtfertigen, sie einfach fallenzulassen (vgl. Teil 1, Fußnote

12), andererseits aber kann man sie bei der Auswahl eines Qualitätsmaßes für die CA-Ergebnisse dann walten lassen, wenn ein solches Maß gefunden ist, das keine derartigen Probleme bereitet. Letzteren Vorteil bieten Rangkorrelationen, die zudem — wie schon bei der Besprechung der Studie von Wittink % Cattin gesagt — inhaltlich sinnvoll sind, da sie erfassen, wie gut auf der Basis der empirisch erhobenen Rangwerte der „wahre Rang“ eines Stimulus vorhergesagt wird.

Für die vorliegende Untersuchung wurde also die Spearman-Rangkorrelation zur Erfassung der Robustheit verwendet[46]: Es wurden für jeden einzelnen Fall i ($i = 1, 2, \ldots, 100$) in jeder der 90 Zellen des experimentellen Designs die 27 „wahren“ Werte $\tilde{y}_j$ (vgl. $\boxed{2}$ in Abb. 2.5) des vollen faktoriellen Designs mit den 27 aus den Koeffizientenschätzungen vorhergesagten Werten $\hat{y}_j$ korreliert. Man erhält so pro Zelle des experimentellen Designs 100 Korrelationswerte r_i (der Index i entspricht der in Teil 1 verwendeten Notation). In Anhang B sind — ungeachtet der Einwände gegen solche Mittelwertbildung, und sozusagen lediglich zur Veranschaulichung — die Mittelwerte der Korrelationen für die 90 Zellen des experimentellen Designs abgedruckt.

Die Korrelationswerte r_i wurden dann in Fisher's-Z-Werte Z_i transformiert[47]. Die Gründe dafür wurden bereits bei der Besprechung der Studie von Wittink & Cattin (S. 66ff) genannt: Die Z-Werte sind intervallskaliert, so daß ihre Zellenmittelwerte als sinnvolle Größen berechnet werden können.

Mit den Z-Werten als abhängigen Variablen kann dann zur Überprüfung der Effekte der spezifizierten experimentellen Faktoren eine Varianzanalyse unternommen werden. Da bei der Datengenerierung die ursprünglichen Teilnutzenwerte β_{lk} nur einmal für 100 Fälle gezogen (Schritt $\boxed{1}$ in Abbildung 2.5) und die Daten in jeder experimentellen Zelle auf der Grundlage dieser Werte generiert wurden, sind die Messungen der abhängigen Variable (Z-Werte) in den verschiedenen Zellen als Meßwiederholungen zu betrachten: Es ist so als ob jeder einzelne Fall wiederholt — unter jeder der insgesamt 90 experimentellen Bedingungen — getestet worden wäre. Es muß also eine Varianzanalyse mit kompletter Meßwiederholung über alle Zellen hinweg durchgeführt werden.

Meine Erläuterungen treffen an dieser Stelle auf das Problem, daß eine einigermaßen gründliche Darstellung dieser Auswertungsmethodik — der ANOVA im allgemeinen und der Meßwiederholungsanalyse im besonderen, nebst der komplizierten statistischen Tests, die hier zur Testung der Voraussetzungen

[46] Der Vollständigkeit halber die Formel zur Berechnung dieser Korrelation:
$r_{\tilde{y}\hat{y}} = 1 - \frac{6\sum_j d_j^2}{n(n^2-1)}$, wobei d_j die Differenz der Rangplätze von $\tilde{y}_j$ und $\hat{y}_j$ ist (nach Bortz 1979, S. 284).

[47] Auch hier der Vollständigkeit halber die Berechnungsformel:
$Z_i = \frac{1}{2} \ln \left(\frac{1+r_i}{1-r_i} \right)$ (Nach Bortz 1979, S. 260.)

durchzuführen sind, — wohl den Rahmen der Darstellung einer Untersuchung sprengen würde. Der Aufwand würde sicherlich denjenigen, der z.B. für LINMAP getrieben wurde, noch weit übertreffen — was angesichts des Umstands, daß die Varianzanalyse eigentlich nicht thematischer Gegenstand der vorliegenden Untersuchung ist, kaum angemessen wäre. Dagegen scheint es verlockend, diesen Aufwand unter Hinweis auf entsprechende Darstellungen in der Fachliteratur[48] ganz und gar zu vermeiden und dieses mit dem Argument zu begründen, das schon eingangs der Besprechung der OLS-Regression (Abschnitt 1.2.2) angeführt wurde und das für die Varianzanalyse erst recht zutrifft, nämlich daß diese Methode sozusagen zu den Standards der sozialwissenschaftlichen Methodenlehre gehört und ihre Bekanntheit deshalb vorausgesetzt werden kann. Ich wähle im folgenden einen Mittelweg, der die grundsätzliche Vertrautheit mit der varianzanalytischen Methode voraussetzt und nur möglichst knapp die Besonderheiten beleuchtet, die sich für die Anwendung im vorliegenden Fall ergeben.

„In Versuchsplänen ohne Meßwiederholung wird die Fehlervarianz durch die Unterschiedlichkeit der unter einer Faktorstufe (Faktorstufenkombination) beobachteten Vpn bestimmt“ (Bortz 1979, S. 407). Die Besonderheit der Varianzanalyse mit Meßwiederholung ergibt sich nun daraus, daß jeweils dieselben a priori Unterschiede zwischen Versuchspersonen die Messungen unter allen Faktorstufenkombinationen (derjenigen Faktoren, für die wiederholte Messungen durchgeführt werden) beeinflussen, so daß die Fehlerkomponenten zwischen den einzelnen Stichproben nicht voneinander unabhängig sind, wodurch eine grundlegende Bedingung zur Durchführung der üblichen F-Tests, mit denen varianzanalytische Effekte auf Signifikanz überprüft werden, verletzt ist (vgl. Bortz 1979, S. 344). Das Grundprinzip der Varianzanalyse mit Meßwiederholung (über alle Faktorenstufenkombinationen) läßt sich nun am einfachsten so beschreiben: Es wird eine Varianzanalyse durchgeführt, die außer den experimentellen Faktoren noch einen weiteren Faktor „Versuchsperson“ als Faktor *mit zufälligen Effekten* enthält (im folgenden mit V bezeichnet). Betrachtet man die experimentellen Faktoren als Faktoren mit festen Effekten, dann resultiert als Prüfvarianz für die Varianzen zu Lasten der zu untersuchenden (Haupt- und Interaktions-) Effekte (der experimentellen Faktoren) jeweils die Varianz zu Lasten der Interaktion des jeweiligen Effekts mit dem Faktor V (vgl. dazu Bortz 1979, S. 546ff), d.h. der F-Bruch, mit dem ein Effekt getestet wird, enthält im Zähler die geschätzte Varianz zu Lasten dieses Effekts (aus dem Quotienten der entsprechenden Quadratsumme durch die entsprechenden Freiheitsgrade) und im Nenner die geschätzte Varianz zu Lasten der Inter-

[48] Ich stütze mich bei meinen weiteren Ausführungen v.a. auf die Darstellungen von Bortz (1979, S. 293-576). Bei Winer (1971) und bei Glaser (1978) finden sich ebenfalls ausführliche Behandlungen, während Eimer (1978) eine etwas knappere Einführung anhand konkreter Beispiele gibt.

Für die Effekte		Für die Prüfvarianzen		F-Bruch
Quadratsumme	df	Quadratsumme	df	$(s_X^2 = SS_X/df_X)$
Varianz „innerhalb“ Versuchspersonen:				
SS_A	4	$SS_{A\times V}$	396	$s_A^2/s_{A\times V}^2$
SS_B	2	$SS_{B\times V}$	198	$s_B^2/s_{B\times V}^2$
SS_C	2	$SS_{C\times V}$	198	$s_C^2/s_{C\times V}^2$
SS_D	1	$SS_{D\times V}$	99	$s_D^2/s_{D\times V}^2$
$SS_{A\times B}$	8	$SS_{A\times B\times V}$	792	$s_{A\times B}^2/s_{A\times B\times V}^2$
$SS_{A\times C}$	8	$SS_{A\times C\times V}$	792	$s_{A\times C}^2/s_{A\times C\times V}^2$
$SS_{A\times D}$	4	$SS_{A\times D\times V}$	396	$s_{A\times D}^2/s_{A\times D\times V}^2$
$SS_{B\times C}$	4	$SS_{B\times C\times V}$	396	$s_{B\times C}^2/s_{B\times C\times V}^2$
$SS_{B\times D}$	2	$SS_{B\times D\times V}$	198	$s_{B\times D}^2/s_{B\times D\times V}^2$
$SS_{C\times D}$	2	$SS_{C\times D\times V}$	198	$s_{C\times D}^2/s_{C\times D\times V}^2$
$SS_{A\times B\times C}$	16	$SS_{A\times B\times C\times V}$	1584	$s_{A\times B\times C}^2/s_{A\times B\times C\times V}^2$
$SS_{A\times B\times D}$	8	$SS_{A\times B\times D\times V}$	792	$s_{A\times B\times D}^2/s_{A\times B\times D\times V}^2$
$SS_{A\times C\times D}$	8	$SS_{A\times C\times D\times V}$	792	$s_{A\times C\times D}^2/s_{A\times C\times D\times V}^2$
$SS_{B\times C\times D}$	4	$SS_{B\times C\times D\times V}$	394	$s_{B\times C\times D}^2/s_{B\times C\times D\times V}^2$
$SS_{A\times B\times C\times D}$	16	$SS_{A\times B\times C\times D\times V}$	1584	$s_{A\times B\times C\times D}^2/s_{A\times B\times C\times D\times V}^2$
Varianz „zwischen“ Versuchspersonen:				
SS_V	99	entfällt		entfällt

Tabelle 2.4: Quadratsummen, Freiheitsgrade und zugehörige Prüfgrößen bei der durchgeführten Meßwiederholungsanalyse.

aktion des Effekts mit V. Bei dieser Varianzanalyse wird die Varianz also in eine Varianz „zwischen den Versuchspersonen“ (Varianz zu Lasten des Faktors V) und in eine Varianz „innerhalb der Versuchspersonen“ aufgeteilt. Erstere interessiert nicht, ihre Elimination ist der eigentliche Sinn der Meßwiederholungsanalyse: Es werden individuell konstant wirkende a priori Unterschiede zwischen den Versuchspersonen eliminiert. Die Varianz „innerhalb“ wird weiter unterteilt in Varianzen zu Lasten der interessierenden Effekte und in eine Residualvarianz, welche auf individuell unterschiedliche Reaktionsweisen auf die variierten experimentellen Bedingungen (Interaktionen des Faktors V mit den Effekten der experimentellen Faktoren) und auf unsystematische Zufallseinflüsse zurückzuführen ist[49].

Tabelle 2.4 gibt für die hier thematische Varianzanlyse mit Meßwiederholung

[49]Diese letztere Varianzquelle findet in gebräuchlichen Darstellungen der Meßwiederholungsanalyse — z.B. Bortz (1979, S. 407ff), Eimer(1978, S. 188ff) — kaum Beachtung, d.h. ich habe keine Darstellung gefunden, in der die darauf zurückführbare Varianz eigens ermittelt worden wäre. Dies mag daran liegen, daß zur Bestimmung der F-Brüche zur Signifikanzprüfung nur die jeweiligen Interaktionen mit V interessieren.

über die Faktoren A, B, C, D die Quadratsummen der zu untersuchenden Effekte, deren Freiheitsgrade, die zugehörigen Quadratsummen zur Schätzung der Prüfvarianz und deren Freiheitsgrade und die F-Brüche für die Effekte wieder. Im Anhang C wird ausführlich gezeigt, wie die Quadratsummen berechnet werden können. Mit Tabelle 2.4 und Anhang C soll allerdings lediglich gezeigt werden, wie die „klassische“ Quadratsummenbildung, die in den genannten Darstellungen der Varianzanalyse beschrieben wird, im vorliegenden Fall vorzunehmen wäre. Tatsächlich wurden die Quadratsummen und F-Brüche natürlich nicht nach den angegebenen Formeln „von Hand“ berechnet, sondern die Varianzanalyse wurde vermittels der SAS-Prozedur GLM (SAS Institute 1989, S. 891ff) durchgeführt und dieses Programm analysiert Daten im Rahmen des allgemeinen linearen Modells (ALM). Die Varianzanalyse wird so also nicht über den „klassischen“ Algorithmus, sondern als Submodell des ALM berechnet, d.h. die univariaten Ergebnisse (siehe unten) resultieren aus einer OLS-Schätzung. Eine ausführliche Darstellung dieser Vorgehensweise liefert Werner (1993), die Meßwiederholungsanalyse wird dort im Kapitel 7.3 behandelt. Man hätte im vorliegenden Fall die Varianzanalyse z.B. auch mit der SAS-Prozedur ANOVA durchführen können, wodurch der „klassische“ Algorithmus zu Anwendung gelangt wäre: Die Ergebnisse wären diesselben geblieben — es wurde schon auf S. 41 darauf hingewiesen, daß beide Vorgehensweisen gleichwertig sind. Für den Anwender bestehen hier praktisch gar keine Unterschiede, die SAS-Statements, die er eingeben muß, sind bis auf eben den Namen der Prozedur genau dieselben.

Als Voraussetzung für die Gültigkeit der F-Tests in einer Varianzanalyse mit Meßwiederholung wird in den meisten Darstellungen (z.B. Bortz 1979, S. 437ff; Winer 1971, S. 594ff) die Homogenität der (zum jeweiligen F-Test, d.h. zum Effekt, den er testet, gehörigen) Varianz-Kovarianz-Matrix angegeben. Gemeint ist: Unter jeder Stufe eines Faktors, über den die Meßwiederholung läuft, müssen die Meßwerte dieselbe Varianz aufweisen und zwischen jedem Paar von Stufen dieses Meßwiederholungsfaktors muß dieselbe Kovarianz der Meßwerte bestehen. Die Voraussetzung wird mit dem Box-Test überprüft, der eine ermittelte Varianz-Kovarianz-Matrix gegen die Nullhypothese ihrer Homogenität testet. Werner (1993, S. 448: Kap. 7.3.1) weißt darauf hin, daß diese Voraussetzung bzw. der Box-Test zu restriktiv ist: Es genügt, daß für jedes Paar von Stufen eines Meßwiederholugsfaktors die Varianz der Differenzen der Meßwerte aus beiden Stufen gleich ist. Sind z.B. im vorliegenden experimentellen Design mit a und a' zwei unterschiedliche Stufen des Faktors A benannt, dann ergibt sich diese Varianz der Differenzen $Z_{ia} - Z_{ia'}$ aus: $\sigma^2_{Z_a - Z_{a'}} = \sigma^2_{Z_a} + \sigma^2_{Z_{a'}} - 2\sigma_{Z_a Z_{a'}}$. Für alle $a \neq a'$ muß also $\sigma^2_{Z_a - Z_{a'}}$ einen konstanten Wert annehmen, d.h. die Einträge $\sigma^2_{Z_a}, \sigma^2_{Z_{a'}}, \sigma_{Z_a Z_{a'}}$ in der Varianz-Kovarianz-Matrix müssen lediglich so beschaffen sein, daß für jedes Stufenpaar a, a' die angegebene Bedingung erfüllt ist. Diese Bedingung wird meist als Huynh-Feldt-Bedingung oder als h.o.t.d.v.-

Annahme („*h*omogeneity *o*f *t*reatment *d*ifference *v*ariances“) bezeichnet und die Varianz-Kovarianz-Matrizen, die diese erfüllen, werden Typ-H-Matrizen genannt. Die Huynh-Feldt-Bedingung muß also für die Varianz-Kovarianz-Matrix eines jeden Meßwiederholungsfaktors und einer jeden Kombination von Meßwiederholungsfaktoren erfüllt sein, damit die zugehörigen F-Tests gelten.

Als Alternative zur Überprüfung der besprochenen Voraussetzung wurden Freiheitsgradkorrekturen für die F-Tests von Meßwiederholungsfaktoren bzw. -faktorkombinationen entwickelt: Ist die h.o.t.d.v-Annahme nicht erfüllt, dann ist zu erwarten, daß die F-Tests progressiv werden, d.h. daß das „tatsächliche“ α-Niveau höher liegt als dasjenige, welches durch den Test gesetzt wird. Da nämlich die Messungen unter den einzelnen Faktorstufen abhängig sind, besitzen sie (bzw. die entsprechenden Quadratsummen) weniger Freiheitsgrade, als die bei der Bildung der F-Brüche ermittelten (vgl. Tabelle 2.4) — man kann dieses dadurch korrigieren, daß man zwar die F-Brüche wie gewohnt berechnet, die resultierenden F-Werte allerdings an einer F-Verteilung mit nach unten korrigierten Zähler- und Nennerfreiheitsgraden testet. Eine solche Korrektur unternimmt der sogenannte konservative F-Test nach Geisser & Greenhouse (vgl. z.B. Bortz 1979, S. 445; Werner 1993, S. 447: Kap. 7.3.1), der, wie der Name schon sagt, sehr konservativ ist, also durch Reduktion der Freiheitsgrade das α-Niveau soweit nach oben korrigiert, daß durch Anwendung dieses Tests die Überprüfung der obigen Voraussetzung überflüssig wird. Die Konservativität dieses Tests bringt natürlich den Nachteil geringer Effizienz mit sich — die Meßlatte für Signifikanz eines Effekts liegt sozusagen sehr hoch. Weniger konservativ, dafür aber auch unsicherer in Bezug auf die zu treffenden Entscheidungen, ist die ϵ-Korrektur der Freiheitsgrade (vgl. wieder Bortz ebd.: Gleichung (9.30); Werner ebd.: Gleichung (7.3.1)): Die Freiheitsgrade werden mit einem Korrekturfaktor ϵ multipliziert, der so berechnet ist, daß er im Falle vollständiger Unabhängigkeit der Meßwerte in den verschiedenen Stufen des Meßwiederholungsfaktors gleich 1 wird und im Falle vollständiger Abhängigkeit zu derselben Korrektur führt, wie der konservative F-Test.

Es wird nun im allgemeinen ein gestuftes Vorgehen bei der Entscheidung über Signifikanz bzw. Nichtsignifikanz eines Effekts vorgeschlagen. Bortz (1979, S. 446) z.B. empfiehlt zunächst den konservativen F-Test, im Falle von dessen Nichtsignifikanz die ϵ-Korrektur und im Falle erneuter Nichtsignifikanz den unkorrigierten F-Test — sollte dieser dann ebenfalls nicht signifikant sein, dann wird die H_1 verworfen (d.h. die Nichtsignifikanz des Effekts akzeptiert), im Falle seiner Signifikanz wäre dann die Voraussetzung bezüglich der Varianz-Kovarianz-Matrix zu überprüfen (bei Bortz mit dem Box-Test), wird dieser Test nicht signifikant, gilt die Voraussetzung als erfüllt, wird er signifikant, kann aufgrund des F-Tests keine Entscheidung getroffen werden: In diesem Fall sollte der Effekt durch einen multivariaten Test der Mittelwerte der einzelnen Faktorstufen (bzw. Faktorstufenkombinationen) — z.B.

durch Hotellings-T^2-Test — auf Signifikanz getestet werden bzw. es kann dann eine multivariate Varianzanalyse, wobei jede Wiederholungsmessung jeweils eine abhängige Variable bildet, gerechnet werden. Bei solchen multivariaten Testungen wird der Korreliertheit der Meßwerte in den Meßwiederholungen von vorneherein Rechnung getragen, so daß keine zusätzlichen Bedingungen wie die obige h.o.t.d.v.-Annahme überprüft werden müssen. Um eine gewisse Ausuferung meiner Ausführungen zu vermeiden, verzichte ich hier auf weitere Darlegungen zu diesen multivariaten Verfahren und verweise auf entsprechende Beschreibungen in der Literatur (z.B. Bortz 1979, S. 697ff).

Es lag nahe, für die vorliegende Untersuchung hinsichtlich der Entscheidung über Signifikanzen der Effekte die Vorgehensweise zu wählen, die im „User's Manual“ für die SAS-Prozedur GLM vorgeschlagen wird (SAS Institute 1989, S. 954). Die SAS Prozeduren GLM und ANOVA produzieren standardmäßig bei Spezifikation von Meßwiederholungsfaktoren univariate und multivariate Tests. Die ersteren sind der „normale“ F-Test, die konservative Freiheitsgradkorrektur nach Geisser & Greenhouse (in der Ausgabe als „Greenhouse-Geisser-Epsilon“ bezeichnet) und die Freiheitsgradkorrektur um das oben erwähnte ϵ (hier als „Huynh-Feldt Epsilon“ benannt). Es werden insgesamt 4 multivariate Tests ausgegeben, der wohl bekannteste unter ihnen ist Wilks Λ (siehe dazu Glaser 1978, S. 276: Gleichung (7.47)). Abrufbar ist ein „sphericity test“ mit dem die Nullhypothese, daß es sich bei der zu einem Effekt gehörigen Varianz-Kovarianz-Matrix um eine Typ-H-Matrix handelt, überprüft werden kann. Der folgende Entscheidungsablauf wird vorgeschlagen: Ist der Sphärizitätstest nicht signifikant, dann können die normalen univariaten F-Tests verwendet werden, im Falle der Signifikanz sollten die Epsilon-Korrekturen herangezogen werden — solange diese Signifkanz nicht allzu drastisch ausfällt: *„ . . . in cases where the sphericity test is dramatically rejected ($p \leq 0.0001$) all these univariate tests should be interpreted cautiously“* (SAS Institute 1989, ebd.). In letzerem Falle empfiehlt es sich also, nur noch die multivariaten Ergebnisse zu verwenden.

Ein besonderes Problem bilden bei der vorliegenden Auswertung Einzelvergleiche von Mittelwerten der Faktorstufen bzw. Faktorstufenkombinationen. Es wurde im Abschnitt 2.1.3 bereits darauf hingewiesen, daß sich hinsichtlich des Faktors A bereits a priori Vergleiche ergeben: Es interessiert zum einen der Vergleich zwischen Datengenerierung mit dem kompensatorischen Modell (A1) und Datengenerierung durch Modelle mit dominantem Attribut (A2, A3), zum anderen der zwischen Dominant-Attribut-Bedingung und Interaktionsbedingungen (A4, A5). Nun rechnet zwar — wie schon gesagt — die SAS-Prozedur GLM die Varianzanalyse als Submodell des ALM und enthält damit grundsätzlich die Möglichkeit, die **L**-Matrix der allgemeinen Linearhypothese so zu konstruieren, daß durch die Partialhypothesen interessierende Kontraste berechnet werden, womit es, wie Werner (1993, S. 293) schreibt, möglich wird, *„das*

Problem multipler Testungen wesentlich zu entschärfen". Aber das zur Spezifizierung spezieller Kontraste notwendige CONTRAST-Statement bezieht sich immer auf solche Effekte, die im MODEL-Statement aufgeführt werden, und d.h. auf Effekte von Faktoren *ohne* Meßwiederholung[50] (vgl. SAS Institute 1989, S. 891ff.). Für die im REPEATED-Statement definierten Meßwiederholungsfaktoren besteht nur eine eingschränkte Auswahl von orthogonalen Kontrasten, die vorab spezifiziert werden können. Genauso verhält es sich auch mit dem Abruf der in den Textbüchern zur Varianzanalyse normalerweise behandelten Tests für (a-priori- oder a-posteriori-) Einzelvergleiche mittels des TEST-Statements. Eine Alternative böte hier vielleicht die BMDP-Prozedur 4V (Dixon 1988, S. 1045ff), die eine Spezifizierung von Kontrasten für Meßwiederholungsfaktoren erlaubt. Die Darstellung der Prozedur im Manual ist allerdings etwas knapp gehalten, so daß es zusätzlicher Dokumentation bedurft hätte, um zu verstehen, wie das vorliegende experimentelle Design mit dieser Prozedur optimal zu bewältigen ist, weswegen sie nicht verwendet wurde. Im übrigen scheinen alle in Augenschein genommenen BMDP-Prozeduren, die Varianzanalysen mit Meßwiederholung erlauben (2V, 3V, 4V, 5V, vgl. Dixon 1988), genausowenig wie die SAS-Prozeduren in der Lage, Tests für Einzelvergleiche bei Meßwiederholungsfaktoren durchzuführen — es hat den Anschein, als ob die Defizite der Computerprogramme (bzw. zumindest bei deren Dokumentation) hinsichtlich der Testung von Einzelvergleichen bei Meßwiederholung mit einer gewissen Unklarheit einhergehen, die in den (von mir zu Rate gezogenen) Textbüchern zur Varianzanalyse bezüglich dieses Themas herrscht. Dort nämlich wird das Thema ebenfalls kaum ausführlicher behandelt.

Zum Glück sind die Verfahren zum Test von a-posteriori-Vergleichen durchaus auch „von Hand" durchführbar: Ist die Varianznalyse durchgeführt und sind damit die für diese Tests notwendigen Kennwerte berechnet, dann liefert z.B. die in der SUGI-Bibliothek enthaltene Prozedur DUNCAN, welche unter der SAS-Version 5 abrufbar ist, die meisten der fraglichen Tests (SAS Institute 1986, S. 47ff). Auf Grund der Empfehlungen von Werner (1993, S. 293ff: Kap. 4.10) fiel die Wahl des Testverfahrens auf Tukey's HSD („*H*onestly *S*ignifikant *D*ifference"). Dieser Test beruht — wie auch der bekanntere Newman-Keuls-Test — auf der q-Statistik („studentized range statistic"), die im vorliegenden

[50] Jedenfalls geht aus dem SAS-Manual nicht hervor, wie das CONTRAST-Statement auf Meßwiederholungsfaktoren bezogen werden könnte, und auch diverse Versuche nach dem Trial-and-Error-Prinzip, dieses zu erreichen, erwiesen sich als fruchtlos. Eine Möglichkeit könnte vielleicht darin bestehen, Kontraste mit Meßwiederholung vermittels des MANOVA-Statements und der Option „M=" zu spezifizieren, dieses erschien mir aber durch das Manual nur unzureichend dokumentiert, so daß bei diesem Vorgehen das Gefühl geblieben wäre, auf theoretisch sehr unsicherem Grund zu stehen.

Fall die folgende Form annimmt:

$$q = \frac{\bar{Z}_{\text{groß}} - \bar{Z}_{\text{klein}}}{\sqrt{s^2_{\text{res}}/m}}$$

Dabei ist $\bar{Z}_{\text{groß}}$ der größere und $\bar{Z}_{\text{klein}}$ der kleinere der beiden zu vergleichenden Mittelwerte, s^2_{res} ist die Prüfvarianz des Effekts (vgl. Tabelle 2.4), dessen Mittelwerte verglichen werden und m ist hier die Mittelwertsbildung zugrundeliegende Stichprobengröße. Die Stichprobenverteilung eines q-Wertes ist (näherungsweise) die Range-Verteilung („studentized range distribution“) mit den Parametern r und df. Letzteres sind die Freiheitsgrade der Prüfvarianz und r ist der Abstand in der Rangreihe zwischen den beiden Mittelwerten $\bar{Z}_{\text{groß}}$ und $\bar{Z}_{\text{klein}}$: Würde z.B. für die Mittelwerte der 5 Stufen des Faktors A gelten $\bar{Z}_{A1} < \bar{Z}_{A2} < \bar{Z}_{A3} < \bar{Z}_{A4} < \bar{Z}_{A5}$, dann betrüge der Abstand r zwischen $\bar{Z}_{A1}$ und $\bar{Z}_{A5}$ 5, der zwischen $\bar{Z}_{A1}$ und $\bar{Z}_{A4}$ 4, der zwischen $\bar{Z}_{A2}$ und $\bar{Z}_{A5}$ ebenfalls 4 usw.[51]. Beim Tukey-HSD-Test wird nun als kritischer Wert für jeden möglichen Vergleich zweier Mittelwerte (innerhalb eines Effekts) der kritische q-Wert bei maximalem Abstand r herangezogen, d.h. r wird gleich der Anzahl der jeweiligen Faktorstufen bzw. der jeweiligen Faktorstufenkombinationen gesetzt und derjenige q-Wert gesucht, der die oberen 5% oder 1% der Range-Verteilung für dieses r abschneidet. Eine Tabelle mit kritischen q-Werten für verschiedene r und df enthält Winer (1971, S. 870f) und Glaser (1978, S. 328f). Auf diese Weise läßt sich eine kritische Mittelwertsdifferenz berechnen: $D_{\text{krit}} = q_{\text{krit}}\sqrt{s^2_{\text{res}}/n}$. Eine ausführlichere Darstellung der q-Statistik gibt und ihrer Anwendung bei Mittelwertsvergleichen gibt Winer (1971, S. 185ff; der HSD-Test ist auf S. 198 dargestellt, die Anwendung bei Meßwiederholung wird am Beispiel auf S. 528 gezeigt).

Abschließend sei noch einmal darauf hingewiesen, daß die durchgeführte Varianzanalyse mit Meßwiederholung die Meßwiederholungsfaktoren als solche mit festen Effekten behandelt. Dieses ist genaugenommen nicht korrekt, da bei den Faktoren A, B, C die ausgewählten Bedingungen nicht alle realisierbaren Möglichkeiten der experimentellen Faktoren darstellen[52]. Die Unkorrektheit ergab sich infolge der schon erwähnten Schwächen der SAS-Prozedur(en)

[51] Allgemein: $r = \text{Rangplatz}(\bar{Z}_{\text{groß}}) - \text{Rangplatz}(\bar{Z}_{\text{klein}}) + 1$.

[52] Die Betrachtung von D als Faktor mit festen Effekten ergibt sich aus einer durch den Stand der Forschung bedingten vorab gegebenen Einschränkung des Interesses an verschiedenen CA-Methoden auf den Vergleich von LINMAP und OLS, d.h. die beiden Verfahren wurden sozusagen nicht als Repräsentanten der Gruppen metrischer bzw. nichtmetrischer Methoden ausgewählt, sondern um ihrer selbst willen. Eventuell könnte auch C als Faktor mit festen Effekten angesehen werden, insofern die beiden ausgewählten reduzierten faktoriellen Designs die einzigen mir bekannten Möglichkeiten der orthogonalen Reduktion eines vollen 3^3-Designs darstellen. Auch wenn mir vielleicht weitere Reduktionsmöglichkeiten entgangen sind, so dürften es derer doch nicht viele sein.

bezüglich Meßwiederholungen: Auch das RANDOM-Statement, das die Einrichtung von Faktoren mit zufälligen Effekten erlaubt, kann nur auf solche Faktoren bezogen werden, die zuvor mit dem MODEL-Statement spezifiziert wurden. Eine Bewältigung auch von Meßwiederholungsfaktoren mit zufälligen Effekten verspricht die BMDP-Prozedur 3V (Dixon 1988, S. 1025ff): Bei näherem Hinsehen zeigt sich allerdings, daß dieses einfach dadurch geschieht, daß eine „normale“ Varianzanalyse ohne gesondert spezifizierte Meßwiederholungen gerechnet, dabei jedoch der Faktor Versuchsperson als solcher mit zufälligen Effekten gebildet und jeder Effekt mit Meßwiederholung unter dem Versuchspersonenfaktor geschachtelt wird. Eine solche Varianzanalyse enthält dann außer den eigentlich interessierenden Effekten noch die Schachtelungseffekte, durch deren Quadratsummen die jeweiligen Residualquadratsummen erfaßt werden. Werner (1993, S. 455: Kap. 8.3) beschreibt dieses Vorgehen näher und zeigt, daß es natürlich auch mit der Prozedur GLM bewerkstelligt werden kann. Dabei können dann im Prinzip die Faktoren bzw. die im MODEL-Statement enthaltenen Effekte nach Belieben als „fixed“ oder „random“ deklariert werden (wobei allerdings auf erwünschten Output, nämlich die Freiheitsgradkorrekturen und multivariaten Ergebnisse, verzichtet werden muß). Um die Analyse abzurunden, wurde sowohl mit GLM, als auch mit der Prozedur 3V diese Varianzanalyse mit zufälligen Effekten der Faktoren A,B, D durchgeführt — ohne Ergebnis, denn das vorliegende varianzanalytische Design erforderte dafür Speicherkapazitäten, die auf den verfügbaren Rechenanlagen nicht voranden waren[53]. Auch hier scheinen wieder die Unzulänglichkeiten der EDV-Programme sozusagen mit weißen Flecken in der varianzanalytischen Literatur einherzugehen, auch das Thema Meßwiederholung mit zufälligen Effekten findet in der zu Rate gezogenen Literatur nur knappe Beachtung.

Es ist wohl angesichts dieser Schwierigkeiten vertretbar, die Unkorrektheit in Kauf zu nehmen und die Faktoren A, B, C, D als solche mit festen Effekten zu behandeln. Für feste Effekte gilt: *„Die Prüfung der generellen Nullhypothese führt zu einer Generalisierung auf genau die Ausprägungen von B* [das meint hier die unabhängige Variable], *die in der Untersuchung enthalten sind“* (Glaser 1978, S. 83). Dagegen gilt für zufällige Effekte: *„Man darf jetzt das Ergebnis des Signifikanztests auf die gesamte unabhängige Variable und nicht nur auf die in der Untersuchung vorkommenden Bedingungen generalisieren“* (ebd.). Die Entscheidung über feste oder zufällige Effekte betrifft also die Interpretation bzw. den Geltungsbereich der Ergebnisse, Generalisierungen auf andere als die behandelten Bedingungen sind genaugenommen nicht inferenzstatistisch abgesichert. Nun könnte man aber argumentieren, daß solche Generalisierun-

[53]Dies gilt einschließlich der Großrechenanlage des Heidelberger Universitätsrechenzentrums: Dort wurden die Programme mit Rückmeldungen über die notwendige Speicherkapazität abgebrochen, die bei mir allerdings gewisse Zweifel hinterließen, es handelte sich jeweils um ungefähr 70MB!

gen nicht das vordringlichste Ziel der durchgeführten Untersuchung darstellen: Es geht letztlich darum, Licht in das Dunkel mehr oder weniger vager Vorstellungen zu bringen, welche bezüglich der Robustheit der beiden untersuchten Methoden der CA bestehen. Zum einen münden diese Vorstellungen in die Hypothesen, die am Ende des Abschnitts 2.1.2 spezifiziert wurden. Sind die theoretischen Vorstellungen richtig, dann wird sich die abhängige Variable unter den speziellen getesteten Bedingungen entsprechend der Erwartungen verhalten. Der so zunächst interessierende Schluß ist der von der allgemeinen Erwartung auf den speziellen Fall und nicht der vom speziell getesteten Fall auf die allgemeine Regel. Oder anders ausgedrückt: Es geht um die mögliche Widerlegung der theoretischen Vorstellungen durch die getesteten Bedingungen. Zum anderen werden natürlich auch bis zu einem gewissen Maße generalisierende Schlußfolgerungen angestrebt: Da die Einflüsse, denen die Robustheit der CA unterworfen ist, zum Teil theoretisch nur schwer faßbar sind, liegt der Sinn einer Simulationsstudie auch darin, vermittels der Daten diese Einflüsse abzutesten, um daraus ad hoc theoretische Vorstellungen abzuleiten — hierbei wäre dann die statistische Absicherung generalisierender Aussagen erwünscht. Jedoch ist dabei zu bedenken, daß bei solchen Ad-hoc-Erklärungen prinzipiell eine gewisse Vorsicht angebracht ist (vgl. z.B. Opp 1976, S. 148ff) — im vorliegenden Fall sollte diese Vorsicht darin bestehen, nachträgliche Interpretationen der Ergebnisse nur in explorativer, hypothesengenerierender Weise vorzunehmen. Bei dieser sowieso schon vorsichtigen Interpretationsweise — so könnte man argumentieren — sind dann auch die vorsichtigen Schlußfolgerungen, die sich aus den Ergebnissen zu festen Effekten ziehen lassen, zu gebrauchen und angemessen.

2.2.2 Ergebnisse

In Tabelle 2.5 sind die Ergebnisse der uni- und multivariaten Tests der geprüften Effekte der Meßwiederholungsanalyse zusammengefaßt. Auch die Resultate des bei der SAS-Prozedur GLM abrufbaren Sphärizitätstests werden dort wiedergegeben: Wie in Abschnitt 2.2.1 besprochen, dient dieser Test zur Entscheidung, welcher der von GLM gelieferten Tests der Effekte verwendet werden soll, das in dieser Hinsicht relevante Ergebnis ist in der Tabelle jeweils unterstrichen. Die detailliertere Darstellung der Ergebnisse erfolgt zunächst für diejenigen Effekte, für die am Ende des Abschnitts 2.1.2 Vorhersagen getroffen wurden.

Der Haupteffekt A: Die hohe Signifikanz des Effekts (in allen multi- und univariaten Tests) deckt sich mit der Vorhersage. Für die 5 Stufen A1 bis A5 ergaben sich die folgenden Mittelwerte:

A1	A2	A3	A4	A5
(kompensatorisch)	(schwach dominant)	(stark dominant)	(schwache Interaktion)	(starke Interaktion)
1.8373	1.8939	1.6561	1.8159	1.8629

Wie schon auf den ersten Blick zu sehen ist, kommt die Signifikanz des Effekts sozusagen nur zum Teil aus dem Grund zustande, aus dem sie vorhergesagt wurde: Zwar verschlechtert die Bedingung mit hochdominantem Attribut A3 die Güte der Ergebnisse, jedoch nicht die Modellmisspezifikation unter den Interaktionsbedingungen A4, A5. Tatsächlich sind bei den Tests der Mittelwertsvergleiche nach Tukey's HSD alle Vergleiche mit dem Mittelwert von A3 signifikant mit $p < 0.01$, sonst ist nur noch der Vergleich von A2 mit A4 signifikant mit $p < 0.05$. Auch wurden 4 orthogonale Kontraste dergestalt spezifiziert, daß jeweils die Bedingung A1 — also das kompensatorische Modell — mit jeder der 4 anderen Bedingungen kontrastiert wurde: Der Kontrast mit A2 ergab Signifikanz auf dem 5%-Level ($p < 0.04$), der mit A3 auf dem 0.1%-Level ($p < 0.0001$), würde hier das α-Niveau der multiplen Tests nach Bonferroni korrigiert, dann bliebe nur noch die Signifikanz des Kontrasts mit A3.

Der Hypothese bezüglich des Haupeffektes von A wird also durch die Daten nur teilweise entsprochen. Die völlige Widerlegung der Hypothese wird man aus diesen Ergebnissen wohl nicht ableiten können, denn eine einfache Erklärung des Ausbleibens signifikanter Mittelwertsdifferenzen zwischen kompensatorischer Bedingung und Interaktionsbedingungen liegt mit Blick auf die Datengenerierung auf der Hand: Das tatsächliche numerische Gewicht der „starken" Interaktion steht in keinem Verhältnis zu dem der „starken" Dominanz, denn den relativ hohen Beträgen, die zur Erzeugung der Dominanz zu den gezogenen Teilnutzenwerten β_{1k} addiert wurden (vgl. Abbildung 2.5, S. 83), stehen relativ kleine (wenn auch verdoppelte) Beträge gegenüber, die sich aus der Multiplikation der Teilnutzenwerte des ersten mit denen des zweiten Attributs ergeben. Folglich fallen hier die Interaktionen nicht so sehr ins Gewicht und möglicherweise hätten sich relevante Verschlechterungen der Analyseergebnisse unter den Interaktionen ergeben, wenn bei der Datengenerierung eine höhere „Interaktionsstärke" gewählt worden wäre. So erscheint das gewählte Ausmaß der Interaktion — welches v.a. nach dem Gesichspunkt der realistischen Erwartbarkeit in empirisch erhobenen Datensätzen festgelegt wurde — im nachhinein als untersuchungstechnisch ungünstig.

Der Haupteffekt B: Auch hier entspricht die hohe Signifikanz der Vorhersage. Die Mittelwerte lauten:

B1	B2	B3
(10% Fehler)	(20% Fehler)	(33% Fehler)
2.0844	1.8068	1.5485

Effekt	Sphärizität	F (univariat)	Wilks Λ (multivariat)
A	$p < 0.05$	$p < 0.001^\star$	$p < 0.001$
B	n.s.	$p < 0.001$	$p < 0.001$
C	$p < 0.0001$	$p < 0.001$	$p < 0.001$
D	entfällt	$p < 0.001$	entfällt
A×B	n.s.	$p < 0.001$	$p < 0.001$
A×C	$p < 0.0001$	$p < 0.001$	$p < 0.001$
A×D	$p < 0.01$	$p < 0.001^\star$	$p < 0.001$
B×C	$p < 0.0001$	$p < 0.05$	n.s.
B×D	$p < 0.05$	n.s.	n.s.
C×D	n.s.	$p < 0.001$	$p < 0.001$
A×B×C	$p < 0.05$	$p < 0.001^\star$	$p < 0.01$
A×B×D	$p < 0.0001$	$p < 0.001$	$p < 0.001$
A×C×D	$p < 0.0001$	$p < 0.001$	$p < 0.01$
B×C×D	$p < 0.01$	$p < 0.001^\star$	$p < 0.001$
A×B×C×D	$p < 0.0001$	$p < 0.05$	n.s.

$\star$: konservativer F-Test bzw. ϵ-Korrektur.

Tabelle 2.5: Ergebnisse der varianzanalytischen Tests.

Mit dem Anstieg der Fehlervarianz sinkt die durchschnittliche Güte der Analyseergebnisse. Alle möglichen Mittelwertsvergleiche sind nach Tukey's HSD signifikant ($p < 0.01$) und auch die F-Werte der beiden orthogonalen Kontraste, die spezifiziert wurden (B1 mit B2, B2 mit B3), zeigen jeweils Signifikanzen von $p < 0.001$).

Der Haupteffekt C: Auch hier ist die erwartete Signifikanz des Haupteffektes eingetroffen, allerdings in einer Deutlichkeit, die angesichts des in der Literatur vorherrschenden Tenors, wonach reduzierte faktorielle Designs bedenkenlos verwendet werden können, etwas überrascht: Sämtliche (uni- und multivariaten) Tests des Effekts waren hochsignifikant mit $p < 0.001$. Die Mittelwerte zeigen die erwartete Tendenz:

C1	C2	C3
(27 Stimuli)	(18 Stimuli)	(9 Stimuli)
2.0545	1.8568	1.5284

Auch hier sind alle möglichen Mittelwertsvergleiche nach Tukey's HSD signifikant ($p < 0.01$) und für die beiden spezifizierten orthogonalen Kontraste (C1 mit C2, C2 mit C3) gilt dasselbe wie für die des Haupteffekts von B, ihre F-Werte sind jeweils signifikant mit $p < 0.001$.

Der Haupteffekt D: Die hier erzielte hohe Signifikanz ($p < 0.001$) entspricht keinesfalls den Erwartungen, auch sieht der Unterschied der beiden Mittelwerte auf den ersten Blick nicht so dramatisch aus:

D1	D2
(OLS)	(LINMAP)
1.8380	1.7884

Man muß bedenken, daß zur Bildung eines jeden dieser Mittelwerte jeweils 4500 Beobachtungen herangezogen wurden, bei dieser Größe der Stichprobe wird die bei oberflächlicher Betrachtung geringfügige Differenz der beiden Werte bedeutsam.

Eine hohe Signifikanz des Haupteffekts von D war, wie gesagt, nicht zu erwarten und zwar v.a. nicht aufgrund der vorne (S. 66ff) dargestellten Ergebnisse der Untersuchung von Wittink & Cattin (1982): Zwar ist dort insgesamt die Anzahl der Bedingungskombinationen, unter denen ANOVA besser abschneidet, als LINMAP, größer als die derjenigen, bei denen es sich umgekehrt verhält, so daß insgesamt ein moderater Vorteil der metrischen Methode resultieren müßte, jedoch lassen die dort wiedergegebenen Werte nicht vermuten, daß dieser Vorteil allzu deutlich ausfällt. Umso weniger war beim vorliegenden experimentellen Design solches zu erwarten, gibt es hier doch insgesamt eher mehr Faktorstufenkombinationen, für die ein Vorteil zugunsten von LINMAP vorhergesagt wurde. Denn im Grunde bestand für 4 von 5 Stufen des Faktors A, zumindest aber für 2 (A3, A5), die Erwartung eines generellen Vorteils der nichtmetrischen Methode und nur bei kompensatorischer Datengenerierung wurde erwartet, daß OLS besser abschneide.

Die Interaktion A×D: Hier wurde die vorgefundene Signifikanz ($p < 0.001$ in allen Tests) dringlichst erwartet, die Hypothese, daß die metrische Methode bei kompensatorischer Datengenerierung bessere Ergebnisse produziere, während bei nicht-kompensatorischer Generierung die nichtmetrische Vorteile besitze, spielte bei der Planung der Untersuchung eine gewichtige Rolle. Der Blick auf die Mittelwerte zeigt jedoch, daß durch die Ergebnisse die Hypothese ganz und gar nicht bestätigt wird:

	A1	A2	A3	A4	A5
D1	1.8489	1.9555	1.6961	1.8148	1.8748
D2	1.8257	1.8324	1.6160	1.8170	1.8510

Mit einer Ausnahme (bei schwacher Interaktion) liegen die OLS-Werte immer über den LINMAP-Werten, gerade bei den Bedingungen mit dominantem Attribut wird der Abstand besonders deutlich. Der Test von Mittelwertsdifferenzen nach Tukey's HSD unterstreicht dieses: Nur unter den beiden Bedingungen mit dominantem Attribut wird der Abstand zwischen durchschnittlichen OLS- und LINMAP-Resultaten signifikant, unter den anderen Bedingungen der Datengenerierung gibt es keine signifikanten Unterschiede zwischen den beiden Analysemethoden. Die folgende Tabelle gibt die Signifikanzen aller möglichen Mittelwertsvergleiche nach Tukey's HSD wieder, darin zeigt „S" Signifikanz auf dem 1%-Niveau, „s" Signifikanz auf dem 5%-Niveau und „–" keine Signifikanz an:

	A1 D1	A1 D2	A2 D1	A2 D2	A3 D1	A3 D2	A4 D1	A4 D2	A5 D1	A5 D2
A1,D1		–	S	–	S	S	s	–	–	–
A1,D2	–		S	–	S	S	–	–	S	–
A2,D1	S	S		S	S	S	S	S	S	S
A2,D2	–	–	S		S	S	–	–	S	–
A3,D1	S	S	S	S		S	S	S	S	S
A3,D2	S	S	S	S	S		S	S	S	S
A4,D1	s	–	S	–	S	S		–	S	s
A4,D2	–	–	S	–	S	S	–		S	–
A5,D1	–	S	S	S	S	S	S	S		–
A5,D2	–	–	S	–	S	S	s	–	–	

Wie zu sehen ist, bilden v.a. die Mittelwerte der Faktorstufenkombinationen A2,D1 , A3,D1 und A3,D2 Extremwerte in dem Sinne, daß alle Vergleiche mit ihnen hochsignifikant sind. Ebenfalls ist zu sehen, daß kaum signifikante Unterschiede zwischen den Ergebnissen bei kompensatorischer Datengenerierung und denen bei Vorliegen einer Interaktion aufgetreten sind: Es scheint sich zu zeigen, was schon durch die Ergebnisse zum Haupteffekt A angedeutet ist, daß nämlich die Verfahren gegenüber den getesteten Interaktionsbedingungen robust sind, daß die Güte der Ergebnisse unter diesen Bedingungen im Großen und Ganzen der unter der kompensatorischen Datengenerierung entspricht.

Hinsichtlich des Interaktionseffekts A×D ist also nicht nur nicht das eingetroffen, was vorhergesagt wurde, sondern das genaue Gegenteil davon: OLS produziert bei Datengenerierung mit dominantem Attribut signifikant bessere Ergebnisse als LINMAP. Damit stehen hier die Ergebnisse in frappierendem Widerspruch nicht nur zu denen von Wittink & Cattin (1981), sondern zu der allgemein in der einschlägigen Literatur immer wieder vertretenen Auffassung (z.B. auch bei Cattin & Bliemel 1978), wonach solche nicht-kompensatorischen Daten die eigentliche Domäne nichtmetrischer Verfahren gegenüber den me-

trischen darstellen. Die Vorhersage bezüglich des Effekts A×D jedenfalls muß als widerlegt gelten.

Die Interaktion B×D: Die vorhergesagte Signifikanz ist nicht eingetroffen, weder bei den uni- noch bei den multivariaten Tests. In den Mittelwerten der Faktorstufenkombination ist auch beim besten Willen kein irgendwie geartetes Interagieren erkennbar:

	B1	B2	B3
D1	2.1163862	1.8262875	1.5715631
D2	2.0525148	1.7874648	1.5254588

Die Interaktion A×B×D: Vorhergesagt wurde Signifikanz die darauf zurückzuführen ist, daß der erwartete Vorteil von LINMAP bei nicht-kompensatorischer Datengenerierung mit wachsendem Fehler wieder schwindet. Nun konnte aber gar kein genereller Vorteil für LINMAP bei nicht-kompensatorischer Datengenerierung beobachtet werden, so daß die Hypothese eigentlich schon deswegen nicht mehr haltbar ist. Folglich kann es kaum noch überraschen, daß der Blick auf die Zellenmittelwerte ein ganz anderes Bild offenbart:

	A1	A2	A3	A4	A5
B1,D1	2.1362	2.2475	1.8492	2.1672	2.1816
B1,D2	2.1172	2.1137	1.7635	2.1227	2.1451
B2,D1	1.8608	2.0217	1.7016	1.7250	1.8221
B2,D2	1.8284	1.8743	1.6384	1.8015	1.7945
B3,D1	1.5498	1.5972	1.5376	1.5523	1.6107
B3,D2	1.5315	1.5091	1.4361	1.5269	1.6134

Mit einer einzigen Ausnahme (A4,B2,D1 vs. A4,B2,D2) liegen die OLS-Werte immer über den entsprechenden LINMAP-Werten, die folgende Tabelle, die die Signifikanzen nach Tukey's HSD für diese Vergleiche enthält, zeigt, daß der Unterschied immer bei dominantem Attribut signifikant zugunsten von OLS ausfällt, ansonsten weist nur die besagte Ausnahme, also das bessere Abschneiden von LINMAP bei mittlerem Fehler und schwacher Interaktion, Signifikanz auf:

	A1	A2	A3	A4	A5
B1: OLS vs. LINMAP	–	S	S	–	–
B2: OLS vs. LINMAP	–	S	s	S	–
B3: OLS vs. LINMAP	–	S	S	–	–

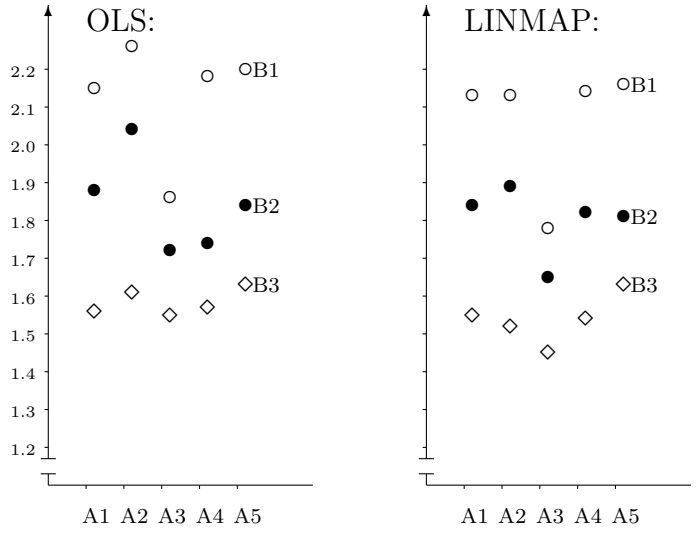

Abbildung 2.6: Mittelwerte der Faktorkombination A×B×D.

Die Frage ist, auf welche Weise die 3 Faktoren überhaupt interagieren. Aus den oben aufgelisteten Mittelwerten ist die Natur dieser Interaktion nicht ohne weiteres zu ersehen, weshalb zur Veranschaulichung der Effekt vermittels Abbildung 2.6 graphisch dargestllt wird. Dort allerdings ist auch keine deutliche Interaktion zu erkennen: Es zeigt sich ein typischer Verlauf der Werte unter den Bedingungen B1 und B2 über die 5 Stufen des Faktors A hinweg, der sowohl bei OLS, als auch bei LINMAP erkennbar ist, dieser Verlauf — das typische Absinken der Werte auf der Stufe A3 bei mittlerem und schwachem Fehler — ist in der Bedingung mit starkem Fehler (B3) deutlich abgschwächt. Diese Abschwächung tritt bei OLS etwas deutlicher auf (vgl. den Range der OLS- bzw. LINMAP-Mittelwerte bei starkem Fehler: 0.0731 vs. 0.1773). Man könnte also vielleicht sagen, daß bei starkem Fehler die Haupteffektwirkung von A — nämlich eben das Absinken der Werte bei A3 — aufgehoben wurde, diese Aufhebung scheint bei OLS stärker stattgefunden zu haben, als bei LINMAP.

Insgesamt aber könnte man mit Blick auf Abbildung 2.6 fragen, wie überhaupt die hohe Signifikanz des Interaktionseffektes zustandekommen konnte. Es sei hier wieder auf die hohe Anzahl von Fällen verwiesen, nämlich immerhin noch 300 pro Zelle dieser Faktorenstufenkombination (die Residualvarianz des Effekts hat 792 Freiheitsgrade) — es ist so möglich, daß auch kleine Unterschiede bedeutsam werden. Jedenfalls entspricht auch bei diesem Effekt das Ergebnis ganz und gar nicht den Erwartungen.

Die Interaktion A×C×D: Es wurde am Ende von Abschnitt 2.1.2 sozusagen nur eine sehr schwache Erwartung formuliert, wonach der erwartete Vorteil von LINMAP bei nicht-kompensatorischer Datengenerierung durch die Reduktion des faktoriellen Designs Verstärkung erfahren sollte. Die bisherigen Ergebnisse zeigen umgekehrt gerade bei dominantem Attribut Nachteile für LINMAP (vgl. A×D), so daß es nicht verwunderlich ist, wenn ähnliches sich auch hier zeigt. Tatsächlich ist diese Interaktion nur in den univariaten Tests mit $p < 0.001$ hochsignifikant, bei den multivariaten Tests, denen aufgrund des Ergebnisses des Sphärizitätstests mehr vertraut werden sollte, wird das 0.1%-Level knapp verfehlt ($p < 0.002$). Die Zellenmittelwerte zeigen in der Tat, daß auch hier die Voraussage nicht eingetroffen ist, auch hier schneidet OLS in nahezu allen Vergleichen der Faktorstufenkombinationen besser ab, als die entsprechende LINMAP-Zelle:

	A1	A2	A3	A4	A5
C1,D1	2.1071	2.2114	1.8133	2.1148	2.1392
C1,D2	2.1018	2.0738	1.7711	2.0740	2.1385
C2,D1	1.8895	1.9948	1.7294	1.7881	1.9421
C2,D2	1.8791	1.8998	1.6610	1.8725	1.9113
C3,D1	1.5503	1.6603	1.5456	1.5416	1.5431
C3,D2	1.4962	1.5235	1.4159	1.5046	1.5033

Merkwürdigerweise sitzt in dieser Tabelle die einzige Ausnahme an derselben Stelle, wie in der obigen Mittelwertstabelle für A×B×D: In der Bedingung mit schwacher Interaktion (A4) liegt bei auf 18 Stimuli reduziertem faktoriellem Design (C2) der LINMAP-Mittelwert über dem von OLS. Und genauso wie dort ist auch hier dieser Unterschied signifikant, wie die Tabelle der HSD-Signifikanzen der Vergleiche zwischen OLS- und LINMAP-Mittelwerten zeigt:

	A1	A2	A3	A4	A5
C1: OLS vs. LINMAP	–	S	–	–	–
C2: OLS vs. LINMAP	–	S	s	S	–
C3: OLS vs. LINMAP	–	S	S	–	–

Auch hier treten die signifikanten Vorteile von OLS wieder bei dominantem Attribut zutage, wobei diese Signifikanzen bei mäßig dominantem Attribut konstant bei allen Größen des faktoriellen Designs auftreten, bei stark dominantem Attribut dagegen nur mit sinkender Größe des Designs. Eine graphische Darstellung wie in Abbildung 2.6 macht hier nur wenig Sinn, da sie sich kaum von der eben jener Abbildung unterscheiden würde, wenn man nur die Stufen des Faktors B durch die des Faktors C ersetzt. Es fällt auf, daß hier die Vergrößerung der Fehlervarianz sich offenbar in genau derselben Art und

Weise auf die Ergebnisse ausgewirkt hat, wie die Reduktion des faktoriellen Designs.

Damit wurden alle diejenigen Effekte betrachtet, für die am Ende von Abschnitt 2.1.2 explizite Hypothesen formuliert wurden. Die Zwischenbilanz enttäuscht ein wenig: Die sozusagen zentrale Vorstellung, die der Untersuchung zugrundelag, war die, daß die große Robustheit metrischer Verfahren in der CA, wie sie z.B. von Carmone et al. (1978) und Cattin & Bliemel (1978) berichtet wird, dann ins Wanken gerät, wenn keine kompensatorische Datengenerierung mehr vorliegt, während umgekehrt die nichtmetrischen Verfahren gegenüber diesem Umstand robust sein müßten. Die Ergebnisse von Wittink & Cattin (1981) scheinen letzteres für LINMAP zu bestätigen. Die Hypothesen der vorliegenden Untersuchung wurden aufgrund der Vorstellungen aufgestellt, welche zur Erklärung der in Abschnitt 2.1.1 dargestellten Ergebnisse der früheren Simulationsstudien gebildet wurden. Das völlige Nichteintreten aller Prophezeiungen scheint zunächst diese Vorstellungen zu widerlegen und steht in besonders krassem Kontrast zu den Ergebnissen von Wittink & Cattin[54]. Wenn die Ergebnisse bis hierher auch in vielerlei Hinsicht ein diffuses Bild offenbaren, so scheint doch eines sehr deutlich zutage zu treten: OLS ist generell und ganz besonders ausgerechnet bei den Dominant-Attribut-Bedingungen gegenüber LINMAP überlegen!

Im folgenden werden diejenigen signifikanten Effekte betrachtet, für die keine Hypothesen formuliert wurden:

Die Interaktion A×B: Die spezifische Interaktionswirkung, obwohl in allen (uni- und multivariaten) Tests hochsignifikant, ist aus den Daten kaum zu ersehen:

	A1	A2	A3	A4	A5
B1	2.1267	2.1806	1.8064	2.1449	2.1634
B2	1.8446	1.9480	1.6700	1.7632	1.8083
B3	1.5406	1.5532	1.4918	1.5396	1.6170

Auf allen Stufen von A herrscht bei den Vergleichen der verschiedenen Fehlerstufen diesselbe Rangfolge B1≻B2≻B3 (die Rangfolge in den Spalten der obigen Tabelle), während die Rangfolge der verschiedenen Stufen von A unter den Stufen von B (die Rangfolge in den Zeilen der obigen Tabelle) variiert

[54]Das Mißtrauen, das dieser Kontrast gegenüber den eigenen Ergebnissen auslöste, führte dazu, daß die vielen Schritte der Datenverarbeitung, die zu ihnen führten, mehrmals überprüft wurden, um auszuschließen, daß durch irgendein Versehen bei einem dieser Schritte die Ergebnisse verfälscht wurden. Ich kann an dieser Stelle versichern, daß ein solches Versehen nicht stattgefunden hat und *diese* Ergebnisse tatsächlich so zustandegekommen sind, wie es dargestellt wurde.

— wobei allerdings immer A3 das Schlußlicht bildet. Auf letzterer Variation dürfte wohl die Signifikanz des Effektes beruhen. Allerdings ist aus diesen Veränderungen der Rangfolge der A-Stufen keine irgendwie geartete systematische Tendenz ersichtlich, durch welche die Interaktion interpretierbar würde. Eine solche systematische Tendenz scheint darin zu bestehen, daß auf der letzten Stufe von B, also bei der größten Fehlervarianz, die Unterschiede zwischen den Stufen von A gewissermaßen nivelliert sind. So nimmt der Range der Werte in den Zeilen der obigen Tabelle von oben nach unten ab: Die Rangewerte betragen 0.3742, 0.278 und 0.1252, die ersteren beiden sind nach Tukey's HSD signifikant mit $p < 0.001$, der letztere ist nicht signifikant. Dieser Umstand ist im Grunde schon in Abbildung 2.6 graphisch dargestellt, man muß dort nur die beiden Schaubilder für OLS und LINMAP zu einem zusammenfassen: Auf der unteren Stufe von B ist der typische Verlauf der Mittelwerte über die Stufen von A abgeschwächt und nähert sich einer Geraden.

Die Interaktion A×C: Hier gilt im Grunde dasselbe, wie für den soeben besprochenen Effekt, trotz hoher Signifikanz in allen Tests „zeigt" sich die Interaktionswirkung bei Betrachtung der Zellenmittelwerte kaum:

	A1	A2	A3	A4	A5
C1	2.1044	2.1426	1.7922	2.0944	2.1388
C2	1.8843	1.9473	1.6952	1.8303	1.9267
C3	1.5233	1.5919	1.4991	1.5231	1.5232

Hier ist sogar die Rangfolge der Zellen innerhalb einer Zeile (also die Rangfolge der Stufen von A) invariant über die Zeilen hinweg. Auch hier hilft der Blick auf die Mittelwertsvergleiche zum Verständnis der Wechselwirkung: Während in den ersten beiden Zeilen die Differenz zwischen dem höchsten und dem niedrigsten Zellenmittelwert — also zwischen A2,C1 und A3,C1 bzw. zwischen A2,C2 und A3,C2 — jeweils mit $p < 0.01$ signifikant nach Tukey's HSD ist, besteht in der letzten Zeile keine solche Signifikanz (ist die Differenz zwischen A2,C3 und A3,C3 nicht signifikant). Man erkennt hier also ein gewissermaßen zur Wechselwirkung A×B paralleles Bild, das in Abbildung 2.7 graphisch dargestellt ist: Es zeigt sich wieder der typische Verlauf über die Stufen von A für die Bedingungen C1 und C2, wie er ähnlich auch in Abbildung 2.6 zutage tritt (dort jeweils für die Bedingungen B1 und B2) und wie er auch bei allen anderen bisher besprochenen Interaktion bei graphischer Darstellung sich zeigen würde. Und so, wie in Abbildung 2.6 die „ungünstigste" Bedingung B3 diesen typischen Verlauf nivelliert, liegen auch hier unter C3 die Werte praktisch auf einer Linie.

Die Interaktion C×D: Einmal mehr ergaben sich hier hohe Signifikanzen für eine Wechselwirkung, welche als solche aus dem bloßen Anblick der entsprechenden Zellenmittelwerte nicht unbedingt offenbar wird:

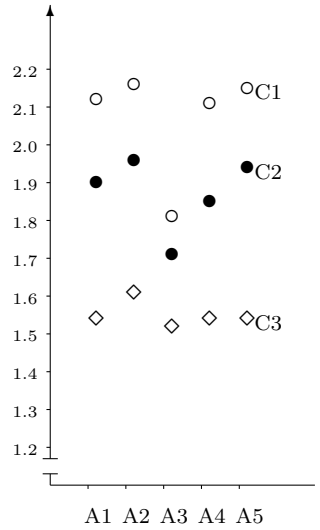

Abbildung 2.7: Mittelwerte der Faktorkombination A×C.

	C1	C2	C3
D1	2.0571966	1.8688228	1.5682174
D2	2.0358974	1.8447822	1.4887589
HSD:	–	s	S

Die Tabelle enthält in der letzten Zeile die Signifikanzen für die spaltenweisen Differenzen der Zellenmittelwerte, diese geben Aufschluß über die Natur der Interaktion: Zwar produziert LINMAP bei allen faktoriellen Designs schlechtere Ergebnisse als OLS, jedoch ist der Unterschied bei vollem faktoriellen Design nicht signifikant, beim Design mit 18 Stimuli nur signifikant auf dem 5%-Level und schließlich bei höchster Reduktion auf 9 Stimuli signifikant auf dem 1%-Niveau. Damit also tritt der Schwund in der Güte der Analyseergebnisse, der bei Reduktion des faktoriellen Designs bei beiden Methoden sehr drastisch ausfällt (die zeilenweisen Mittelwertsdifferenzen sind in obiger Tabelle natürlich hochsignifikant), bei LINMAP nocheinmal in verstärktem Maße auf.

Die Interaktion B×C×D: Auch diese Interaktion wies in allen (uni- und multivariaten) Tests sehr hohe Signifikanz ($p < 0.001$) auf und wie bei allen zuvor besprochenen Interaktionen ist auch hier die Wechselwirkung nicht ohne weiteres an den Mittelwerten abzulesen:

	B1	B2	B3
C1,D1	2.3730	2.0744	1.7841
C1,D2	2.3105	2.0260	1.7590
C2,D1	2.1709	1.8266	1.6088
C2,D2	2.1201	1.8532	1.5709
C3,D1	1.8052	1.5777	1.3216
C3,D2	1.7268	1.4831	1.2563

Betrachtet man jedoch die Paarvergleiche zwischen OLS- und LINMAP-Mittelwerten unter den verschiedenen Kombinationen B×C, so wird deutlich, warum dieser Effekt signifikant wurde:

	B1	B2	B3
C1: OLS vs. LINMAP	S	s	–
C2: OLS vs. LINMAP	s	–	–
C3: OLS vs. LINMAP	S	S	S

Nach dieser Tabelle sind die Differenzen zwischen OLS- und LINMAP-Resultaten, die aus dem auf 9 Stimuli reduzierten faktoriellen Design berechnet wurden, immer hochsigifikant, während beim weniger reduzierten bzw. vollen faktoriellen Design der Unterschied zwischen den Methoden nur bei kleinem und moderatem Fehler signifikante Ausmaße annimmt. Man könnte also — angesichts des Umstandes, daß weder die Interaktion B×C, noch B×D signifikant war — sagen, daß der Faktor B, nämlich das Ausmaß der Fehlervarianz mit der zuletzt besprochenen Wechselwirkung C×D interagiert.

Die Interaktion A×B×C: Bei diesem Effekt trat hohe Signifikanz ($p < 0.001$) nur bei den univariaten Tests auf, und zwar sowohl beim normalen F-Test, als auch beim konservativen nach Geisser & Greenhouse. Bei den multivariaten Tests war der Effekt nur mit $p < 0.01$ signifikant. Ich erspare an dieser Stelle mir und dem Leser die weitere Darstellung wie bei den obigen Effekten: Betrachtet man hier die Ergebnisse, dann ist kein spezifischer systematischer Effekt der Tripelinteraktion zu sehen, es zeigt sich lediglich, daß die schon besprochene Nivellierung der Unterschiede zwischen den verschiedenen Stufen von A, die beim auf 9 Stimuli reduzierten faktoriellen Design einerseits und bei großem Fehler andererseits auftritt und die bei graphischer Darstellung aus der Abschwächung des typischen Verlaufs über die Stufen von A hinweg sichtbar wird (vgl. Abbildungen 2.6 und 2.7), zwischen den verschiedenen Faktorstufenkombinationen, in denen eine der beiden Abschwächungsbedingungen B3 und C3 beteiligt ist, variiert, d.h. unter den 5 Kombinationen B1,C3; B2,C3; B3,C1; B3,C2; B3,C3 tritt der besagte Abschwächungseffekt in unterschiedlich starkem Ausmaß auf. Irgendeine systematische Tendenz, die interpretierbar

wäre, ist dabei aber beim besten Willen nicht zu erkennen, weswegen hier auf eine weitere Betrachtung dieses Effekts verzichtet wird.

Damit sind alle signifkanten Effekte besprochen. Es ist bei dieser Besprechung wohl deutlich geworden, das nicht unbedingt jede der vielen hochsignifikanten Wechselwirkungen per se einen besonders bedeutsamen inhaltlichen Zusammenhang wiedergibt: Besonders bei den Interaktionen 2. Ordnung springt oft kein spezifischer Wechselwirkungseffekt ins Auge, welcher auf einen qualitativ neuen inhaltlichen Zusammenhang, der erst durch das Zusammenwirken aller 3 Faktoren entsteht, verweist. Stattdessen ist über die bloße Potenzierung der beteiligten Effekte niederer Ordnung hinaus nur ein diffuses Interagieren derselben erkennbar, das zu scheinbar marginalen Verzerrungen der Ergebnisse, die sich bei dieser bloßen Potenzierung ergeben würden, führt, auf die jedoch die Signifikanztests äußerst sensibel reagieren. Überhaupt zeigen die vielen hohen Signifikanzen die Sensibilität dieser Tests, welche auch ein Charakteristikum der Meßwiederholungsanalyse darstellt: *„Die Verkleinerung der Fehlervarianz durch die Eliminierung der a priori Vpn-Unterschiede hat natürlich zur Konsequenz, daß Mittelwertsunterschiede von Varianzanalysen mit Meßwiederholung eher aufgedeckt werden als von Varianzanalysen ohne Meßwiederholung“* (Bortz 1979, S. 407). Es beruhigt, daß die hohen Signifikanzen meist auch in den multivariaten Tests auftreten, welche, da ihre F-Werte jeweils eine gegenüber den univariaten Tests deutlich reduzierte Anzahl von Nennerfreiheitsgraden besitzen, weniger sensibel gegenüber kleinen Mittelwertsunterschieden sind, wie die univariaten Tests. Der Unterschied wird an den beiden bisher noch nicht besprochenen Effekten deutlich: Sowohl $\mathsf{B}\times\mathsf{C}$, als auch die Interaktion 3. Ordnung $\mathsf{A}\times\mathsf{B}\times\mathsf{C}\times\mathsf{D}$ weisen bei den univariater Testung Signifikanz auf dem 5%-Niveau (konservativer F-Test oder ϵ-Korrektur) auf, sind jedoch bei den multivariaten Tests nicht signifikant. In beiden Fällen wird die univariate Signifikanz nicht interpretiert, entsprechend der schon im Abschnitt 2.2.1 besprochenen Empfehlung im SAS Manual, den univariaten Ergebnissen dann prinzipiell zu mißtrauen, wenn der Sphärizitätstest mit $p < 0.0001$ hochsignifikant wird, was bei beiden Wechselwirkungen der Fall war.

Es sei hier zum Abschluß der Darstellung der Ergebnisse noch darauf hingewiesen, daß bei aller Signifikanz in Effekten und Einzelvergleichen, die berichtet wurde, sich doch die Spearman-Korrelationen, die als Maß für die Güte der Analyseergebnisse verwendet wurden, insgesamt alle auf einem sehr hohen Level befinden: Nur 18 von 9000 Werten liegen unter 0.5 (der Minimalwert ist 0.274 unter der Bedingung $\mathsf{A5,B3,C3,D2}$), nur 48 Werte liegen unter 0.6 (vgl. auch Anhang B). Man könnte deshalb, differenziertere Ergebnisse außer Acht lassend, sagen, daß eigentlich beide Verfahren fast immer akzeptable Ergebnisse liefern und sich im Großen und Ganzen als sehr robust erwiesen haben. Bei der differenzierten Betrachtung geht es also weniger darum, ob ein Verfahren geeignet ist, als darum, welches optimale Ergebnisse liefert.

2.2.3 Diskussion und Ausblick

Die Ergebnisse der Varianzanalyse sind, vom Standpunkt der theoretischen Erwartungen aus betrachtet, sehr enttäuschend und sie bereiten mit Blick auf die in Abschnitt 2.1.1 dargestellten Ergebnisse v.a. der Untersuchung von Wittink & Cattin (1981) einiges Kopfzerbrechen: Wie ist es möglich, daß in der vorliegenden Untersuchung LINMAP durch die metrische Methode sozusagen um Längen geschlagen wird, während von Wittink & Cattin deutliche Belege dafür geliefert wurden, daß LINMAP bei Datengenerierung mit dominantem Attribut zumindest dann die klar besseren Ergebnisse liefert, wenn ein moderater Fehler vorliegt? Zunächst einmal aber sollen die systematischen Tendenzen, die an den oben berichteten Ergebnissen gewissermaßen über alle Effekte hinweg ins Auge stechen, benannt und diskutiert werden.

Zunächst einmal fällt der in allen Interaktionen, an denen A beteiligt ist, immer wiederkehrende, schon als typisch bezeichnete Verlauf der Zellenmittelwerte über die Stufen von A auf. Dieser ist in den Abbildungen 2.6 und 2.7 graphisch dargestellt und zeigt sich deutlich in den Zellenmittelwerten für den Haupteffekt A (vgl. S.100): Während die Mittelwerte bei kompensatorischem Modell und bei beiden Interaktionsbedingungen im allgemeinen auf demselben Niveau liegen, liegt der bei stark dominantem Attribut deutlich darunter, während der Spitzenwert bei schwach dominantem Attribut erzielt wird. Zwar unterscheidet sich letzterer beim Haupteffekt nur von einem der 3 „mittleren“ Mittelwerte signifikant, aber bei differenzierterer Betrachtung in Interaktionseffekten hebt sich dieser Mittelwert unter solchen Bedingungen, in denen der typische Verlauf deutlicher zutage tritt, oft signifikant von den Mittelwerten der kompensatorischen und Interaktionsbedingungen nach oben ab. Das aber bedeutet, daß die beiden Bedingungen mit dominantem Attribut sich völlig unterschiedlich, nämlich geradezu konträr auf die Güte der Analyseergebnisse auswirken! Offensichtlich fehlt den Vorstellungen über die Auswirkungen dominanter Attribute, wie sie z.B. in den im Abschnitt 2.1.2 angestellten Überlegungen zum Ausdruck kommen und auch der Untersuchung von Wittink & Cattin zugrundelagen, noch ein tieferes theoretisches Verständnis der Zusammenhänge. Die Auffassung, daß mit wachsender Dominanz eines Attributs die Rangwerte zusehends weniger „pseudometrisch“ werden und sich deswegen mit wachsender Dominanz ein kontinuierlich in eine Richtung wirkender Effekt auf die Ergebnisse einer CA zeigen müsse, greift entweder in der ersten, oder in der zweiten oder in beiden darin enthaltenen Hypothesen zu kurz.

Vielleicht bedürfte der behauptete einfache Zusammenhang zwischen Dominanz und metrischen Qualitäten von Rangwerten genauerer Untersuchung. Wie schon im Abschnitt 2.1.2 anhand des Zahlenbeispiels in Tabelle 2.1 erläutert wurde, verteilen sich die metrischen Gesamtnutzenwerte y_j bei Dominanz eines Attributs mit k Stufen auf k Gruppen um die Teilnutzenwerte der Stufen

des dominanten Attributs, so daß innerhalb einer jeden Gruppe nur geringe Abstände von im Rang aufeinanderfolgenden Werten, zwischen den Gruppen aber sehr große Abstände auftreten. Die Vorstellung, die der Hypothesenbildung zugrunde lag, war, daß dadurch bei der Umwandlung der y_j in Rangwerte p_j letztere die tatsächlichen Abstände in der subjektiven Präferenz für die einzelnen Stimuli verzerren, weil dabei sowohl alle Abstände innerhalb der Gruppen, als auch die zwischen ihnen gleichgesetzt werden. Nun könnte man diese allgemeine Vorstellung bezweifeln:

- Zum einen ist es denkbar, daß bei Dominanz eines Attributs die y_j eine Struktur erhalten, die die metrischen Qualitäten der Rangwerte gegenüber denen bei kompensatorischer Datengenerierung noch erhöht. Bei letzterer nämlich — das willkürlich gewählte Zahlenbeispiel in Tabelle 2.1 zeigt auch dieses —, resultiert mit zunehmender Zahl von Attributen und/oder Stufen eine der Normalverteilung ähnliche Struktur, d.h. die y_j kumulieren um den Mittelwert, die Abstände von im Rang aufeinanderfolgenden Werten werden zu den Enden hin größer. Es verhält sich also mit der kompensatorischen Datengenerierung nicht so, daß die Abstände aufeinanderfolgender Gesamtnutzenwerte immer annähernd gleich werden, wie es sein müßte, damit die metrischen Qualitäten der Rangwerte optimal würden — dieses war auch nicht die Vorstellung, die der Hypothesenbildung zugrundelag, sondern es wurde angenommen, daß die Struktur der y_j beim kompensatorischen Modell günstiger wäre, als beim Modell mit dominantem Attribut (oder mit nichtspezifizierter Interaktion). Nun könnte man dagegen vermuten, daß bei der durch ein dominantes Attribut bedingten Gruppierung der y-Werte gewissermaßen die Struktur, die aus kompensatorischer Datengenerierung resultiert, entzerrt wird: Die Kumulierung der Werte um eine Mitte wird aufgehoben. Es könnte darum sein, daß bei moderater Dominanz die metrischen Qualitäten der Rangwerte noch verbessert werden — während dann, wenn die Dominanz ein gewisses Ausmaß überschreitet, wieder eine Verschlechterung resultiert, aus den in Abschnitt 2.1.2 erwogenen Gründen. Da diese Überlegung nur schwer in Worten darzulegen ist, soll ein einfaches Zahlenbeispiel verdeutlichen, was gemeint ist (von der Fehlerkomponente wird in diesem Beispiel einmal abgesehen): Gegeben seien 2 Attribute mit jeweils 3 Stufen, zunächst soll bei kompensatorischer Datengenerierung jedes Attribut die Teilnutzenwerte -1, 0, und 1 haben. Es resultieren die Gesamtnutzenwerte

  -2, -1, -1, 0, 0, 0, 1, 1, 2.

  Wird nun eines der Attribute dominant mit den Teilnutzenwerten -2, 0, 2, dann resultiert

  -3, -2, -1, -1, 0, 1, 1, 2, 3.

  Erhält das dominante Attribut die Teilnutzenwerte -3, 0, 3, dann erhält man y_j, für die bei Umwandlung in Rangwerte keinerlei „metrische Information“

verlorenginge:

-4, -3, -2, -1, 0, 1, 2, 3, 4.

Erhöht man die dominanten Teilnutzenwerte weiter z.B. auf -6, 0, 6, dann entspricht die resultierende Struktur der y_j den in Abschnitt 2.1.2 für die Dominant-Attribut-Bedingungen gebildeten Erwartungen:

-7, -6, -5, -1, 0, 1, 5, 6, 7.

Vielleicht ist damit der typische Mittelwertsverlauf über die Stufen A1,A2 und A3 erklärbar: Die Bedingung A2 könnte in der dargestellten Weise als moderate Dominanz eine Verbesserung der Ergebnisse bewirkt haben, während für die hohe Dominanz die erwartete Verschlechterung aus den erwarteten Gründen eingetreten ist. Selbstverständlich resultiert aus den Ergebnissen der Untersuchung keine Beweiskraft für eine solche ad hoc Erklärung, diese müßte in zukünftigen Arbeiten überprüft werden.

- Zum anderen könnte man argwöhnen, daß die Auswirkung der Dominanz eines Attributs auf die metrischen Qualitäten der Rangwerte von der Anzahl Stufen des dominanten Attributs abhängt. Bei der 3^3-Struktur der vorliegenden Untersuchung bewirkt die Dominanz die Verteilung der y_j-Werte auf 3 Gruppen, d.h. bei den Abständen aufeinanderfolgender Werte gibt es 2 große Abstände zwischen den Gruppen und innerhalb der Gruppen insgesamt 24 kleine. Die Vorhersage bezog sich gewissermaßen auf die großen Abstände, durch diese sollte die Struktur der y_j so beschaffen sein, daß bei Transformation in Rangwerte viel „metrische Information“ verlorengeht. Nun stellt sich die Frage, ob nur 2 solcher Ausreißer bei den Abständen überhaupt genügen, um den vorhergesagten Effekt zu erzielen. Hätte das dominante Attribut mehr Stufen, so wüchse damit der Anteil der Ausreißer an den Abständen und vielleicht würde damit jeglicher Effekt der Dominanz verstärkt. Vielleicht hat also in der vorliegenden Untersuchung die geringe Anzahl der Stufen nicht ausgereicht, um die erwarteten Wirkungen zu zeitigen, vielleicht wären bei einer stark erhöhten Stufenzahl die Erwartungen angemessener.

Möglicherweise greift aber auch die Vorstellung zu kurz, daß die Auswirkung der Dominanz auf die Güte von CA-Ergebnissen ausschließlich über die metrischen Qualitäten der Rangwerte vermittelt wird. Diese Vorstellung besitzt allerdings eine gewisse Plausibilität wenn es um den Einfluß auf OLS-Ergebnisse geht, denn sie greift hier auf denjenigen Umstand zurück, der bei dieser Anwendung der OLS-Regression die Besonderheit ausmacht: Würde die Regression mit einem metrischen Kriterium — also mit den y_j — gerechnet und wäre damit die Voraussetzung bezüglich der Skalendignität der abhängigen Variable erfüllt, dann wäre eigentlich nicht einzusehen, wieso der größere Einfluß eines der Attribute irgendeine Auswirkung auf die Qualität der Berechnungen haben sollte. Dieses läßt sich für die OLS-Regression, deren Eigenschaften und Qualitäten vielfach erforscht und dokumentiert sind, behaupten, auf unsicher-

erem Grund stünde eine solche Behauptung bei den metrischen Verfahren, über die allesamt weniger veröffentlicht ist und die in diesem Sinne noch weniger verstanden sind. Vielleicht also war die Hypothesenbildung bei der vorliegenden Untersuchung zu sehr an der OLS-Regression orientiert, ist sozusagen der Verlockung erlegen, die die größere Durchsichtigkeit des Algorithmus für theoretische Spekulationen bietet. Tatsächlich wurde bei der Argumentation im Zusammenhang mit den Gleichungen 2.1.1 und 2.1.2 vor allem auf die BLUE-Qualitäten der OLS-Schätzer Bezug genommen: Der Schluß, daß dann, wenn die best-Qualität aufgrund der Verletzungen von OLS-Voraussetzungen nicht mehr gegeben ist, die LINMAP-Resultate sich automatisch im Vergleich zu denen von OLS verbessern müßten, war natürlich von Anfang an nicht logisch zwingend. Die Ergebnisse der Untersuchung scheinen anzuzeigen, daß er falsch war. Möglicherweise also wirkt sich Dominanz eines Attributs auf LINMAP noch vermittels eines anderen Umstands, als dem der metrischen oder nichtmetrischen Qualitäten der Rangwerte, aus — um dieses zu untersuchen wären allerdings zunächst genauere und tiefere Einsichten in die theoretischen Qualitäten dieses Algorithmus vonnöten.

Zur vollständigen Diskussion des typischen Verlaufs der Mittelwerte über die Stufen von A (d.h eigentlich des Haupteffekts von A) gehört auch die Betrachtung des Umstands, daß in den Ergebnissen an keiner Stelle irgendwelche beachtenswerten Unterschiede zwischen der kompensatorischen und den Interaktionsbedingungen auftreten: Offenbar waren in der vorliegenden Untersuchung beide Verfahren gegenüber den generierten Interaktionen gleichermaßen robust. Dieses Ergebnis deckt sich nicht mit dem von Carmone et al. (1976), das eine deutliche Verschlechterung bei Vorliegen einer nichtspezifizierten Interaktionsbedingung zeigt (vgl. S. 55ff). Nun besitzen aber den Ergebnisse von Carmone et al. aufgrund des Umstands, daß dort nur ein Fall generiert wurde, nicht gerade große Beweiskraft und es ist auch fraglich, inwieweit die dortige Interaktion vierter Ordnung mit der hier generierten erster Ordnung vergleichbar ist. Für das Nichteintreten signifikanter Unterschiede zur kompensatorischen Bedingung ist in der vorliegenden Untersuchung eventuell der Grund verantwortlich zu machen, der schon bei der Darstellung der Ergebnisse zum Haupteffekt A besprochen wurde (S. 100): Das numerische Gewicht der Summanden, die bei der Generierung der $\tilde{y}$-Werte durch die Interaktion entstehen (vgl. Abbildung 2.5, S. 83), war wahrscheinlich nicht groß genug, um Wirkungen zu erzielen: Vielleicht wäre unter der Bedingung A5 eine spürbare Verschlechterung eingetreten, wenn die starke Interaktion mit einem deutlich höheren Multiplikanten als 2 generiert worden wäre. Man kann also hier, mit aller Vorsicht, die bei generalisierenden Schlußfolgerungen aus den Ergebnissen dieser Untersuchung angebracht sind (siehe S. 97f), sagen, daß die Ergebnisse zeigen, daß aus einer moderaten unberücksichtigten Interaktion nicht unbedingt bedeutsame Verschlechterungen der Schätzungen subjektiver Präferenz-

werte folgen müssen.

Eine weitere systematische Tendenz in den in Abschnitt 2.2.2 berichteten Ergebnissen besteht in der Abschwächung des systematischen Verlaufs der Mittelwertsunterschiede über die Stufen von A unter Bedingungen, die durchschnittlich sehr niedrige Z-Werte verursachen, nämlich bei großer Fehlervarianz und beim auf 9 Stimuli reduzierten faktoriellen Design (und natürlich erst recht in der Kombination dieser beiden Bedingungen). Nun wurde bei der Hypothesenbildung argumentiert, daß bei nicht-kompensatorischer Datengenerierung OLS den erwarteten Vorteil von LINMAP mit wachsendem Fehler ausgleichen müsse, weil in dem Maße, wie die eigentliche Fehlervarianz σ_ϵ^2 gegenüber der „uneigentlichen“ σ_δ^2 wüchse, auch hier wieder die OLS-Voraussetzungen nahezu erfüllt seien — in gewisser Hinsicht hätte man aus diesem Gedankengang heraus auch eine Vorhersage der eingetretenen Nivellierung aller Unterschiede zwischen den Modelltypen bei großem Fehler zumindest für die OLS-Resultate ableiten können. Es soll nun natürlich nicht die Behauptung aufgestellt werden, daß so die der Hypothesenbildung zugrunde gelegten theoretischen Vorstellungen wenigstens ein winziges Stück weit bestätigt worden wären, aber vielleicht kann der besagte Gedankengang auf die Spur einer Erklärung führen. Zwar ist nämlich in diesem das Größenverhältnis, nicht aber der Zusammenhang zwischen σ_ϵ^2 und σ_δ^2 berücksichtigt. Eine allgemeine Vorstellung von diesem Zusammenhang ist nicht leicht zu bilden, es sind vielfältige Möglichkeiten denkbar, wie die Struktur der $\tilde{y}$-Werte durch die Addition von Fehlerwerten verändert wird. Vielleicht aber, so könnte man allerdings nur sehr vage vermuten, bewirkt diese Veränderung eine ähnliche Entzerrung der Gesamtnutzenwerte, wie sie oben für den Fall leichter Dominanz eines Attributs erwogen wurde. Durch die Addition von Fehlerwerten ist die Varianz der y-Werte gegenüber den fehlerfreien $\tilde{y}_j$ vergrößert, d.h. die Verteilung der y_j ist gewissermaßen auseinandergezogen, wodurch, wenn die $\tilde{y}_j$ infolge eines dominanten Attributs gruppiert waren, die Abstände zwischen den Gruppen bei den y-Werten verringert, die innerhalb der Gruppen vergrößert werden. Auf diese Weise könnte mit wachsendem Fehler eine Angleichung der y-Strukturen bei kompensatorischer und Dominant-Attribut-Bedingungen bewirkt worden sein. Allerdings tritt die Nivellierung der Ergebnisse über die A-Stufen nur beim 30%-Fehler deutlich hervor und nicht auf der mittleren Fehlerstufe.

Und außerdem tritt die Nivellierung ja auch unter der höchsten Reduktion des faktoriellen Designs auf, was durch obige Überlegung natürlich nicht erklärt werden kann. Es ist also insgesamt nur zu vermerken, daß diejenige(n) Einwirkung(en), welche den Unterschied in der Güte der Analyseergebnisse unter den getesteten Möglichkeiten des datengenerierenden Modells verursacht hat (haben), offensichtlich von Bedingungen, die dieser Güte besonders abträglich sind, weitgehend überlagert und aufgehoben wird (werden). Salopp formuliert: Unterschiedliche Generierungsmodelle produzieren unterschiedlich gute oder

gleich schlechte Ergebnisse.

Die dritte systematische Tendenz in den Ergebnissen, die Beachtung verdient, äußert sich im signifikanten Haupteffekt von C: Durchgehend kehrt in allen Effekten, an denen C beteiligt ist, die deutliche Reduktion der Mittelwerte im Verein mit der Reduktion des faktoriellen Designs wieder. Dieses Ergebnis ist das einzige, das in dieser Form erwartet wurde, und zwar wegen der Verminderung der Freiheitsgrade bei der Parameterschätzung: Es stehen mit zunehmender Reduktion des Designs zusehends weniger Beobachtungen zur Schätzung zur Verfügung. Trotzdem ist das Ergebnis bemerkenswert, widerspricht es doch einem Grundtenor in der CA-Literatur, wonach orthogonale reduzierte Designs problemlos verwendbar seien. Dieser Grundtenor hat offenbar dazu geführt, daß in praktischen Anwendungen der CA fast immer reduzierte Designs verwendet werden — was natürlich auch deshalb unvermeidbar ist, weil bei jeder größeren Conjoint-Struktur das volle faktorielle Design so viele Stimuli enthält, daß eine Versuchsperson *zu* viele Präferenzurteile abgeben müßte, so daß sie wahrscheinlich überfordert und damit der Fehleranteil wachsen würde. Wie schon im Abschnitt 2.1.3 geschildert, ist z.B. die komerziell vertriebene Programmversion von LINMAP (wahrscheinlich) aufgrund eines Programmierfehlers gar nicht in der Lage, größere faktorielle Designs zu bewältigen, was bisher offenbar noch niemandem aufgefallen ist. Es wurde schon kurz erwähnt (Abschnitt 1.1.2), daß die Ausarbeitung orthogonaler reduzierter faktorieller Designs in den allgemeineren theoretischen Kontext der Entwicklung varianzanalytischer Versuchspläne fällt, innerhalb dieses Kontexts ist es sozusagen gesicherte Erkenntnis, daß orthogonale reduzierte Designs eine effektive und unverzerrte Schätzung derjenigen Effekte erlauben, zu deren Schätzung sie konstruiert wurden (d.h. normalerweise von Haupteffekten unter Verzicht auf Interaktionen). Vielleicht beruht die Leichtfertigkeit der Verwendung dieser reduzierten Designs in der CA auf letzterem Umstand, dabei wird aber nicht berücksichtigt, daß diese Anwendung genaugenommen außerhalb — oder zumindest nicht ganz und gar innerhalb — des besagten theoretischen Kontexts steht bzw. einen sehr speziellen Fall darstellt: Die Sicherheit der Schätzung von Effekten auf der Basis orthogonaler Designs durch ANOVA bzw. OLS muß hier, da ja genaugenommen die Voraussetzung bezüglich der abhängigen Variablen nicht erfüllt ist, nicht unbedingt gelten, und sie muß erst recht nicht bei den nichtmetrischen Verfahren gelten. Tatsächlich zeigt sich am signifikanten Interaktionseffekt C×D, daß der Schwund in der Güte der Analyseergebnisse bei wachsender Reduktion des faktoriellen Designs bei LINMAP bedeutend rapider stattfindet, als bei OLS.

Die letzte der systematischen Tendenzen, die hier diskutiert werden soll, ist vielleicht die praktisch bedeutsamste: LINMAP produziert insgesamt schlechtere Ergebnisse und erweist sich als weniger robust gegenüber ungünstigen Bedingungen als OLS. Der Ausgangspunkt der ganzen Untersuchung war der

Eindruck — aufgrund der Ergebnisse von Wittink & Cattin (1981) und einer gewissen theoretischen Aufmerksamkeit, die das Verfahren in der Literatur genießt —, daß LINMAP dasjenige nichtmetrische Verfahren sei, daß eine echte Alternative zu den metrischen Verfahren darstellt, insofern es bestimmte „Domänen“ besitzt, in denen es gegenüber jenen überlegen ist. Unter den hier untersuchten Bedingungen befand sich jedenfalls keine solche Domäne! Die OLS-Schätzung erwies sich unter allen getesteten Bedingungen als das weitaus robustere Verfahren, dessen Ergebnisse sich zwar unter widrigen Bedingungen (Fehler, Reduktion des faktoriellen Designs, starke Dominanz eines Attributs) auch verschlechtern, aber meist nicht so dramatisch, wie die, welche LINMAP produziert. Tatsächlich scheint der Unterschied zwischen LINMAP- und OLS-Resultaten immer dann signifikant zu werden, wenn Bedingungen gegeben sind, bei denen auch insgesamt eine Veränderung der Ergebnisse gegenüber der „Referenzkombination“ kompensatorisches Modell mit kleinem Fehler und vollem faktoriellen Design eintritt. So ist der Unterschied z.B. in den Interaktionen A×D, A×B×D und A×C×D immer nur bei dominantem Attribut signifikant (von dem „Ausreißer“ in beiden Tripelinteraktionen unter A4, bei dem LINMAP signifikant besser abschneidet, abgesehen). Und schließlich zeigt sich unter den verschiedenen faktoriellen Designs diese Signifikanz erst mit wachsender Reduktion derselben (vgl. C×D).

Lediglich über die Stufen von B hinweg verhält es sich mit der Robustheit von LINMAP nicht wie in der zuletzt gezeigten Weise: Zwar erwies sich die Interaktion B×D als nichtsignifikant, doch zeigt die Wechselwirkung B×C×D Interessantes: Dort schneidet LINMAP unter allen Fehlerstufen dann signifikant schlechter ab, wenn das auf 9 Stimuli reduzierte faktorielle Design vorliegt, ansonsten jedoch (fast) nur noch unter der Bedingung mit schwachem Fehler. Tendenziell zeigt sich auch bei den Werten der nichtberücksichtigten Wechselwirkung B×D, daß der Unterschied eher bei kleinem Fehler größer ausfällt. Mit Blick auf die in diesem Abschnitt bereits unternommenen Erklärungsansätze erscheint dieses interessant: Wenn man nämlich die Hypothese aufstellt, daß nicht OLS, sondern im Gegenteil LINMAP besonders anfällig auf Rangwerte reagiert, die nur geringe metrische Qualität haben und zweitens die oben kurz erwogene hypothetische Vorstellung hinzunimmt, wonach mit wachsendem Fehler die Verzerrung der metrischen Datenstruktur durch die Transformation in Rangwerte teilweise kompensiert werden kann, dann könnte man damit ein solches Ergebnis erklären. Dann wäre nämlich mit wachsendem Fehler — ungeachtet der Verschlechterung der Analyseergebnisse, die damit zwangsläufig einhergeht — sozusagen der Startvorteil von OLS verspielt, weil die Daten größere metrische Qualität hätten. Selbstverständlich gilt auch gerade für diese Überlegung, daß sie nur hypothetischen Charakter besitzt und aus den Ergebnissen der vorliegenden Untersuchung keinerlei Gültigkeitsanspruch für sie abgeleitet werden kann!

Keiner weiteren Diskussion bedarf wohl die Signifikanz des Haupteffekts von B: Daß sich mit steigender Fehlervarianz die Ergebnisse insgesamt verschlechtern, versteht sich im Grunde von selbst und interessant wäre es nur geworden, wenn diese Verschlechterung — auch in den Einzelvergleichen — nicht signifikant ausgefallen wäre.

Bleibt die Frage, wie es möglich ist, daß die Ergebnisse der vorliegenden Untersuchung so große Divergenzen mit denen, die Wittink & Cattin (1978) berichten, aufweisen. Es sei in diesem Zusammenhang auch nocheinmal darauf hingewiesen, daß Wittink & Cattin alle Ergebnisse aus einem extrem reduzierten faktoriellen Design — mit 27 Stimuli bei 729 Stimuli des vollen 3^6-Designs — gewonnen haben: In der vorliegenden Untersuchung waren, wie dargestellt, gerade unter der Bedingung größter Reduktion des faktoriellen Designs die Unterschiede zwischen OLS und LINMAP — und bei diferenzierterer Betrachtung zwischen OLS und LINMAP bei dominantem Attribut (vgl. die Tabelle der Signifikanzen in den entsprechenden Mittelwertsvergleichen auf S. 106) — signifikant! Der eigentlich einzige möglicherweise bedeutsame Unterschied zwischen beiden Untersuchungen, der ins Auge fällt, ist die Größe der untersuchten Conjoint-Struktur: Bei Wittink & Cattin bestand dieselbe aus mehr, nämlich 6 Attributen. Es stellt sich so die Frage, inwieweit überhaupt die Effekte irgendwelcher unabhängiger Variablen auf die Güte von CA-Ergebnissen unter verschiedenen Conjoint-Strukturen miteinander vergleichbar sind. Oder, anders ausgedrückt: Welchen Einfluß hat die Anzahl der Attribute und wie interagiert diese unabhängige Variable mit anderen? Cattin & Bliemel (1978) fanden in ihrer Untersuchung, daß bei 9 statt 4 Attributen die beobachteten Effekte verstärkt auftraten, wobei allerdings mit der wachsenden Zahl von Attributen eine Verringerung der Fehlervarianz und die Reduktion des faktorielen Designs konfundiert war. Ansonsten scheint die Auswirkung der Anzahl der Attribute noch wenig erforscht, die unter Dominant-Attribut-Bedingungen stark gegenläufigen Tendenzen in der vergleichsweisen Güte von LINMAP-Resultaten und solchen aus metrischer Analyse bei Wittink & Cattin und in der vorliegenden Arbeit aber rücken diese Auswirkung stärker ins Blickfeld: Sie erscheint so als ein wichtiger Gegenstand zukünftiger Untersuchungen. Es ist jedenfalls kaum sinnvoll, Erklärungsversuche über die Diskrepanzen der beiden Untersuchungen anzustellen, solange keine genaueren Erkenntnisse über die Auswirkung dieses möglicherweise entscheidenden Unterschieds in der Größe der Conjoint-Struktur vorliegen.

Abschließend kann man also sagen, daß die Ergebnisse der vorgelegten Unteruchung v.a. zeigen, wie sehr es an grundlegendem Verständnis der bei einer typischen CA mit ordinalskalierter abhängiger Variable auftretenden Zusammenhänge zwischen den Qualitäten der Schätzverfahren und variablen Bedingungen, unter denen solche Analysen stattfinden, mangelt. Damit scheint es, als ob man am Ende der Untersuchung genau dort stünde, wo man auch am

Anfang schon gestanden hat, jedoch wird man zumindest sagen können, daß durch sie immerhin ein vermeintliches Verständnis, das sich in zu simplen theoretischen Vorstellungen äußert, erschüttert wird: Der Zusammenhang zwischen kompensatorischer versus nicht-kompensatorischer Datengenerierung und metrischer versus nichtmetrischer Analyse ist nicht immer so einfach vorauszusagen, wie es vorher den Anschein hatte und es ist auch nicht in allen Fällen unbedenklich, reduzierte Designs zu verwenden. Der Wert der Ergebnisse könnte also darin gesehen werden, daß sie gewissermaßen Etappe auf dem Weg zu differenzierterer Theoriebildung sein könnten, insofern aus dem Scheitern der weniger differenzierten Hypothesen der Blick auf mögliche neue, speziellere gelenkt wird. Nach meiner Einschätzung resultieren aus der vorgelegten Untersuchung v.a. die folgenden spezielleren Fragestellungen:

- Wie wirken sich unterschiedliche Stärken von Dominanz eines (oder mehrerer) Attribute auf die Struktur der metrischen Gesamtnutzenwerte y_j und damit auf das aus, was hier als metrische Qualität von Rangwerten bezeichnet wurde? Taugt der oben angedeutete, aus den vorgelegten Ergebnissen gefolgerte Zusammenhang, wonach bei mäßiger Dominanz eher eine Erhöhung dieser metrischen Rangwertqualitäten bewirkt wird, als Ansatz zur Beantwortung dieser Frage?

- Wie wirken sich unterschiedliche Fehlervarianzen auf die Struktur der y_j aus? Taugt hier die oben vage umrissene, ebenfalls aus den Ergebnissen gefolgerte Überlegung, wonach durch den Fehler unter Umständen die metrische Qualität der Rangwerte verbessert wird, als Erklärungsansatz und, wenn ja, welchen Verlauf nimmt der Zusammenhang zwischen Ausmaß der Fehlervarianz und metrischen Rangwertqualitäten?

- Steht der Zusammenhang zwischen metrischen Rangwertqualitäten und Dominanz eines Attributes in Wechslewirkung mit der Anzahl der Stufen des Attributs und, wenn ja, in welcher?

- Sind die Methoden gegenüber nichtspezifizierten Interaktionen robust bzw. in welcher Anzahl, Stärke und Ornung müssen solche Interaktionen gegeben sein, damit die Robustheit der Verfahren ernsthaft leidet?

- Gibt es eine Wechselwirkung der vorne untersuchten Effekte mit Variationen in der Größe der Conjoint-Struktur?

Mit diesen Fragen beende ich die Diskussion der Ergebnisse und überhaupt die Darstellung der durchgeführten Untersuchung. Ich hoffe daß, wer immer bis hierher gelesen hat, nun nicht zu der Einsicht kommt, daß es ganz umsonst gewesen ist.

2.3 Zusammenfassung

In der vorgelegten Untersuchung sollte die relative Robustheit von LINMAP und der OLS-Regression bei Anwendung im typischen Design der Conjoint-Analyse untersucht werden. Dieses typische Design zeichnet sich dadurch aus, daß Präferenzurteile — sogenannte Gesamtnutzenurteile — über multiattributive Stimuli in Form von Rangwerten erhoben werden und daß anhand dieser Rangwerte der Beitrag einzelner Ausprägungen der relevanten Attribute — Teilnutzenwerte genannt —, durch welche die Stimuli definiert sind, zur subjektiven Präferenz ermittelt werden soll. Dabei wird ein Modell zugrundegelegt, das davon ausgeht, daß die empirisch ermittelten Rangwerte auf der Basis von subjektiven, metrisch strukturierten Gesamtnutzenwerten gebildet werden, die durch die (meist additive) Verbindung der Teilnutzenwerte zustande kommen.

Die Verfahren, die zur Durchführung einer Conjoint-Analyse in Frage kommen, können grundsätzlich in metrische und nichtmetrische unterteilt werden: Erstere machen von der abhängigen Variable metrischen Gebrauch, d.h. sie interpretieren die Rangwerte als intervallskaliert. Es könnte so eventuell in dem Maße, wie bei einer Versuchsperson die Abstände in der subjektiven Präferenz zwischen solchen Stimuli, die im Rang aufeinanderfolgen, schwanken, dieser Gebrauch der abhängigen Variable zu Fehlern bei der Parameterschätzung führen. Andererseits sind die metrischen Verahren — OLS und ANOVA — allgemein leichter verfügbar, bekannter und in der Durchführung weniger aufwendig, so daß sich schon aus praktischen Gründen die Frage nach ihrer Robustheit gegenüber der Verletzung der Voraussetzung bezüglich der abhängigen Variablen stellt.

Für die nichtmetrischen Verfahren ist keine Fehlertheorie formuliert, sie sind ihrer Natur nach deterministisch. In empirischen Conjoint-Analysen wird man jedoch davon ausgehen müssen, daß die Urteile der Versuchspersonen mit einem gewissen Fehler behaftet sind. Es stellt sich hier die Frage, wie robust die nichtmetrischen Verfahren gegenüber dem Einfluß eines Fehlers auf die abhängige Variable sind.

Eine allgemeine Frage gilt der Robustheit von Verfahren der Conjoint-Analyse bei Verwendung reduzierter faktorieller Designs: Diese aus der Varianzanalyse bekannten Versuchspläne erlauben es hier, der Versuchsperson nicht Präferenzurteile über alle Kombinationen, die sich aus den verschiedenen Ausprägungen der Attribute bilden lassen, abzuverlangen, was von großer praktischer Relevanz ist, da ansonsten die Anzahl der präsentierten Stimuli meist in nicht mehr zu bewältigende Höhen wüchse.

Die vorgelegte Untersuchung knüpft an die Ergebnisse früherer Simulationsstudien an, bei denen Präferenzdaten unter Bedingungen, die möglicherweise für die Robustheit von Verfahren relevant sind, generiert und metrische und

nichtmetrische Methoden verglichen wurden. In diesen früheren Simulationsstudien wurde eine generelle Überlegenheit der metrischen Methode gegenüber den meisten nichtmetrischen Verfahren festgestellt. Wenig erforscht ist LINMAP, welches in der Untersuchung von Wittink & Cattin (1981) gegenüber ANOVA bessere Ergebnisse dann produzierte, wenn bei der Generierung der Präferenzurteile ein dominantes Attribut, d.h. ein solches, dessen Einfluß den der anderen Attribute deutlich übertrifft, erzeugt wurde.

In der vorgelegten Untersuchung wurde die Anzahl Rangwerte, die zur Schätzung zur Verfügung stehen — d.h. die Größe des faktoriellen Designs —, das Ausmaß der Fehlervarianz und der Typus des datengenerierenden Modells variiert. Unter der letzteren Variablen wurden verschiedene Möglichkeiten subjektiver Präferenzstrukturen zusammengefaßt, nämlich (1) die eines rein additiven Zusammenhangs von Attributen, die alle ungefähr die gleiche Wichtigkeit für die Bildung der Gesamtpräferenz besitzen (kompensatorische Datengenerierung), (2) die eines rein additiven Zusammenhangs, wobei jedoch ein Attribut besonders dominant ist und (3) die einer Interaktion zweier Attribute zusätzlich zum additiven Zusammenhang aller Attribute, wobei hier wieder von gleich wichtigen Attributen ausgegangen wurde. Sowohl die Stärke der Dominanz eines Attributs, als auch die der Interaktion wurden 2-fach abgestuft. Dabei war die grundlegende, hypothesenbildende Vorstellung die, daß unter der kompensatorischen Bedingung OLS, als metrisches Verfahren, gegenüber LINMAP, als nichtmetrisches, im Vorteil sein müßte, während umgekehrt LINMAP in den nichtkompensatorischen Bedingungen, v.a. bei Vorliegen eines dominanten Attributs, das metrische übertreffen müßte.

Diese Vorstellung konnte durch die Ergebnisse nicht bestätigt werden: OLS übertraf insgesamt LINMAP deutlich, und zwar besonders unter den Dominant-Attribut-Bedingungen! Sowohl das Anwachsen der Fehlervarianz, als auch die Reduktion des faktoriellen Designs führten erwartungsgemäß zu signifikanten Verschlechterungen der Analyseergebnisse. Die Ergebnisse scheinen v.a. den Mangel an Einsicht in die Zusammenhänge, die bei Durchführung einer Conjoint-Analyse wirksam werden, deutlich zu machen: Das Scheitern der wichtigsten Hypothesen richtet den Blick auf differenziertere theoretische Vorstellungen — und damit neue Hypothesen —, aus denen möglicherweise die gefundene Zusammenhänge erklärbar sind.

Literatur

Addelman, S. (1962a). Orthogonal main-effect plans for asymmetrical factorial experiments. *Technometrics*, 4, 21–46.

Anderson, N.H. & Shanteau, J.C. (1977). Weak inference with linear models. *Psychological Bulletin*, 84, 1155–1170.

Arbuckle, J. & Larimer, J (1976). The number of two-way tables satisfying certain additivity axioms. *Journal of Mathematical Psychology*, 13, 89–100.

Backhaus, K. (1990). *Multivariate Analysemethoden* (6. Aufl.). Berlin, Heidelberg: Springer.

Birnbaum, M.H. (1973). The devil rides again: Correlation as an index of fit. *Psychological Bulletin*, 79, 239–242.

Bortz, J. (1979). *Lehrbuch der Statistik*. Berlin, Heidelberg, New York: Springer.

Bose, R.C. & Bush, K.A. (1952). Orthogonal arrays of strength two and three. *Annals of Mathematical Statistics*, 23, 508–524.

Colberg, R.T. (1978). A Monte Carlo evaluation of metric recovery of conjoint measurement algorithms. Research Paper: College of Business Administration, University of Nevada-Reno.

Dixon, W.J. (Hrsg.) (1988). *BMDP statistical software maual*, Vol. 1 & 2. Berkeley, Los Angeles, London: University of California Press.

Eimer, E. (1978). *Varianzanalyse*. Stuttgart: Kohlhammer.

Emery, D.R. & Barron, F.H. (1979). Axiomatic and numerical conjoint measurement: An evaluation of diagnostic efficacy. *Psychometrica*, 44, 195–210.

Glaser, W. (1978). *Varianzanalyse*. Stuttgart: Gustav Fischer Verlag.

Green, P.E. (1974). On the design or choice experiments involving multifactor alternatives. *Journal of Consumer Research*, 1, 61–68.

Green, P.E. & Helsen, K. (1989). Cross-validation assessment of alternatives to individual-level conjoint analysis: A case study. *Journal of Marketing Research*, 26, 346–350.

Green, P.E. & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, 5, 103–123.

Green, P.E. & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, 54(4), 3–19.

Hagerty, M.R. (1985). Improving the predictive power of conjoint analysis: The use of factor analysis and cluster analysis. *Journal of Marketing Research*, 22, 168–184.

Hanushek, E.A. & Jackson, J.E. (1977). *Statistical methods for social scientists*. New York, San Francisco, London: Academic Press.

Hartung, J., Elpelt, B. & Klösener, K.-H. (1986). *Statistik: Lehr und Handbuch der angewandten Statistik*. München: Oldenbourg.

Hull, C.L. (1952). *A behavior system*. New Haven: Yale University Press.

Johnson, R.M. (1973). Varieties of conjoint measurement. Working Paper, Chicago: Market Facts, Inc.

Johnson, R.M. (1974). Tradeoff analysis of consumer values. *Journal of Marketing Research*, 11, 121–127.

Johnson, R.M. (1975). A simple method for pairwise monotone regression. *Psychometrika*, 40, 163–168.

Kamakura, W. (1988). A least squares procedure for benefit segmentation with conjoint experiments. *Journal of Marketing Research*, 25, 157–167.

Krantz, D.H. & Tversky, A. (1971). Conjoint measurement analysis of composition rules in psychology. *Psychological Review*, 78, 151–169.

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–27.

Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115–129.

Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society*, 27, 251–263.

McFadden, D. (1976). Quantal choice analysis: a survey. *Annals of Economic and Social Measurement*, 5, 363–390.

Mullet, G.M. & Karson, M.J. (1986). Percentiles of LINMAP conjoint indices of fit for various orthogonal arrays: A simulation study. *Journal of Marketing Research*, 23, 286–290.

Nickerson, C.A. & McClelland, G.B. (1984). Scaling distortion in numerical conjoint measurement. *Applied Psychological Measurement*, 8, 183–198.

Nygren, T.E. (1985a). Axiomatic and numeric conjoint measurement: A comparison of three methods for obtaining subjective workload (SWAT) rankings. *Proceedings of the IEEE 1985 NAECON conference*, 2, 878–883.

Nygren, T.E. (1985b). An examination of conditional violations of axioms for additive conjoint measurement. *Applied Psychological Measurement*, 9, 249–264.

Nygren, T.E. (1986). A two-stage algorithm for assessing violations of additivity via axiomatic and numerical conjoint analysis. *Psychometrika*, 51, 483–491.

Opp, K.-D. (1976). *Methodologie der Sozialwissenschaften.* Reinbek b. Hamburg: Rowohlt.

Orth, B. (1974). *Einführung in die Theorie des Messens.* Stuttgart: Kohlhammer.

Roberts, F.S. (1979). *Measurement theory with applications to decision making, utility, and the social sciences.* Reading, MA: Addison-Wesley.

Roskam, E.E. (1974). *Unidimensional conjoint measurement (UNICON) for multi-faceted designs.* Nijmwegen: Psychological Laboratory, University of Nijmwegen.

SAS Institute (Hrsg.) (1986). *SUGI supplemental library user's guide, vers. 5.* Cary, NC: SAS Institute Inc.

SAS Institute (Hrsg.) (1989). *SAS/STAT user's guide, vers. 6*, Vol. 2 (4. Aufl.). Cary, NC: SAS Institute Inc.

SAS Institute (Hrsg.) (1990). *SAS language: Reference, vers. 6* (1. Aufl.). Cary, NC: SAS Institute Inc.

Shanteau, J.C. (1977). Correlation as a deceiving measure of fit. *Bulletin of the Psychonomic Society*, 10, 134–136.

Spence, K.W. (1956). *Behavior theory and conditioning.* New Haven: Yale University Press.

Srinivasan, V. & Shocker, A.D. (1973a). Linear programming techniques for multidimensional analysis of preferences. *Psychometrika*, 38, 337–369.

Srinivasan, V. & Shocker, A.D. (1973b). Estimating the weights for multiple attributes in a composite criterion using pairwise judgements. *Psychometrika*, 38, 473–493.

Timmermans, Harry (1980). Unidimensional conjoint measurement models and consumer decision-making. *Area (Publication of the Institute of British Geographers)*, 12, 291–300.

Umesh, U.N. & Mishra, S. (1990). A Monte Carlo investigation of conjoint analysis index-of-fit : Goodness of fit, significance and power. *Psychometrika*, 55, 33–44.

Werner, J. (1993). Lineare Statistik. In Vorbereitung.

Winer, B.J. (1971). *Statistical principles in experimental design* (2. Aufl.). New York: McGraw-Hill.

Wittink, D.R. & Cattin, P. (1989). Commercial use of conjoint analysis: An update. *Journal of Marketing*, 53(3), 91–96.

Wittink, D.R. & Cattin, T. (1981). Alternative estimation methods for conjoint analysis: A Monte Carlo study. *Journal of Marketing Research*, 18, 101–106.

Zeleny, M. (1976). On the inadequacy of the regression paradigm used in the study of human judgement. *Theory and Decision*, 7, 57–65.

Anhang

A SAS-Jobs

Es werden hier natürlich nicht alle die vielen SAS-Jobs abgedruckt, die im Verlauf der Untersuchung gerechnet wurden, sondern nur einige ausgewählte, die transparent machen sollen, *wie* die wichtigsten der geschilderten Generierungs- und Auswertungsschritte durchgeführt wurden.

Zunächst der Job, mit dem der eine Satz von 100×9 Teilnutzenwerten, der der weiteren Datengenerierung zugrunde gelegt wurde, gezogen wurde. Über alle darin enthaltenen Statements informiert SAS Institute (1990). Das abschließende PUT-Statement hat den Sinn, eine leicht weiterzuverwendende Ausgabe der gezogenen Werte zu erzeugen:

```
data a;
retain x1 1 x2 2 x3 3 x4 4 x5 5 x6 6 x7 7 x8 8 x9 9;
do xx = 1 to 100;
call rannor(x1,pa1);
call rannor(x2,pa2);
call rannor(x3,pa3);
call rannor(x4,pb1);
call rannor(x5,pb2);
call rannor(x6,pb3);
call rannor(x7,pc1);
call rannor(x8,pc2);
call rannor(x9,pc3);
output; end; run;
data b; set a;
put (pa1-pa3)(7.5,+1) / (pb1-pb3)(7.5,+1) / (pc1-pc3)(7.5,+1); run;
proc means; var pa1 pa2 pa3 pb1 pb2 pb3 pc1 pc2 pc3; run;
```

Es folgt nun ein Beispiel für die Weiterverarbeitung der gezogenen Teilnutzenwerte. In diesem Falle werden die Rangwerte für die Bedingungskombination A3,B2,C1 erzeugt. Dazu werden zunächst aus den ursprünglichen Teilnutzenwerten die „wahren“ Teilnutzenwerte gebildet, zu diesen werden Fehlerwerte gezogen und addiert. Um die resultierenden Werte mittels der SAS-Prozedur RANK in Rangwerte transformieren zu können, muß zunächst die Datenmatrix mittels PROC-Transpose gedreht werden. Die Matrix der Rangwerte wird dann mit erneuter PROC TRANSPOSE wieder zurückgedreht, um die Rangwertedatei in einer Form ausgeben zu können (abschließendes PUT-Statement), wie

sie zum einlesen im LINMAP-Programm nötig ist (über die Prozeduren RANK und TRANSPOSE informiert SAS Institute 1989):

```
data old;
input x1 x2 x3 #2 pb1 pb2 pb3 #3 pc1 pc2 pc3;
if x1=max(of x1-x3) then do; pa1=x1+15; end; else do;
if x1=min(of x1-x3) then pa1=x1-15;
else pa1=x1; end;
if x2=max(of x1-x3) then do; pa2=x2+15; end; else do;
if x2=min(of x1-x3) then pa2=x2-15;
else pa2=x2; end;
if x3=max(of x1-x3) then do; pa3=x3+15; end; else do;
if x3=min(of x1-x3) then pa3=x3-15;
else pa3=x3; end;
s1=pa1+pb1+pc1;
s2=pa1+pb1+pc2;
s3=pa1+pb1+pc3;
s4=pa1+pb2+pc1;
s5=pa1+pb2+pc2;
s6=pa1+pb2+pc3;
s7=pa1+pb3+pc1;
s8=pa1+pb3+pc2;
s9=pa1+pb3+pc3;
s10=pa2+pb1+pc1;
s11=pa2+pb1+pc2;
s12=pa2+pb1+pc3;
s13=pa2+pb2+pc1;
s14=pa2+pb2+pc2;
s15=pa2+pb2+pc3;
s16=pa2+pb3+pc1;
s17=pa2+pb3+pc2;
s18=pa2+pb3+pc3;
s19=pa3+pb1+pc1;
s20=pa3+pb1+pc2;
s21=pa3+pb1+pc3;
s22=pa3+pb2+pc1;
s23=pa3+pb2+pc2;
s24=pa3+pb2+pc3;
s25=pa3+pb3+pc1;
s26=pa3+pb3+pc2;
s27=pa3+pb3+pc3;
n1=_n_;
cards;
```

Die im ersten Job gezogenen Werte werden eingelesen.

```
;
run;
data d2b; set old;
ev = 0.25*var(of s1-s27);
sd2b1 = s1+sqrt(ev)*rannor(801);
sd2b2 = s2+sqrt(ev)*rannor(802);
sd2b3 = s3+sqrt(ev)*rannor(803);
sd2b4 = s4+sqrt(ev)*rannor(804);
sd2b5 = s5+sqrt(ev)*rannor(805);
sd2b6 = s6+sqrt(ev)*rannor(806);
sd2b7 = s7+sqrt(ev)*rannor(807);
sd2b8 = s8+sqrt(ev)*rannor(808);
sd2b9 = s9+sqrt(ev)*rannor(809);
sd2b10 = s10+sqrt(ev)*rannor(810);
sd2b11 = s11+sqrt(ev)*rannor(811);
sd2b12 = s12+sqrt(ev)*rannor(812);
sd2b13 = s13+sqrt(ev)*rannor(813);
sd2b14 = s14+sqrt(ev)*rannor(814);
sd2b15 = s15+sqrt(ev)*rannor(815);
sd2b16 = s16+sqrt(ev)*rannor(816);
sd2b17 = s17+sqrt(ev)*rannor(817);
sd2b18 = s18+sqrt(ev)*rannor(818);
sd2b19 = s19+sqrt(ev)*rannor(819);
sd2b20 = s20+sqrt(ev)*rannor(820);
sd2b21 = s21+sqrt(ev)*rannor(821);
sd2b22 = s22+sqrt(ev)*rannor(822);
sd2b23 = s23+sqrt(ev)*rannor(823);
sd2b24 = s24+sqrt(ev)*rannor(824);
sd2b25 = s25+sqrt(ev)*rannor(825);
sd2b26 = s26+sqrt(ev)*rannor(826);
sd2b27 = s27+sqrt(ev)*rannor(827);
put (sd2b1-sd2b6)(8.5,+1) / (sd2b7-sd2b12)(8.5,+1) /
 (sd2b13-sd2b18)(8.5,+1) / (sd2b19-sd2b24)(8.5,+1) /
 (sd2b25-sd2b27)(8.5,+1) +2 n1 3.;
run;
proc means data=d2b; var sd2b1-sd2b27;
proc transpose data=d2b out=b prefix=obs;
var sd2b1-sd2b27;
proc rank data=b out=rd2b;
proc transpose data=rd2b out=r2d2b prefix=r;
data d2b2; set r2d2b; x=_n_;
```

```
put x 3. +1 (r1-r22)(2.,+1) / (r23-r27)(2.,+1);
run;
```

Der nächste Job führt die OLS-Analyse bei den oben gegenrierten Datensatz durch. Ein entsprechender LINMAP-Job kann hier nicht abgedruckt werden, da das Programm menügesteuert abläuft. Em Ende des Jobs werden die vorhergesagten Gesamtnutzenwerte in 2 eigens dafür eingerichteten Datensätzen gebildet und wieder mittels PUT in der erwünschten Form ausgegeben (weil dieser Job auf dem PC gerechnet wurde, mußte dieser Schritt aufgrund begrenzter Speicherkapazität zweigeteilt werden):

```
data a;
input nr r1-r22 #2 r23-r27;
cards;
```

Die im letzten Job generierten Rangdaten werden eingelesen.

```
;
proc transpose data=a out=b prefix=y;
var r1-r27;
run;
data c;
input pa1 pa2 pb1 pb2 pc1 pc2;
cards;
 1  0  1  0  1  0
 1  0  1  0  0  1
 1  0  1  0 -1 -1
 1  0  0  1  1  0
 1  0  0  1  0  1
 1  0  0  1 -1 -1
 1  0 -1 -1  1  0
 1  0 -1 -1  0  1
 1  0 -1 -1 -1 -1
 0  1  1  0  1  0
 0  1  1  0  0  1
 0  1  1  0 -1 -1
 0  1  0  1  1  0
 0  1  0  1  0  1
 0  1  0  1 -1 -1
 0  1 -1 -1  1  0
 0  1 -1 -1  0  1
 0  1 -1 -1 -1 -1
```

```
-1 -1  1  0  1  0
-1 -1  1  0  0  1
-1 -1  1  0 -1 -1
-1 -1  0  1  1  0
-1 -1  0  1  0  1
-1 -1  0  1 -1 -1
-1 -1 -1 -1  1  0
-1 -1 -1 -1  0  1
-1 -1 -1 -1 -1 -1
;
run;
data d;
merge b c;
run;
proc reg data=d outest=e;
model y1-y50 = pa1 pa2 pb1 pb2 pc1 pc2 / p;
run;
proc reg data=d outest=f;
model y51-y100 = pa1 pa2 pb1 pb2 pc1 pc2 / p;
run;
data g;
set e;
prs1=intercep+pa1+pb1+pc1;
prs2=intercep+pa1+pb1+pc2;
prs3=intercep+pa1+pb1-pc1-pc2;
prs4=intercep+pa1+pb2+pc1;
prs5=intercep+pa1+pb2+pc2;
prs6=intercep+pa1+pb2-pc1-pc2;
prs7=intercep+pa1-pb1-pb2+pc1;
prs8=intercep+pa1-pb1-pb2+pc2;
prs9=intercep+pa1-pb1-pb2-pc1-pc2;
prs10=intercep+pa2+pb1+pc1;
prs11=intercep+pa2+pb1+pc2;
prs12=intercep+pa2+pb1-pc1-pc2;
prs13=intercep+pa2+pb2+pc1;
prs14=intercep+pa2+pb2+pc2;
prs15=intercep+pa2+pb2-pc1-pc2;
prs16=intercep+pa2-pb1-pb2+pc1;
prs17=intercep+pa2-pb1-pb2+pc2;
prs18=intercep+pa2-pb1-pb2-pc1-pc2;
prs19=intercep-pa1-pa2+pb1+pc1;
prs20=intercep-pa1-pa2+pb1+pc2;
prs21=intercep-pa1-pa2+pb1-pc1-pc2;
```

```
prs22=intercep-pa1-pa2+pb2+pc1;
prs23=intercep-pa1-pa2+pb2+pc2;
prs24=intercep-pa1-pa2+pb2-pc1-pc2;
prs25=intercep-pa1-pa2-pb1-pb2+pc1;
prs26=intercep-pa1-pa2-pb1-pb2+pc2;
prs27=intercep-pa1-pa2-pb1-pb2-pc1-pc2;
put (intercep pa1 pa2 pb1 pb2 pc1 pc2)(8.5,+1) /
(prs1-prs9)(7.4,+1) / (prs10-prs18)(7.4,+1) / (prs19-prs27)(7.4,+1);
run;
data h;
set f;
prs1=intercep+pa1+pb1+pc1;
prs2=intercep+pa1+pb1+pc2;
prs3=intercep+pa1+pb1-pc1-pc2;
prs4=intercep+pa1+pb2+pc1;
prs5=intercep+pa1+pb2+pc2;
prs6=intercep+pa1+pb2-pc1-pc2;
prs7=intercep+pa1-pb1-pb2+pc1;
prs8=intercep+pa1-pb1-pb2+pc2;
prs9=intercep+pa1-pb1-pb2-pc1-pc2;
prs10=intercep+pa2+pb1+pc1;
prs11=intercep+pa2+pb1+pc2;
prs12=intercep+pa2+pb1-pc1-pc2;
prs13=intercep+pa2+pb2+pc1;
prs14=intercep+pa2+pb2+pc2;
prs15=intercep+pa2+pb2-pc1-pc2;
prs16=intercep+pa2-pb1-pb2+pc1;
prs17=intercep+pa2-pb1-pb2+pc2;
prs18=intercep+pa2-pb1-pb2-pc1-pc2;
prs19=intercep-pa1-pa2+pb1+pc1;
prs20=intercep-pa1-pa2+pb1+pc2;
prs21=intercep-pa1-pa2+pb1-pc1-pc2;
prs22=intercep-pa1-pa2+pb2+pc1;
prs23=intercep-pa1-pa2+pb2+pc2;
prs24=intercep-pa1-pa2+pb2-pc1-pc2;
prs25=intercep-pa1-pa2-pb1-pb2+pc1;
prs26=intercep-pa1-pa2-pb1-pb2+pc2;
prs27=intercep-pa1-pa2-pb1-pb2-pc1-pc2;
put (intercep pa1 pa2 pb1 pb2 pc1 pc2)(8.5,+1) /
(prs1-prs9)(7.4,+1) / (prs10-prs18)(7.4,+1) / (prs19-prs27)(7.4,+1);
run;
```

Schließlich noch der SAS-Job, mit dem die Meßwiederholungsanalyse gerechnet

wurde. Auch dieser ist einer von vielen, die probeweise und zum Vergleich der Ergebnisse unterschiedlicher Prozeduren und Optionen mit SAS und BMDP gerechnet wurden, jedoch derjenige, der die vorne berichteten Ergebnisse lieferte:

```
data a;
input c1-c5 / c6-c10 / c11-c15 / c16-c20 / c21-c25 / c26-c30 /
      c31-c35 / c36-c40 / c41-c45 / c46-c50 / c51-c55 / c56-c60 /
      c61-c65 / c66-c70 / c71-c75 / c76-c80 / c81-c85 / c86-c90 /
      c91-c95 / c96-c100;
array a c1-c100;
do over a;
a= 0.5*log((1+a)/(1-a));
end;
cards;
```

Die Spearman-Korrelationen werden eingelesen.

```
;
run;
proc transpose data=a out=b prefix=Z;
var c1-c100;
run;
*proc means data=b n mean std;
*run;
proc glm data=b outstat=gaga;
model Z1-Z90= / nouni e1 e3;
repeated D 2 contrast(1), A 5 contrast(1),
         B 3 profile, C 3 profile / printe printm summary;
run;
```

B Mittelwerte der Spearman-Korrelationen

			D1	D2
A1	B1	C1	0.9782067	0.9783840
		C2	0.9663282	0.9675619
		C3	0.9351247	0.9278543
	B2	C1	0.9642015	0.9633323
		C2	0.9440722	0.9418387
		C3	0.8921903	0.8853911
	B3	C1	0.9338457	0.9329857
		C2	0.9004630	0.8964469
		C3	0.8302670	0.8126480
A2	B1	C1	0.9781813	0.9768788
		C2	0.9698840	0.9660017
		C3	0.9434384	0.9239271
	B2	C1	0.9656892	0.9600543
		C2	0.9547113	0.9461929
		C3	0.9225252	0.8934284
	B3	C1	0.9296309	0.9170128
		C2	0.9027691	0.8891111
		C3	0.8424936	0.8070164
A3	B1	C1	0.9391369	0.9387588
		C2	0.9302481	0.9287343
		C3	0.9253841	0.9104274
	B2	C1	0.9349672	0.9314314
		C2	0.9220693	0.9194644
		C3	0.8987786	0.8747044
	B3	C1	0.9156888	0.9107185
		C2	0.8987384	0.8848922
		C3	0.8203187	0.7900996
sf A4	B1	C1	0.9796026	0.9784947
		C2	0.9695484	0.9683835
		C3	0.9365518	0.9305678
	B2	C1	0.9616554	0.9600797
		C2	0.8802072	0.9431599
		C3	0.8856714	0.8735267
	B3	C1	0.9325521	0.9296889
		C2	0.9023312	0.8969759
		C3	0.8327900	0.8290145
A5	B1	C1	0.9810800	0.9804692
		C2	0.9703819	0.9690089
		C3	0.9308918	0.9276359
	B2	C1	0.9580769	0.9578724
		C2	0.9405892	0.9397934
		C3	0.8900467	0.8757036
	B3	C1	0.9436245	0.9449735
		C2	0.9120582	0.9083714
		C3	0.8395025	0.8289056

C Quadratsummen der Meßwiederholungsanalyse

Die Bildung der Quadratsummen folgt den Regeln zur Konstruktion varianzanalytischer Versuchspläne, die Glaser (1978, S. 246ff) aufstellt. Für die Stufen der experimentellen Faktoren A, B, C, D werden dazu die Indizes a, b, c, d eingeführt ($a = 1, \ldots, 5; b = 1, 2, 3; c = 1, 2, 3; d = 1, 2$), für den Versuchspersonenfaktor V — d.h. hier für die generierten Fälle — bleibt der Laufindex i ($i = 1, 2, \ldots, 100$). Zunächst müssen aus den Z-Werten die folgenden Größen berechnet werden. Die Ausdrücke auf der linken Seite der Summenformeln folgen der von Glaser eingeführten Schreibweise (ebd., S. 69ff):

$$
\begin{aligned}
(VABCD) &= \sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{b=1}^{3}\sum_{a=1}^{5}\sum_{i=1}^{100} Z_{iabcd}^2 \\
(vABCD) &= \frac{1}{100}\sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{b=1}^{3}\sum_{a=1}^{5}\left(\sum_{i=1}^{100} Z_{iabcd}\right)^2 \\
(aVBCD) &= \frac{1}{5}\sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{b=1}^{3}\sum_{i=1}^{100}\left(\sum_{a=1}^{5} Z_{iabcd}\right)^2 \\
(bVACD) &= \frac{1}{3}\sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{a=1}^{5}\sum_{i=1}^{100}\left(\sum_{b=1}^{3} Z_{iabcd}\right)^2 \\
(cVABD) &= \frac{1}{3}\sum_{d=1}^{2}\sum_{b=1}^{3}\sum_{a=1}^{5}\sum_{i=1}^{100}\left(\sum_{c=1}^{3} Z_{iabcd}\right)^2 \\
(dVABC) &= \frac{1}{2}\sum_{c=1}^{3}\sum_{b=1}^{3}\sum_{a=1}^{5}\sum_{i=1}^{100}\left(\sum_{d=1}^{2} Z_{iabcd}\right)^2 \\
(vaBCD) &= \frac{1}{500}\sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{b=1}^{3}\left(\sum_{a=1}^{5}\sum_{i=1}^{100} Z_{iabcd}\right)^2 \\
(vbACD) &= \frac{1}{300}\sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{a=1}^{5}\left(\sum_{b=1}^{3}\sum_{i=1}^{100} Z_{iabcd}\right)^2 \\
(vcABD) &= \frac{1}{300}\sum_{d=1}^{2}\sum_{b=1}^{3}\sum_{a=1}^{5}\left(\sum_{c=1}^{3}\sum_{i=1}^{100} Z_{iabcd}\right)^2 \\
(vdABC) &= \frac{1}{200}\sum_{c=1}^{3}\sum_{b=1}^{3}\sum_{a=1}^{5}\left(\sum_{d=1}^{2}\sum_{i=1}^{100} Z_{iabcd}\right)^2 \\
(abVCD) &= \frac{1}{15}\sum_{i=1}^{100}\sum_{d=1}^{2}\sum_{c=1}^{3}\left(\sum_{a=1}^{5}\sum_{b=1}^{3} Z_{iabcd}\right)^2 \\
(acVBD) &= \frac{1}{15}\sum_{i=1}^{100}\sum_{d=1}^{2}\sum_{b=1}^{3}\left(\sum_{a=1}^{5}\sum_{c=1}^{3} Z_{iabcd}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
(adVBC) &= \frac{1}{10}\sum_{i=1}^{100}\sum_{c=1}^{3}\sum_{b=1}^{3}\left(\sum_{a=1}^{5}\sum_{d=1}^{2}Z_{iabcd}\right)^2 \\
(bcVAD) &= \frac{1}{9}\sum_{i=1}^{100}\sum_{a=1}^{5}\sum_{d=1}^{2}\left(\sum_{b=1}^{3}\sum_{c=1}^{3}Z_{iabcd}\right)^2 \\
(bdVAC) &= \frac{1}{6}\sum_{i=1}^{100}\sum_{a=1}^{5}\sum_{c=1}^{3}\left(\sum_{b=1}^{3}\sum_{d=1}^{2}Z_{iabcd}\right)^2 \\
(cdVAB) &= \frac{1}{6}\sum_{i=1}^{100}\sum_{a=1}^{5}\sum_{b=1}^{3}\left(\sum_{c=1}^{3}\sum_{d=1}^{2}Z_{iabcd}\right)^2 \\
(vabCD) &= \frac{1}{1500}\sum_{d=1}^{2}\sum_{c=1}^{3}\left(\sum_{b=1}^{3}\sum_{a=1}^{5}\sum_{i=1}^{100}Z_{iabcd}\right)^2 \\
(vacBD) &= \frac{1}{1500}\sum_{d=1}^{2}\sum_{b=1}^{3}\left(\sum_{c=1}^{3}\sum_{a=1}^{5}\sum_{i=1}^{100}Z_{iabcd}\right)^2 \\
(vadBC) &= \frac{1}{1000}\sum_{c=1}^{2}\sum_{b=1}^{3}\left(\sum_{d=1}^{2}\sum_{a=1}^{5}\sum_{i=1}^{100}Z_{iabcd}\right)^2 \\
(vbcAD) &= \frac{1}{900}\sum_{d=1}^{2}\sum_{a=1}^{5}\left(\sum_{c=1}^{3}\sum_{b=1}^{3}\sum_{i=1}^{100}Z_{iabcd}\right)^2 \\
(vbdAC) &= \frac{1}{600}\sum_{c=1}^{3}\sum_{a=1}^{5}\left(\sum_{d=1}^{2}\sum_{b=1}^{3}\sum_{i=1}^{100}Z_{iabcd}\right)^2 \\
(vcdAB) &= \frac{1}{600}\sum_{b=1}^{3}\sum_{a=1}^{5}\left(\sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{i=1}^{100}Z_{iabcd}\right)^2 \\
(abcVD) &= \frac{1}{45}\sum_{i=1}^{100}\sum_{d=1}^{2}\left(\sum_{c=1}^{3}\sum_{b=1}^{3}\sum_{a=1}^{5}Z_{iabcd}\right)^2 \\
(abdVC) &= \frac{1}{30}\sum_{i=1}^{100}\sum_{c=1}^{3}\left(\sum_{d=1}^{2}\sum_{b=1}^{3}\sum_{a=1}^{5}Z_{iabcd}\right)^2 \\
(acdVB) &= \frac{1}{30}\sum_{i=1}^{100}\sum_{b=1}^{3}\left(\sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{a=1}^{5}Z_{iabcd}\right)^2 \\
(acdVA) &= \frac{1}{18}\sum_{i=1}^{100}\sum_{a=1}^{5}\left(\sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{b=1}^{3}Z_{iabcd}\right)^2 \\
(vabcD) &= \frac{1}{4500}\sum_{d=1}^{2}\left(\sum_{c=1}^{3}\sum_{b=1}^{3}\sum_{a=1}^{5}\sum_{i=1}^{100}Z_{iabcd}\right)^2 \\
(vabdC) &= \frac{1}{3000}\sum_{c=1}^{3}\left(\sum_{d=1}^{2}\sum_{b=1}^{3}\sum_{a=1}^{5}\sum_{i=1}^{100}Z_{iabcd}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
(vacdB) &= \frac{1}{3000}\sum_{b=1}^{3}\left(\sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{a=1}^{5}\sum_{i=1}^{100}Z_{iabcd}\right)^2 \\
(vbcdA) &= \frac{1}{1800}\sum_{a=1}^{5}\left(\sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{b=1}^{3}\sum_{i=1}^{100}Z_{iabcd}\right)^2 \\
(abcdV) &= \frac{1}{90}\sum_{i=1}^{100}\left(\sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{b=1}^{3}\sum_{a=1}^{5}Z_{iabcd}\right)^2 \\
(vabcd) &= \frac{1}{9000}\left(\sum_{d=1}^{2}\sum_{c=1}^{3}\sum_{b=1}^{3}\sum_{a=1}^{5}\sum_{i=1}^{100}Z_{iabcd}\right)^2
\end{aligned}
$$

Damit können nun die auf die einzelnen Effekte des Meßwiederholungsplans mit den Faktoren A, B, C, D und V zurückführbaren Quadratsummen so gebildet werden, wie es in den nachfolgenden Tabellen dargestellt ist. Die Berechnung folgt den Regeln 2.1 bis 2.4 von Glaser (1978, S.248f), allerdings ist hier der Versuchspersonenfaktor nicht mit O bzw. o, sondern mit V bzw. v benannt.

Quadratsummen der Haupteffekte:		
SS_V	=	$(abcdV) - (vabcd)$
SS_A	=	$(vbcdA) - (vabcd)$
SS_B	=	$(vacdB) - (vabcd)$
SS_C	=	$(vabdC) - (vabcd)$
SS_D	=	$(vabcD) - (vabcd)$

Quadratsummen der Interaktionen 1. Ordnung:		
$SS_{V\times A}$	=	$(bcdVA) - (vbcdA) - (abcdV) + (vabcd)$
$SS_{V\times B}$	=	$(acdVB) - (vacdB) - (abcdV) + (vabcd)$
$SS_{V\times C}$	=	$(abdVC) - (vabdC) - (abcdV) + (vabcd)$
$SS_{V\times D}$	=	$(abcVD) - (vabcD) - (abcdV) + (vabcd)$
$SS_{A\times B}$	=	$(vcdAB) - (vacdB) - (vbcdA) + (vabcd)$
$SS_{A\times C}$	=	$(vbdAC) - (vabdC) - (vbcdA) + (vabcd)$
$SS_{A\times D}$	=	$(vbcAD) - (vabcD) - (vbcdA) + (vabcd)$
$SS_{B\times C}$	=	$(vadBC) - (vabdC) - (vacdB) + (vabcd)$
$SS_{B\times D}$	=	$(vacBD) - (vabcD) - (vacdB) + (vabcd)$
$SS_{C\times D}$	=	$(vabCD) - (vabcD) - (vabdC) + (vabcd)$

Quadratsummen der Interaktionen 2. Ordnung:		
$SS_{V\times A\times B}$	$=$	$(cdVAB) - (vcdAB) - (acdVB) - (bcdVA) + (vacdB) + (vbcdA) + (abcdV) - (vabcd)$
$SS_{V\times A\times C}$	$=$	$(bdVAB) - (vbdAC) - (abdVC) - (bcdVA) + (vabdC) + (vbcdA) + (abcdV) - (vabcd)$
$SS_{V\times A\times D}$	$=$	$(bcVAD) - (vbcAD) - (abcVD) - (bcdVA) + (vabcD) + (vbcdA) + (abcdV) - (vabcd)$
$SS_{V\times B\times C}$	$=$	$(adVBC) - (vadBC) - (abdVC) - (acdVB) + (vabdC) + (vacdB) + (abcdV) - (vabcd)$
$SS_{V\times B\times D}$	$=$	$(acVBD) - (vacBD) - (abcVD) - (acdVB) + (vabcD) + (vacdB) + (abcdV) - (vabcd)$
$SS_{V\times C\times D}$	$=$	$(abVCD) - (vabCD) - (abcVD) - (abdVC) + (vabcD) + (vabdC) + (abcdV) - (vabcd)$
$SS_{A\times B\times C}$	$=$	$(vdABC) - (vadBC) - (vbdAC) - (vcdAB) + (vabdC) + (vacdB) + (vbcdA) - (vabcd)$
$SS_{A\times B\times D}$	$=$	$(vcABD) - (vacBD) - (vbcAD) - (vcdAB) + (vabcD) + (vacdB) + (vbcdA) - (vabcd)$
$SS_{A\times C\times D}$	$=$	$(vbACD) - (vabCD) - (vbcAD) - (vbdAC) + (vabcD) + (vabdC) + (vbcdA) - (vabcd)$
$SS_{B\times C\times D}$	$=$	$(vaBCD) - (vabCD) - (vacBD) - (vadBC) + (vabcD) + (vabdC) + (vacdB) - (vabcd)$

Quadratsummen der Interaktionen 3. Ordnung:		
$SS_{V\times A\times B\times C}$	$=$	$(dVABC)-(vdABC)-(adVBC)-(bdVAC)-$ $(cdVAB)+(vadBC)+(vbdAC)+(vcdAB)+$ $(abdVC)+(acdVB)+(bcdVA)-(vabdC)-$ $(vacdB)-(vbcdA)-(abcdV)+(vabcd)$
$SS_{V\times A\times B\times D}$	$=$	$(cVABD)-(vcABD)-(acVBD)-(bcVAD)-$ $(cdVAB)+(vacBD)+(vbcAD)+(vcdAB)+$ $(abcVD)+(acdVB)+(bcdVA)-(vabcD)-$ $(vacdB)-(vbcdA)-(abcdV)+(vabcd)$
$SS_{V\times A\times C\times D}$	$=$	$(bVACD)-(vbACD)-(abVCD)-(bcVAD)-$ $(cdVAC)+(vabCD)+(vbcAD)+(vbdAC)+$ $(abcVD)+(abdVC)+(bcdVA)-(vabcD)-$ $(vabdC)-(vbcdA)-(abcdV)+(vabcd)$
$SS_{V\times B\times C\times D}$	$=$	$(aVBCD)-(vaBCD)-(abVCD)-(acVBD)-$ $(adVBC)+(vabCD)+(vacBD)+(vadBC)+$ $(abcVD)+(abdVC)+(acdVB)-(vabcD)-$ $(vabdC)-(vacdB)-(abcdV)+(vabcd)$
$SS_{A\times B\times C\times D}$	$=$	$(vABCD)-(vaBCD)-(vbACD)-(vcABD)-$ $(vdABC)+(vabCD)+(vacBD)+(vadBC)+$ $(vbcAD)+(vbdAC)+(vcdAB)-(vabcD)-$ $(vabdC)-(vacdB)-(vbcdA)+(vabcd)$

Quadratsumme der Interaktion 4. Ordnung:		
$SS_{V\times A\times B\times C\times D}$	$=$	$(VABCD)-(vABCD)-(aVBCD)-(bVACD)-$ $(cVABD)-(dVABC)+(vaBCD)+(vbACD)+$ $(vcABD)+(vdABC)+(abVCD)+(acVBD)+$ $(adVBC)+(bcVAD)+(bdVAC)+(cdVAB)-$ $(vabCD)-(vacBD)-(vadBC)-(vbcAD)-$ $(vbdAC)-(vcdAB)-(abcVD)-(abdVC)-$ $(acdVB)-(bcdVA)+(vabcD)+(vabdC)+$ $(vacdB)+(vbcdA)+(abcdV)-(vabcd)$