Erziehung und Didaktik

Gerhard Heilig

Schülereinstellungen zum Fach Erdkunde

Geographiedidaktische Forschungen Band 10

DIETRICH REIMER VERLAG BERLIN

Herausgeber:
Hochschulverband für Geographie
und ihre Didaktik
Schriftleitung:
Prof. Dr. Jürgen Nebel

Gerhard Heilig

Schülereinstellungen zum Fach Erdkunde

Methodische Verbesserungen bei der
Analyse geographie-didaktischer Erhebungen
durch multivariate Verfahren

Mit einem Vorwort von Helmut Schrettenbrunner

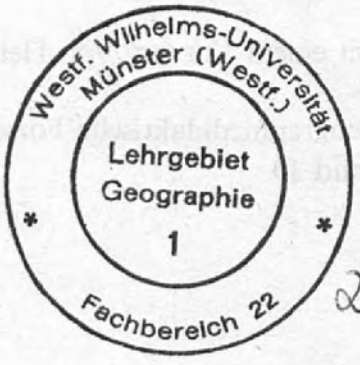
Geographiedidaktische Forschungen
Band 10

Dietrich Reimer Verlag

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Heilig, Gerhard:
Schülereinstellungen zum Fach Erdkunde: method. Verbesserungen bei d. Analyse geographie-didakt. Erhebungen durch multivariate Verfahren / Gerhard Heilig. Hrsg.: Hochsch.-Verb. für Geographie u. ihre Didaktik. Mit e. Vorw. von Helmut Schrettenbrunner. – Berlin: Reimer, 1984. –
(Geographiedidaktische Forschungen; Bd. 10)
ISBN 3–496–00728–1

NE: GT



Vorwort

Die vorliegende Arbeit basiert auf der umfangreichsten Serie von Schulversuchen im Fach Geographie, die je in Europa durchgeführt wurde und über 6000 Schülerfälle einschließt. Vergleichbar ist nur ein amerikanisches Projekt (High School Geography Project, HSGP), das in den 60er Jahren in den USA für das Fach im Rahmen der Social Studies erstellt wurde.

Das Deutsche Raumwissenschaftliche Curriculum-Forschungsprojekt (RCFP) wurde in den Jahren 1972–1978 von Bund und Ländern finanziert und war 1979–1981 Gegenstand eines DFG-Projektes, bei dem der Verfasser der vorliegenden Arbeit wissenschaftlicher Mitarbeiter war.

Sämtliche Originaldaten lagen also zwar schon vor, doch zerstückelt in Einzelprojekte und deshalb datentechnisch mit sehr vielen Problemen behaftet. Es ist dem Verfasser gelungen, die großen Probleme eines solchen Datensatzes zu lösen und letzten Endes dazu beizutragen, daß das Instrumentarium des RCFP ausreichend überprüft werden konnte.

Es kann gezeigt werden, daß einige gravierende Mängel zwar bestehen (z. B. fehlende Repräsentativität, fehlende Vor- und Nachtests, Nichtüberprüfung von theoretisch angenommen Dimensionen von Tests), gleichzeitig aber auch welche Überprüfung der Meßqualität vorgenommen werden müßte (Reliabilitäts- und Dimensionsanalysen).

Mit den erwähnten Einschränkungen werden die Einstellungsbatterien und Polaritätsprofile zum Fach oder zur RCFP-Einheit interpretiert. Dabei ergeben sich wichtige Aussagen für die Didaktik der Geographie hinsichtlich der Einstellungen von Jugendlichen zwischen 11 und 19 Jahren zum Fach Erdkunde, wobei allerdings trotz einer recht positiven Ausprägung offen bleiben muß, ob nicht andere Fächer noch positiver gesehen werden, da eine Vergleichbarkeit aus methodischen Mängeln nicht durchgefürht werden kann.

Anhand der multivariaten Analysen zwischen nicht-metrischen Daten gelingt es dem Verfasser auch graphisch sehr anschaulich zu belegen, welche Zusammenhänge bestehen, die gegenüber den bisherigen bivariaten Analysen neue Erkenntnisse bringen hinsichtlich der Einstellung der Schüler, der schulischen Erfahrung von Schülern und der Einstufung von RCFP-Einheiten.

Die vorliegende Arbeit ist ein Musterbeispiel für eine empirische Fachdidaktik der Geographie, die mit quantitativen Verfahren arbeitet. Sie enthält alle Schritte und viele Verfahren, die notwendig sind, wenn empirisch geforscht werden soll, und kann deswegen über den konkreten Fall der untersuchten RCFP-Einheiten hinaus als Leitfaden dienen, um Forschungsanlagen zu verbessern und Auswertungsmethoden adäquater als bisher einzusetzen.

Helmut Schrettenbrunner, im April 1983

1. Einführung in die Arbeit und Stand der Forschung

Die folgende Arbeit möchte zur methodischen Weiterentwicklung der empirischen Forschung beitragen, die im Rahmen der Geographiedidaktik einen Nachholbedarf gegenüber anderen sozialwissenschaftlichen Fachrichtungen aufweist.

Sie geht dabei von der Auffassung aus, daß sich methodische Probleme nur anhand konkreter *inhaltlicher Analysebeispiele* vortragen und klären lassen. Aus diesem Grund enthält die Arbeit eine Sekundäranalyse einer geographiedidaktischen Erhebung. Es handelt sich dabei um die bundesweite Evaluationsuntersuchung des Raumwissenschaftlichen Curriculum Forschungsprojektes (RCFP), bei der eine Reihe erdkundlicher Unterrichtseinheiten in der Unterrichtspraxis erprobt wurden. Vor und nach den Erprobungen wurden mittels Fragebogen u. a. verschiedene Schülereinstellungen zum Fach Erdkunde allgemein und zu den jeweiligen RCFP-Unterrichtseinheiten erhoben. Besonders für den Bereich dieser Schülereinstellungen sollen Sekundäranalysen zeigen, wie durch den Einsatz neuer Techniken und Auswertungsverfahren der bisherige Forschungsstand erweitert werden kann.

Die Perspektive der Arbeit ist primär methodisch und erst in zweiter Linie inhaltlich. Die Analysen haben den Charakter von Beispielen; der vorliegende Datensatz wird nicht systematisch durchanalysiert, sondern punktuell herangezogen, um bestimmte Techniken und Methoden am empirischen Material zu demostrieren.

Aus dieser methodischen Perspektive ergibt sich im einzelnen folgender Aufbau der Arbeit:

- *Kapitel 1* bringt einen Überblick über empirische Untersuchungen zur Geographiedidaktik in der Bundesrepublik Deutschland. Dabei wird deutlich, daß viele Arbeiten erhebliche methodische Unzulänglichkeiten aufweisen, oder einfach hinter dem Stand quantitativer Forschung in anderen Sozialwissenschaften hinterherhinken. Dies bestätigt die These vom „methodischen Nachholbedarf", die mehrere Geographiedidaktiker Ende der 70iger Jahre vertreten hatten, auch für die neueren Arbeiten.
- *Kapitel 2* präzisiert die Problemstellung der vorliegenden Arbeit. Dazu werden zunächst einige grundlegende Thesen zum empirischen Forschungsprozeß entwickelt, die als Maßstab der methodischen Kritik (in Kapitel 1) dienen, und die eine Richtschnur für die Auswahl methodischer Verbesserungsvorschläge in den folgenden Kapiteln darstellen. Aus diesen Thesen werden schließlich die Ziele der Arbeit abgeleitet.
- *Kapitel 3* befaßt sich mit der Datenbasis, auf die wir in der vorliegenden Arbeit zurückgreifen. Es wird dargestellt, wie aus den 10 ursprünglichen

Einzeluntersuchungen des RCFP ein Gesamtdatensatz aggregiert wurde. Dabei werden datentechnische und methodische Fragen besprochen, die bei der Zusammenführung einzelner Datensätze zu beachten sind. Es folgt ein Überblick über die Variablen, die für den Gesamtdatensatz ausgewählt wurden. Schließlich diskutieren wir noch einige Probleme der Repräsentativität und des Untersuchungsdesigns der RCFP-Erhebungen.

- Die *Kapitel 4, 5, und 6* widmen sich verschiedenen Aspekten der Meßproblematik. Am Beispiel einer Einstellungsbatterie und zweier Polaritätsprofile wird vorgeführt, daß die Meßqualität von Einstellungsvariablen mit relativ einfachen Verfahren überprüft und (auch nachträglich noch) verbessert werden kann. Bei den Verfahren handelt es sich um die „klassischen" Kennwertanalysen zur Überprüfung der Dimensionalität.
- *In Kapitel 7* werden einige Einstellungsvariablen einer bivariaten statistischen Analyse unterworfen. Dabei wird gezeigt, warum eine bivariate Analyse allenfalls für eine erste Inspektion des Datensatzes geeignet ist, nicht aber für die tiefergehende Analyse von (kausalen) Variablenzusammenhängen. Diese geringe analytische Kraft von bivariaten Verfahren bleibt auch dann bestehen, wenn sie durch inferrenzstatistische Tests abgesichert werden. Mit ihnen gelingt weder die Eliminierung von Scheinzusammenhängen, noch die Aufdeckung verdeckter Zusammenhänge. Es wird gezeigt, daß bivariate Analysen, bei den in der Geographiedidaktik üblichen Untersuchungsanlagen, im allgemeinen höchst zweifelhafte Ergebnisse erbringen.
- Die Problematik wird in *Kapitel 8* vertieft. Dabei konzentriren wir uns auf nicht-metrische Variablen. Auch bei ihnen ist eine Mehrvariablenanalyse möglich, und zwar durch die verschiedenen Techniken der Kontingenztabellenanalyse. Anhand der Analyse von Partial-Kontigenztabellen wird vorgeführt, wie bei nichtmetrischen Variablen multiple Zusammenhänge analysiert werden können. Nur so lassen sich durch Einführung zusätzlicher Kontrollvariablen Scheinkorrelationen aufdecken. Diese traditionelle Methode wird erweitert durch ein neues Verfahren zur schrittweisen Analyse multidimensionaler Kreuztabellen: die sog. „Clark-Higgins-Koch-Procedure" zur Variablenselektion. Das Verfahren stellt eine Analogie zur multiplen Regression bei metrischen Variablen dar.
- Eines der neuesten Modelle zur Analyse nicht-metrischer Variablensätze wird im *9. Kapitel* der Arbeit vorgeführt. Es handelt sich um einen speziellen Regressionsansatz, mit dem mehrdimensionale Kontigenztabellen analysiert werden. Dieser sog. „GSK-Ansatz (nach Grizzley, Starmer, Koch) stellt die grundlegenden Strukturen eines multiplen – oder auch multivariaten – Variablenzusammenhanges in der Form weniger „Parameter" dar. Am Beispiel von drei Einstellungsvariablen aus der RCFP-Erhebung wird demonstriert, wie sich damit der multiple Effekt mehrerer nichtmetrischer (Struktur)-Variablen auf eine nichtmetrische abhängige Einstellungsvariable analysieren läßt.

1.1 Zum Stand der quantitativen Didaktik der Geographie

Quantitative Forschung in der Didaktik der Geographie steht in Deutschland am Anfang. Der methodische Nachholbedarf gegenüber der empirischen Forschung in Psychologie und Soziologie ist beträchlich. Es werden z. T. statistische Verfahren benutzt, die schon vor Jahrzehnten überholt waren. So schrieb Schrettenbrunner im Jahre 1976: „Überprüft man die vorliegende deutsche Literatur zur Didaktik der Geographie nach dem Gesichtspunkt der Quantifizierung von Aussagen, dann ergibt sich, daß die Lehrbücher zur Didaktik (oder Methodik) ausnahmslos auf dem Stand vor der Quantifizierung stehen, und das schon deshalb, weil ihr Anliegen doch recht umfangreich das ganze Fach umspannen will. Da aber auch nur sehr wenige Artikel zur Didaktik auf dem einfachsten quantitativen Stand sind (Absolutwerte, Prozentwerte von Unterrichtsbeobachtungen), können die Lehrbücher also im wesentlichen nur auf der Basis der Erfahrung und Plausibilität beruhen." (Schrettenbrunner 1976, S. 5).

Im März des selben Jahres fand an der Pädagogischen Hochschule Freiburg ein Symposium über „Quantitative, empirische Methoden in der Didaktik der Geographie" statt. In seinem Einführungsreferat gab H. Haubrich eine Bestandsaufnahme der empirischen Geographiedidaktik in der Bundesrepublik Deutschland. Das Ergebnis war niederschmetternd. Mit leichter Untertreibung stellte Haubrich fest: „Ohne Zweifel liegt ein Defizit geographiedidaktischer und insbesonders empirischer Forschung vor" (Haubrich 1977a, S. 25). Haubrich hatte mehrere didaktische und fachdidaktische Zeitschriften der letzten 10 Jahre auf empirische Arbeiten hin durchgesehen. Außerdem war an allen 48 Standorten (in der Bundesrepublik) mit geographiedidaktischen Lehrstühlen eine Umfrage durchgeführt worden, wobei 37 Forschungsvorhaben erfaßt wurden. Die Analyse dieser beiden Erhebungen ergab u. a., daß „auf dem Gebiet der Forschungsmethoden . . . eine große Heterogenität und Unsicherheit vorzuliegen" scheint (Haubrich 1977a, S. 24). „Einmal werden Methoden genannt, die keine sind wie z. B. Lehrplanvergleich, Videoaufzeichnung und Diskrepanzmotivation, einmal sind die Angaben nur sehr pauschal wie deskriptive Statistik, EDV-Auswertung, Unterrichtsevaluierung, einmal erscheinen die Angaben sehr konservativ wie geisteswissenschaftliche, hermeneutische Verfahren, Leistungstests, Befragungstests; und einmal muten sie sehr progressiv an wie z. B. nichtmetrische Multidimensionale Skalierung und Interaktionsprozeßanalyse. Bei zahlreichen Angaben zu Forschungsmethoden wird deutlich, daß nicht unterschieden wird zwischen Methoden zur Datenerhebung und Methoden zur Datenauswertung" (Haubrich 1977a, S. 24f.).

Haubrich referierte diese Ergebnisse auch in einem Aufsatz in der Geographischen Rundschau (Haubrich 1977b). Hier zieht er eine deutliche Konsequenz aus dieser entmutigenden Situationsbeschreibung. „Die bisherige einseitige Theoriediskussion in der Didaktik der Geographie muß all-

mählich durch theorieorientierte empirische Forschung ergänzt und weiterentwickelt werden" (Haubrich 1977b, S. 26) – lautete seine These.

Allerdings sieht Haubrich auch die Gefahren einer forcierten empirischen Orientierung in der Didaktik der Geographie: „Wird die empirische Didaktik mit einer Sprache für Experten arbeiten, so daß man unter sich bleibt und nur füreinander redet und schreibt" (Haubrich 1977b, S. 27). Außerdem befürchtet Haubrich, daß neue quantitative Methoden mehr wissenschaftliche Exaktheit nur auf Kosten der inhaltlichen Reichweite erbringen.

Die Fachdidaktiker waren also Ende der 70iger Jahre (zumindest teilweise) zu der Überzeugung gekommen, daß die geisteswissenschaftlich orientierte Didaktik an einem toten Punkt angekommen war: So lautete das Motto des schon erwähnten Symposions: „Laßt es uns einmal im Bewußtsein ihrer Grenzen mit den Methoden der sog. exakten Wissenschaften versuchen!" (Haubrich 1977a, S. 31).

Genau das versucht diese Arbeit zu tun.

1.2 Vorliegende empirische Untersuchungen zum Fach Erdkunde

Es sind vor allem vier Themenbereiche, in denen in der Bundesrepublik empirische Forschungsarbeiten zum Fach Erdkunde unternommen wurden:

- Evaluationen von Unterrichtseinheiten,
- Lernweganalysen,
- Untersuchungen zu psycho-sozialen Rahmenbedingungen des Erdkundeunterrichts,
- Analysen von Unterrichtsformen und -medien,
- empirische Arbeiten über methodische Probleme.

Ich werde zunächst einen kurzen Überblick über diese Arbeiten geben. Im nächsten Abschnitt (1.3) werden dann jene Arbeiten, die für unser Thema interessant sind, nach methodischen Gesichtspunkten genauer analysiert.

(1) Unterrichtsevaluationen

Empirische Evaluationen von erdkundlichen Unterrichtseinheiten wurden in der Bundesrepublik Deutschland bisher im wesentlichen vom RCFP durchgeführt (vergl. *Fürstenberg/Jungfer* 1980). Das methodische Niveau dieser Arbeiten ist relativ niedrig: Über einfache Kreuztabellen und Häufigkeitsauszählungen wird kaum hinausgegangen. Vor allem aber entspricht das verwendete Forschungsdesign nicht einmal in Ansätzen einem streng wissenschaftlichen Vorgehen. Die Evaluationen des RCFP orientieren sich statt dessen mehr an pragmatischen Kriterien zur Revision und Implemen-

tation der betreffenden Unterrichtseinheiten. Dennoch kann man diese Evaluationen als empirische Untersuchungen betrachten, da zumindest eine systematische und kontrollierte (quantitative) Datenerhebung stattgefunden hatte.

Neben den eigentlichen RCFP-Untersuchungen gibt es einige Evaluationen mit Unterrichtseinheiten, die im Umfeld des RCFP entwickelt, aber dann nicht in die RCFP-Erprobung aufgenommen wurden (*Schrettenbrunner* 1981; *Heilig* 1981; *Schedl* 1981). Bei den Evaluationen dieser Einheiten wurde wesentlich mehr in das Erhebungsdesign investiert (z. B. immer *Vor-* und Nachtest). Außerdem versuchte man die Qualität der Meßinstrumente durch Reliabilitätsanalysen zu überprüfen und zu verbessern.

Im anglo-amerikanischen Sprachraum sind quantitative Evaluationen von erdkundlichen Unterrichtseinheiten nicht nur wesentlich stärker verbreitet, sondern auch auf methodisch höherem Niveau (vergl. *Irwin/Baumgart* 1978).

Natürlich gibt es außerdem eine große Menge von *Curriculum-Vorschlägen*, sowie etliche *Erfahrungsberichte* zu Unterrichtseinheiten. Schließlich finden sich in den einschlägigen Fachzeitschriften eine Vielzahl von Arbeiten zu verschiedenen *Unterrichtsmedien* (vergl. *Haubrich* 1977b, S. 27). Aber alle diese Arbeiten sind entweder theoretisch-konzeptioneller Art oder sie basieren auf dem vorwissenschaftlichen Niveau der Alltagserfahrung. In beiden Fällen fehlt eine kontrollierte, intersubjektiv nachvollziehbare Datenerhebung und damit die Grundbedingung jeder empirischen Forschungsarbeit.

(2) *Lernweganalysen*

Der zweite Bereich empirischer Untersuchungen zum Fach Erdkunde sind Analysen von Lernprozessen und Lernpfaden, wobei häufig computerunterstützte Unterrichtsprogramme als Modell der vielfach höchst kompliziert verzweigten Lernwege entwickelt wurden. (*Schrettenbrunner* 1976a, 1976b, 1977; *Nolzen* 1977; *Geiger* 1977).

Diese empirischen Arbeiten haben häufig einen wesentlich höheren methodischen Stand als die vorhin genannten Evaluationen: So werden generell nicht nur die Variablen aus dem Lernvorgang selbst (benutzte Lernschritte, Zeit pro Lernschritt, Fehlerzahl) quantitativ erfaßt, sondern auch bedeutsame Hintergrundvariablen (wie Geschlecht des Schülers, Beruf des Vaters, IQ, usw.). Vor allem aber werden diese Variablen mit multiplen statistischen Verfahren (wie Varianz- und Regressionsanalysen) auf Zusammenhänge hin überprüft. Und schließlich gibt es auch Ansätze zu einer Entwicklung „härterer" Theorien des (erdkundlichen) Lernens durch Mathematisierung des Lehralgorithmus mit Hilfe der Graphentheorie (*Schrettenbrunner* 1977). Dabei werden Detailstrukturen des Lernprozesses gleichsam in Form eines Mikro-Modells herausgearbeitet und in verschiedenen

Unterrichtssituationen analysiert (z. B. in verschiedenen Klassenstufen, bei Schülern mit unterschiedlichem IQ, usw.).

Diese Gruppe empirischer Arbeiten knüpft stark an experimentelle Traditionen an. Damit wird ein besonders strenges wissenschaftliches Untersuchungsdesign akzeptiert; allerdings auch, daß die Ergebnisse oft nur sehr geringe Allgemeingültigkeit besitzen (s. *Heipke* 1967).

(3) *Untersuchungen zu psycho-sozialen Rahmenbedingungen des Erdkundeunterrichts.*

Der dritte (etwas heterogene) Bereich empirischer Untersuchungen befaßt sich mit den psycho-sozialen Rahmenbedingungen des Erdkundeunterrichts.

Dazu gehören zunächst einmal eine Reihe auch methodisch relativ anspruchsvoller Untersuchungen zum *„räumlichen Denken"* der Schüler (*Engelhardt* 1973/*Fichtinger* 1974/*Dueck* 1978/*Schrettenbrunner* 1978). Zum Teil werden hier Fragestellungen aus der „Mental-Map-Forschung" der Fachgeographie (die dort eine große Bedeutung haben) übernommen. Die Perspektive dieser Untersuchungen ist stark psychologisch.

Daneben gibt es eine überraschend lange *Tradition empirischer Erhebungen von Schülereinstellungen*, über die wir hier etwas ausführlicher berichten wollen:

Bereits 1905 hatte *W. Stern* eine Erhebung über die Beliebtheit von Schulfächern durchgeführt, wobei u. a. auch das Fach Geographie von den Schülern (als relativ unbeliebt) eingestuft wurde (*Stern* 1905). *Stern* hatte 1461 Knaben und 1095 Mädchen (also 2556 Schüler) danach befragt, welches Fach sie am liebsten hätten. Es waren Schüler aus praktisch allen damals bestehenden Schulgattungen vertreten. Die statistische Auswertung der Erhebung bestand in den Häufigkeitsauszählungen für die Nennungen der verschiedenen Schulfächer sowie deren jeweilige Aufgliederung nach Altersstufen, „höheren" und „niederen" Schulen, Geschlechtern und „Stadt"-„Land"-Schulen. Bereits 2 Jahre später griff *G. Wiederkehr* dieses Thema mit einer Veröffentlichung erneut auf (*Wiederkehr* 1907/08). Er verbesserte die Erhebungsmethode von Stern durch Einführung eines „Versuchsleiters". Während bei Stern die Erhebung der Daten durch die jeweiligen Klassenlehrer erfolgte, führte Wiederkehr (mit einem Mitarbeiter) die ganze Erhebung selbst durch. Wieder zwei Jahre später erschien eine Veröffentlichung von *M. Lobsien*. Darin berichtete er über eine Befragung von nicht weniger als 6248 Schülern (3343 Knaben, 2905 Mädchen) der Kieler Volks- und Mittelschulen von Klasse 1 bis 6. Er benutzte zur Einstufung der Beliebtheit eines Schulfaches im Gegensatz zu seinen Vorgängern bereits eine Rang-Skala: Die Schüler vergaben für die verschiedenen Fächer Rangplätze; das beliebteste Fach erhielt eine 1, das zweitliebste Fach eine 2, usw. (*Lobsien* 1909).

Diese und eine Reihe kleinerer empirischer Untersuchungen wurden damals gründlich diskutiert und in zusammenfassenden Veröffentlichungen

miteinander verglichen, am ausführlichsten von *Paul Hoffmann* (*Hoffmann* 1911). Im Jahre 1924 konnte *G. Lunk* mehr als ein Dutzend empirische Arbeiten über die Beliebtheit verschiedener Schulfächer zitieren.

Zusammenfassend kann man festhalten, daß diese frühen Arbeiten ein überraschend hohes methodisches Niveau hatten. Ihre Datenbasis jedenfalls war sicher fundierter als die der meisten „modernen" Untersuchungen.

Natürlich gibt es auch etliche *neuere empirische Arbeiten* über Schülereinstellungen zu Schulfächern allgemein (vergl. z. B. *Seelig* 1968) und zum Fach Erdkunde im speziellen. Hier geht es nicht mehr nur um Beliebtheit oder Unbeliebtheit des jeweiligen Faches, sondern um eine Vielzahl verschiedener Einstellungsaspekte (vgl. *Gaensslen* 1976).

So befaßt sich *Leusmann* u. a. mit Polaritätsprofilen über schulgeographische Inhalte, Medieneinsatz und Einstellungen zu verschiedenen Unterrichtsformen (*Leusmann* 1976a, 1976b, 1977, 1978). *H. Schrettenbrunner* und *Bauer* befragten die Schüler in ihren Untersuchungen u. a über das Freizeitverhalten (Ausflüge, Fahrten), über die Meinung zu bestimmten Unterrichtsmedien (Schulatlas, Erdkundebuch) und über verschiedene Unterrichtsformen (Diavortrag, Referat, Lehrwanderung) (*Schrettenbrunner* 1969 / *Bauer* 1969). *Schumacher* untersuchte die – für das Fach Erdkunde sicher interessanten – Einstellungen der Schüler zur „Dritten Welt" (*Schumacher* 1974).

Im englischen Sprachraum liegen eine Vielzahl empirischer Untersuchungen über Schülereinstellungen vor.[1] Für uns interessant ist die Arbeit von *T. Haladyna* und *G. Thomas*: Sie unternahmen 1979 einen *nicht*-verbalen Einstellungstest mit 3000 Grundschülern (elementary school), bei dem die Einstellung der Kinder zu 7 Fächern und zur Schule als solcher gemessen wurden (*Haladyna* 1979). Ein für die vorliegende Arbeit besonders interessantes Ergebnis war, daß in höheren Klassen die Einstellungen generell negativer wurden. Die Autoren folgern daraus, daß es im Lauf der Schulkarriere zu einer bedenklichen gefühlsmäßigen Klimaverschlechterung kommt. Wir werden darauf zurückkommen.

Zum Schluß wollen wir noch auf eine Arbeit verweisen, die eine besonders große Zahl unterschiedlicher Schüler- und Lehrereinstellungen zum Unterrichtsfach untersucht hat: Es handelt sich um die Dissertationsarbeit von *G. Bachmaier* (*Bachmaier* 1969). Auch auf diese Arbeit werden wir später noch genauer eingehen.

Die *Einstellungen der Lehrer* zum Fach Erdkunde sind für die Unterrichtssituation mindestens ebenso interessant wie die der Schüler. Trotzdem sind empirische Untersuchungen hier relativ selten. Auf die Arbeit von *Bachmaier* haben wir bereits hingewiesen (*Bachmaier* 1969). Daneben gibt es eine Erhebung über Lehrermeinungen zu Unterrichtsmedien, Curricula und berufsbezogene Ausbildung (*Cloß* 1977). Im Rahmen des RCFP wurden die Erprobungslehrer über Reformbereitschaft hinsichtlich des Faches Erdkunde befragt (*Fürstenberg/Jungfer* 1979). Von anderer Seite wurden diese Daten vertiefenden Analysen unterworfen, die jedoch eher fachgeo-

graphische als fachdidaktische Zielrichtung hatten (*Reichert* 1981). Und eine ebenfalls eher fachgeographische Arbeit untersucht die Diffusion der Reformvorstellungen des RCFP in der bayerischen Lehrerschaft (*Dolansky* 1980).

(4) Analysen von Unterrichtsformen und Unterrichtsmedien

Hierzu gehören teilweise die vorhin schon erwähnten Untersuchungen über Lernweganalysen bei programmiertem Unterricht. Daneben gibt es aber auch einige spezielle Arbeiten über Unterrichtsprozesse im traditionellen Erdkundeunterricht, wobei teilweise die Erhebung quantitativer Daten über Videorecorder erfolgte (*Jäger* 1977, *Haubrich/Nebel* 1977, *Watzka* 1977). Die empirische Analyse von Unterrichtsmedien umfaßt beispielsweise auch Arbeiten zur quantitativen Inhaltsanalyse von Schulbüchern (*Weber* 1974).

(5) *Empirische Untersuchungen zu methodischen Problemen der Unterrichtsforschung zum Fach Erdkunde*

Wie *Haubrich* schon 1976 feststellte (*Haubrich* 1977a; b), besteht eine wesentliche Forschungslücke der quantitativen Geographiedidaktik darin, daß nur selten untersucht wurde, ob und wie sich quantitative Methoden aus Nachbardisziplinen für die Unterrichtsforschung im Fach Erdkunde einsetzen lassen. Das gilt ganz besonders, wenn man sich auf empirische Arbeiten zu diesem Transferproblem konzentriert: So lassen sich zwar einige theoretisch-methodische bzw. konzeptionelle Arbeiten finden (vergl. *Schanz* 1973 / *Jäger* 1977 / *Haubrich*; *Nebel* 1977 / *Schrettenbrunner* 1978), *empirische* Untersuchungen über *methodische* Fragen sind dagegen ziemlich selten. Eine der wenigen ist *Köcks* Untersuchung über Auswahlantwortaufgaben in lernzielorientierten, erdkundlichen Klassenarbeiten (*Köck* 1977). Hier vergleicht er in einer experimentellen Versuchsanordnung „gebundene" und „freie" Aufgabenstellungen bei einer Klassenarbeit. Mit seiner Studie kann er empirisch nachweisen, daß „gebundene" Fragestellungen (Vorgabe der Antwortkategorien) keine höhere Objektivität und Eindeutigkeit der Meßergebnisse erbringt als „freie" Aufgabenstellungen – zumindest bei lernzielorientierten Testaufgaben.

Eine andere methodische Arbeit stammt von *Zwirner*, der sich der Problematik von geographischen Tests widmet (*Zwirner* 1978).

1.3 Methodische Mängel vorliegender empirischer Untersuchungen

Bei der Durchsicht der für die vorliegende Arbeit interessanten empirischen Untersuchungen können wir folgendes feststellen:

Bis etwa 1975 waren empirische Untersuchungen zum Fach Geographie äußerst selten und auf sehr niedrigem methodischen Niveau. Allerdings gab es eine Reihe von empirischen Arbeiten zur allgemeinen Didaktik, in denen u. a. auch auf das Fach Erdkunde Bezug genommen wurde (vgl. *Bachmair* 1969 / *Seelig* 1968), und die ein vergleichsweise höheres methodisches Niveau hatten oder zumindest einen solchen Anspruch erhoben. Ein Sonderfall sind die ganz frühen empirischen Arbeiten (von 1905 bis 1925) über die Beliebtheit von Schulfächern.

Nach 1975 ist dann ein allgemeiner Aufschwung empirischer Arbeiten zu verzeichnen. Das gilt für alle thematischen Bereiche, besonders für Untersuchungen zum räumlichen Denken und über Unterrichtsformen bzw. Unterrichtsmedien. Empirische Arbeiten über methodische Probleme in der geographiedidaktischen Forschung sind fast ausschließlich nach 1975 entstanden. Für alle diese Arbeiten gilt, daß selbst die kompliziertesten multiplen und multivariaten statistischen Analyseverfahren eingesetzt wurden, wie Faktorenanalysen, Diskriminanzanalysen, multiple Regressionen, multidimensionale Skalierungen, multivariate Varianzanalysen usw.

Bedeutet diese Entwicklung nun, daß der methodische Nachholbedarf in der geographiedidaktischen Forschung aufgeholt ist? Man muß diese Frage leider verneinen. Zwar wurden in letzter Zeit verstärkt komplizierte Analyseverfahren benutzt, aber bei genauerem Hinsehen zeigt sich, daß diese Verfahren häufig völlig unreflektiert eingesetzt wurden: Teils entsprach die Datenbasis nicht den Voraussetzungen dieser Verfahren, teils bestehen erhebliche Zweifel, ob die Verfahren überhaupt richtig durchgeführt und interpretiert wurden. Häufig läßt sich eine erhebliche Diskrepanz zwischen Design der Untersuchung und den eingesetzten Analysetechniken feststellen. Außerdem sind die meisten Stichproben unzureichend, und zwar sowohl von der inneren Zusammensetzung als auch vom Umfang her. Auch die mit einem experimentellen Design arbeitenden Untersuchungen (vgl. *Schrettenbrunner* 1981) benutzen meist nur die allereinfachsten Versuchspläne (z. B. ohne Kontrollgruppen), die für eine strenge experimentelle Beweislogik kaum ausreichen. Ein weiterer methodischer Mangel besteht darin, daß bei keiner der neueren empirischen Untersuchungen eine Reliabilitäts- und Validitätsüberprüfung der eingesetzten Variablen (Items) stattfand. Schließlich werden die Ergebnisse komplexer statistischer Analysen sehr häufig nur in einer mathematisch-statistischen Fachterminologie präsentiert, die verhindert, daß auch Nichtmethodiker dieser Resultate rezipieren und kritisieren können. Andererseits kommt es aber auch vor, daß Ergebnisse quantitativer Analysen ohne jede Dokumen-

tation veröffentlicht werden, so daß eine Überprüfung der Berechnungen unmöglich ist.

Dies soll nun an konkreten Beispielen belegt werden. Dabei kommt es uns nicht auf eine streng systematische Bestandsaufnahme methodischer Mängel in den vorliegenden Arbeiten an. Wir haben statt dessen fünf Schwerpunkte gesetzt, die nach unserer Auffassung die kritischen „Angelpunkte" der methodischen Weiterentwicklung sind.

(1) *Diskrepanzen zwischen Datenbasis und verwendeten statistischen Verfahren*

Eine sinnvolle Verwendung der modernen statistischen Analyseverfahren hat eine ganze Reihe von methodischen Voraussetzungen, die sich aus der Logik der Verfahren und der Art und Qualität der verwendeten Daten ergeben.

Vor allem kann kein statistisches Verfahren aus den Daten mehr an Information „herausholen", als in sie durch das Erhebungsdesign oder die Versuchsanlage „hineingesteckt" wurde. Das klingt so selbstverständlich, daß man es kaum hinzuschreiben wagt. Nur leider sind unter den von uns durchgesehenen Arbeiten einige, die genau dies nicht beachten, und durch besonders komplizierte statistische Analyseverfahren ***nachträglich*** mehr in die Daten hineinzudeuten versuchen, als in ihnen steckt.

Sehen wir uns zu diesem Problem *Leusmanns* Untersuchung über „Schülereinstellungen zum Fach Erdkunde, zu Unterrichtstoffen und zu fachspezifischen Erarbeitungsformen" an (*Leusmann* 1977).

Leusmanns Datensatz umfaßt 261 Gymnasialschüler aus den Klassenstufen 7, 9, 11 und 13. Die Daten stammen aus drei verschiedenen Schulen (in Bonn, St. Augustin bei Bonn und in Hamburg).

Mit diesem Datensatz führt *Leusmann* verschiedene komplizierte statistische Analysen durch, u. a. Faktorenanalysen eines Eindrucksdifferentials zum Fach Erdkunde mit 24 Eigenschafts-Polaritäten und je 7-stufiger Skala. Zunächst faktorisiert *Leusmann* das Einstellungsprofil für den Gesamtdatensatz, um die dimensionale Struktur eines „Mittleren Beurteilers" zu erhalten. Er verwendet dazu eine Hauptfaktorenanalyse mit Kommunalitätenschätzung und anschließender obliquer Quartimin-Rotation.

Da *Leusmann* die Hypothese hat, daß sich die dimensionale Struktur der Einstellungen zum Fach Erdkunde durch Reifeprozesse mit den verschiedenen Klassenstufen ändert, spaltet er nun seinen Datensatz in vier Teilstrichproben auf, die jeweils die ***gleiche Klassenstufe*** umfassen. Mit diesen vier Teilstichproben führt *Leusmann* dann nochmals vier gesonderte Faktorenanalysen der gleichen Art wie oben durch. Dadurch kann er überprüfen – oder glaubt zumindest, es überprüfen zu können – ob sich auf allen Klassenstufen ein vergleichbarer dimensionaler Beurteilungsraum ergibt.

Leusmann geht dann sogar noch einen Schritt weiter. Er stellt die vier Faktorenanalysen für die einzelnen Klassenstufen nicht nur einfach neben-

1.3 Methodische Mängel vorliegender empirischer Untersuchungen

Bei der Durchsicht der für die vorliegende Arbeit interessanten empirischen Untersuchungen können wir folgendes feststellen:

Bis etwa 1975 waren empirische Untersuchungen zum Fach Geographie äußerst selten und auf sehr niedrigem methodischen Niveau. Allerdings gab es eine Reihe von empirischen Arbeiten zur allgemeinen Didaktik, in denen u. a. auch auf das Fach Erdkunde Bezug genommen wurde (vgl. *Bachmair* 1969 / *Seelig* 1968), und die ein vergleichsweise höheres methodisches Niveau hatten oder zumindest einen solchen Anspruch erhoben. Ein Sonderfall sind die ganz frühen empirischen Arbeiten (von 1905 bis 1925) über die Beliebtheit von Schulfächern.

Nach 1975 ist dann ein allgemeiner Aufschwung empirischer Arbeiten zu verzeichnen. Das gilt für alle thematischen Bereiche, besonders für Untersuchungen zum räumlichen Denken und über Unterrichtsformen bzw. Unterrichtsmedien. Empirische Arbeiten über methodische Probleme in der geographiedidaktischen Forschung sind fast ausschließlich nach 1975 entstanden. Für alle diese Arbeiten gilt, daß selbst die kompliziertesten multiplen und multivariaten statistischen Analyseverfahren eingesetzt wurden, wie Faktorenanalysen, Diskriminanzanalysen, multiple Regressionen, multidimensionale Skalierungen, multivariate Varianzanalysen usw.

Bedeutet diese Entwicklung nun, daß der methodische Nachholbedarf in der geographiedidaktischen Forschung aufgeholt ist? Man muß diese Frage leider verneinen. Zwar wurden in letzter Zeit verstärkt komplizierte Analyseverfahren benutzt, aber bei genauerem Hinsehen zeigt sich, daß diese Verfahren häufig völlig unreflektiert eingesetzt wurden: Teils entsprach die Datenbasis nicht den Voraussetzungen dieser Verfahren, teils bestehen erhebliche Zweifel, ob die Verfahren überhaupt richtig durchgeführt und interpretiert wurden. Häufig läßt sich eine erhebliche Diskrepanz zwischen Design der Untersuchung und den eingesetzten Analysetechniken feststellen. Außerdem sind die meisten Stichproben unzureichend, und zwar sowohl von der inneren Zusammensetzung als auch vom Umfang her. Auch die mit einem experimentellen Design arbeitenden Untersuchungen (vgl. *Schrettenbrunner* 1981) benutzen meist nur die allereinfachsten Versuchspläne (z. B. ohne Kontrollgruppen), die für eine strenge experimentelle Beweislogik kaum ausreichen. Ein weiterer methodischer Mangel besteht darin, daß bei keiner der neueren empirischen Untersuchungen eine Reliabilitäts- und Validitätsüberprüfung der eingesetzten Variablen (Items) stattfand. Schließlich werden die Ergebnisse komplexer statistischer Analysen sehr häufig nur in einer mathematisch-statistischen Fachterminologie präsentiert, die verhindert, daß auch Nichtmethodiker dieser Resultate rezipieren und kritisieren können. Andererseits kommt es aber auch vor, daß Ergebnisse quantitativer Analysen ohne jede Dokumen-

tation veröffentlicht werden, so daß eine Überprüfung der Berechnungen unmöglich ist.

Dies soll nun an konkreten Beispielen belegt werden. Dabei kommt es uns nicht auf eine streng systematische Bestandsaufnahme methodischer Mängel in den vorliegenden Arbeiten an. Wir haben statt dessen fünf Schwerpunkte gesetzt, die nach unserer Auffassung die kritischen „Angelpunkte" der methodischen Weiterentwicklung sind.

(1) *Diskrepanzen zwischen Datenbasis und verwendeten statistischen Verfahren*

Eine sinnvolle Verwendung der modernen statistischen Analyseverfahren hat eine ganze Reihe von methodischen Voraussetzungen, die sich aus der Logik der Verfahren und der Art und Qualität der verwendeten Daten ergeben.

Vor allem kann kein statistisches Verfahren aus den Daten mehr an Information „herausholen", als in sie durch das Erhebungsdesign oder die Versuchsanlage „hineingesteckt" wurde. Das klingt so selbstverständlich, daß man es kaum hinzuschreiben wagt. Nur leider sind unter den von uns durchgesehenen Arbeiten einige, die genau dies nicht beachten, und durch besonders komplizierte statistische Analyseverfahren *nachträglich* mehr in die Daten hineinzudeuten versuchen, als in ihnen steckt.

Sehen wir uns zu diesem Problem *Leusmanns* Untersuchung über „Schülereinstellungen zum Fach Erdkunde, zu Unterrichtstoffen und zu fachspezifischen Erarbeitungsformen" an (*Leusmann* 1977).

Leusmanns Datensatz umfaßt 261 Gymnasialschüler aus den Klassenstufen 7, 9, 11 und 13. Die Daten stammen aus drei verschiedenen Schulen (in Bonn, St. Augustin bei Bonn und in Hamburg).

Mit diesem Datensatz führt *Leusmann* verschiedene komplizierte statistische Analysen durch, u. a. Faktorenanalysen eines Eindrucksdifferentials zum Fach Erdkunde mit 24 Eigenschafts-Polaritäten und je 7-stufiger Skala. Zunächst faktorisiert *Leusmann* das Einstellungsprofil für den Gesamtdatensatz, um die dimensionale Struktur eines „Mittleren Beurteilers" zu erhalten. Er verwendet dazu eine Hauptfaktorenanalyse mit Kommunalitätenschätzung und anschließender obliquer Quartimin-Rotation.

Da *Leusmann* die Hypothese hat, daß sich die dimensionale Struktur der Einstellungen zum Fach Erdkunde durch Reifeprozesse mit den verschiedenen Klassenstufen ändert, spaltet er nun seinen Datensatz in vier Teilstrichproben auf, die jeweils die *gleiche Klassenstufe* umfassen. Mit diesen vier Teilstichproben führt *Leusmann* dann nochmals vier gesonderte Faktorenanalysen der gleichen Art wie oben durch. Dadurch kann er überprüfen – oder glaubt zumindest, es überprüfen zu können – ob sich auf allen Klassenstufen ein vergleichbarer dimensionaler Beurteilungsraum ergibt.

Leusmann geht dann sogar noch einen Schritt weiter. Er stellt die vier Faktorenanalysen für die einzelnen Klassenstufen nicht nur einfach neben-

einander und vergleicht die sich ergebenden Faktoren durch eine inhaltliche Interpretation. Vielmehr wendet er ein weiteres quantitatives Verfahren an, um die Lösungen der vier Faktorenanalysen objektiv vergleichen zu können, nämlich eine Zielrotation der Ladungsmatrizen (eine sog. „Transformationsanalyse") (vgl. *Rhenius* 1974; *Tarnai* 1978).

Diese eben skizzierte Vorgehensweise wäre im Prinzip methodisch durchaus sinnvoll. Nur scheint *Leumann* (aus Begeisterung für die raffinierten Verfahren vielleicht?) vergessen zu haben, auf welche Datenbasis er diese Vorgehensweise anwendet.

Vergegenwärtigen wir uns nochmals seine Erhebung. 261 Schüler insgesamt umfaßt sein Datensatz. Bei der oben dargestellten zweiten Auswertung benutzt er jedoch nicht diese Stichprobe, sondern die vier Teilstichproben entsprechend den Klassenstufen. Im günstigsten Fall, d. h. also bei völlig gleicher Aufteilung auf die Klassenstufen, besteht *Leusmanns* Datenbasis aus je 65 Schülern, die in Bezug auf die Klassenstufen-Variable homogenisiert sind.

Damit verletzt *Leusmann* zumindest zwei Voraussetzungen für den sinnvollen Einsatz einer Faktorenanalyse:

- Die Personenstichproben (65 Schüler) sind im Verhältnis zur Variablen-Stichprobe (24 Items des Polaritätsprofils) zu klein.
- Die Stichproben sind mit ziemlich hoher Wahrscheinlichkeit im Hinblick auf die Schülereinstellungen extrem verzerrt.

Es gibt eine allgemein anerkannte Faustregel für die Relation zwischen Variablen- und Probandenzahl bei Faktorenanalysen (*Pawlik* 1968, übernommen von *Lukesch* 1974): Die „Anzahl der Versuchspersonen (sollte) das Dreifache der Anzahl der Variablen betragen" (*Lukesch* 1974, S. 283). Das von *Leusmann* untersuchte Profil umfaßt 24 Items, er müßte also mindestens 72 Probanden haben. Dabei ist aber noch zu berücksichtigen, daß sich diese Faustregel auf *normale* Zufallsstichproben bezieht und nicht auf künstlich zusammengestellte Teilstichproben, wie sie *Leusmann* untersucht. Außerdem ist anzunehmen, daß *Leusmanns* Teilstichproben nicht genau gleich groß ausgefallen sind, so daß die notwendige Stichprobengröße vermutlich erheblich unterschritten wurde.

Ein schwerer Mangel von *Leusmanns* Datenbasis liegt jedoch nicht allein in der absoluten Größe der Stichprobe, sondern vor allem in ihrer Struktur, die auf Grund theoretischer Überlegungen zu bewerten ist.

- *Leusmann* untersucht nur Gymasiasten. Nichts berechtigt zu der Annahme, daß seine Ergebnisse beispielsweise für Gesamtschüler auch gelten.
- Er berücksichtigt nur Schulen aus zwei Bundesländern. Die Lehrpläne der Bundesländer im Fach Erdkunde sind sehr unterschiedlich. Deshalb ist zweifelhaft, ob seine Ergebnisse z. B. auch für ein bayerisches Gymnasium aussagekräftig sind.

- Die 65 Schüler seiner Teilstichproben verteilen sich auf nur drei verschiedenen Schulen. Also tragen – im günstigsten Fall – ca. 21 Schüler je Schule und Klassenstufe zum Ergebnis der Faktorenanalyse bei. D. h.: nur 21 Schüler sind in Bezug auf die Variablen Klassenstufe und Schule homogen.
- Daraus ergibt sich auch, daß pro untersuchter Klassenstufe und Schule wahrscheinlich nur ein bestimmter Lehrer den Erdkundeunterricht der Schüler gestaltet hat. Jede Klassenstufe wurde also, im günstigsten Fall, überhaupt nur von drei verschiedenen Lehrern unterrichtet. Bei nur drei verschiedenen Lehrern fällt aber der Unterrichtsstil jedes einzelnen Lehrers viel zu stark ins Gewicht. Einstellungsunterschiede je Klassenstufen könnten deshalb Ergebnis des spezifischen Unterrichtsstils dieser drei Lehrer sein.
- Besonders deutlich zeigt sich die Verzerrung von *Leusmanns* Stichprobe, wenn man bedenkt, daß pro untersuchter Klassenstufe auch nur drei Klassengemeinschaften, jede in einer anderen Schule, in die Faktorenanalyse eingegangen sind. Wenn nun *Leusmann* rein zufällig eine aufgeschlossene Klasse der Unterstufe und zwei desinteressierte der Oberstufe aufgenommen hat, könnten sozusagen Zweidrittel der Einstellungen eine Folge der Variablen Klassengemeinschaft und nicht eine Folge der Klassenstufe sein. Man hätte also die 65 Schüler je Klassenstufe aus *verschiedenen* Klassengemeinschaften *zufällig* auswählen müssen – oder es wäre notwendig gewesen, insgesamt *mehrere* Klassengemeinschaften je Klassenstufe zu berücksichtigen.

Gerade der zuletzt erwähnte Punkt zeigt erneut, daß es gar nicht unbedingt auf die absolute Größe der Stichprobe ankommt, sondern auf deren strukturelle Zusammensetzung.

Kein wie immer geartetes statistisches Prüfverfahren kann die genannten Mängel beheben. Es ist deshalb völlig unberechtigt, wenn *Leusmann* seinen Untersuchungen eine „gewisse statistisch abgesicherte Gültigkeit und Allgemeingültigkeit" zuschreibt. (*Leusmann*, 1977, S. 147).

Zwischen der relativ schlechten Datenbasis *Leusmanns* und den verwendeten Analyseverfahren besteht also ein krasser Gegensatz: Sein quantitativer Vergleich der Ladungsmatrizen aus den vier Faktorenanalysen mittels eines „Zielrotationsverfahrens" suggeriert Scheingenauigkeit. In Wirklichkeit sind bereits die Ergebnisse der vier Faktorenanalysen wegen der nicht kontrollierten Variablen „Klassengemeinschaft" und „Lehrereffekt", sowie der nicht randomisierten Variablen „Schule", höchstwahrscheinlich völlig verzerrt. Ganz abgesehen davon, daß man mit 65 Fällen eben keine Faktorenanalyse von 24 Items vornehmen darf, wenn man irgenwelche inferrenzstatistischen Schlüsse ziehen will. Um es nochmals in aller Deutlichkeit zu sagen: Zwei unbrauchbare Rechenergebnisse werden auch dann nicht sinnvoll, wenn man sie mit einem raffinierten quantitativen Verfahren auf Unterschiede hin untersucht.

Es könnte nun der Eindruck entstanden sein, wir würden es mit dem Erhebungsdesign übergenau nehmen. Die Stichprobe von *Leusmann* sei im Vergleich mit anderen doch relativ gut. Letzteres trifft leider zu.

Es gibt empirische Erhebungen in der geographiedidaktischen Forschung, die weitaus schlimmere Verzerrungen aufweisen als die oben besprochene von *Leusmann*. Ein Beispiel ist eine Lehrerbefragung von *Cloß* und *Sperling* (*Cloß/Sperling* 1977) über Curricula, Medieneinsatz und berufsbezogene Ausbildung bei 152 Lehrern. *Andree Kilchenmann* sagte in einer Diskussion über diese Stichprobe klipp und klar: „Ich würde sagen, daß ihre Stichprobe eigentlich sehr uninteressant ist. Sie enthält Geographielehrer, die sie angeschrieben haben, die aus irgendwelchen Gründen bereit und fähig waren, den Fragebogen zurückzuschicken. Dies ist eine nicht sehr interessante Stichprobe." (*Haubrich* 1977).

Dem ist nichts hinzuzufügen – außer, daß *Cloß* und *Sperling* mit ihrer nicht repräsentativen Datenbasis lediglich einfache Häufigkeitsauszählungen vornahmen und außerdem selbst auf die Verzerrung ihrer Stichproben hinwiesen. Sie versuchten auch nicht durch übertrieben komplexe statistische Analyseverfahren die Mängel ihrer Daten zu vertuschen. Genau diesen Eindruck hat man jedoch bei der oben besprochenen Arbeit von *Leusmann*.

Unser erster methodischer Kritikpunkt richtet sich also nicht gegen einfache Erhebungen an sich, sondern gegen die Diskrepanz zwischen statistischem Analysemodell und Datenbasis. In ihr sehen wir die Gefahr einer Überbewertung der statistischen Raffinesse auf Kosten der inhaltlichen Substanz der Aussagen.

(2) *Widersprüche zwischen der Variablenauswahl und dem Erhebungsdesign, Vernachlässigung wichtiger Strukturvariablen*

Im Jahre 1905 veröffentlichte *William Stern* seine schon erwähnte bemerkenswerte Untersuchung über „Beliebtheit und Unbeliebtheit der Schulfächer" (*Stern* 1905). Seine Analyse beruht nur auf Häufigkeitsauszählungen der abhängigen Variablen für verschiedene Schülergruppen. Seine abhängigen Variablen bestanden in den Fragen „Welches Fach (welche Stunde) hast Du (haben Sie) am liebsten?" und „Welches Fach ist Dir (Ihnen) am wenigsten lieb?" Ausgezählt wurde die relative Häufigkeit der Nennungen einzelner Schulfächer für folgende Schülergruppen:

- Für Knaben der Volkschule,
- für Mädchen der Volksschule,
- für Mädchen der Höheren Schule,
- für die jeweiligen Alterstufen (Oberstufe, Mittelstufe, Unterstufe) bei Knaben der Volksschule,
- für die Altersstufen bei Mädchen der Volksschule,
- für altersgleiche Mädchen je nach der Volksschule und der höheren Schule,

– für altersgleiche Mädchen je nach Land- und Stadtschule und
– für altersgleiche Jungen je nach Land- und Stadtschule.

Heute würde man sagen, *Stern* führte eine mehrdimensionale Kreuztabellenanalyse durch, wobei nach Geschlecht, Klassenstufe, Schultyp und Stadt-Land-Schule aufgegliedert wurde.

2556 Schüler insgesamt waren in der Stichprobe, zusätzlich 141 Schülerinnen eines Lehrerinnenseminars, sozusagen als Kontrollgruppe. *Sterns* Variablensatz bestand praktisch nur aus der abhängigen Variablen „Beliebtheit des Faches" und den erklärenden Strukturvariablen „Geschlecht", „Klassenstufe", „Schultyp" und „Stadt-Land-Gegensatz".

Seine Analyse mit diesen Variablen ergab u. a., daß weibliche Volksschüler das Fach Erdkunde seltener als beliebtestes Fach ausgewählt hatten als männliche Volksschüler. Außerdem wurde es von weiblichen Volksschülern wesentlich häufiger als unbeliebtestes Fach bezeichnet als von männlichen (vgl. *Stern* 1905, S. 277)[2].

Interessant ist auch, daß bei männlichen Schülern der Volksschule sich keine altersspezifischen Unterschiede in der Beliebtheit des Faches Geographie ergaben. Bei Mädchen dagegen nahm die Unbeliebtheit des Faches Geographie leicht ab (vgl. *Stern* 1905, S. 289).[3]

Wir haben hier zwei Ergebnisse dieser frühen Studie referiert, weil sich an ihnen zeigen läßt, daß Art und Auswahl der Variablen in einem vernünftigen Verhältnis zur Datenbasis stehen sollten, wenn man vertrauenswürdige Resultate erhalten möchte. Gewiß, die Analysen von *Stern* verblassen, wenn man heutige Maßstäbe anlegen würde. Heute ließe sich beispielsweise die Beliebtheit der Fächer mit Fragenbatterien sehr viel genauer erfassen. Aber *Sterns* Erklärung seiner, relativ einfach erfaßten, abhängigen Variablen fußt auf einem Satz unabhängiger Strukturvariablen, die alle mit einer soliden empirischen Grundlage versehen sind.

Eine solch vorsichtige Haltung bei der Datenanalyse ist bei vielen modernen Untersuchungen verloren gegangen. Die teilweise erhebliche Diskrepanz zwischen den eingesetzten statistischen Analyseverfahren und der Datenbasis wurde bereits erwähnt.

Darüberhinaus besteht eine Kluft zwischen den benutzten Variablen und ihrer empirischen Absicherung. Diese Behauptung läßt sich belegen, wenn man die Untersuchung von *Stern* mit einer neueren Arbeit zum selben Thema konfrontiert: z. B. mit *Gerd Bachmairs* Analysen der „Einstellungen von Schülern zum Lehrer und zum Unterrichtsfach" (*Bachmair* 1969).

Bachmair untersuchte eine unübersehbare Vielzahl von Einstellungs-Variablen der Schüler zu verschiedenen Schulfächern, zu didaktischen Hilfsmitteln, zur Schule allgemein, zur Beliebtheit des Lehrers, zum Aussehen des Lehrers und seinen Verhaltensweisen. Die Liste der Variablen umfaßte weiterhin: Schülereinstellungen . . .

– zum Einfühlungsvermögen der Lehrer,
– zur Unterrichtsführung der Lehrer,

- zu dominantem Unterrichtsverhalten der Lehrer,
- zur Arbeitsbelastung durch die Lehrer und zu deren Prüfungsverhalten,
- zur Unterrichtsorganisation der Lehrer und
- zum Unterrichtsfach selbst.

Diese Variablen wurde in einem umfangreichen Fragebogen mit insgesamt 11 Polaritätsprofilen, einer Vielzahl von Einzelfragen mit Rating-Skala und von einer Einstellungbatterie mit fast 50 Statements erhoben. Die Schüler waren bis zu 90 Minuten mit dessen Ausfüllung beschäftigt. „Insgesamt gaben die Schüler 133.096 Urteile ab" (*Bachmair* 1969, S. 89).

Dieser Berg von Daten zu allen möglichen Variablen wurde bei 8 verschiedenen Klassen an nur 4 Schulen erhoben. Die Stichprobe umfaßte 60 Mädchen (!) und 132 Jungen, wobei ganze 4 Mädchen (!) aus einer Mittelstufe kamen und nur 25 aus einer Oberstufe. An keiner einzigen der vier Schulen wurden über alle Klassenstufen Daten erhoben. Alle Schulen waren Gymnasien.

Hier die Beschreibung der Stichprobe, entnommen aus der Arbeit von *Bachmair* (S. 87, 89):

	Knaben	Mädchen	Summe
Unterstufe	17	31	48
Mittelstufe	80	4	84
Oberstufe	35	25	60
Summe	132	60	192

Tab.1: Stichprobe von Bachmair nach Klassenstufe und Geschlecht

Diese wohl mehr als dürftige Datenbasis hindert *Bachmair* aber nun keineswegs daran, z. B. Aufgliederungen nach „Klassenstufe" vorzunehmen:

Wir übernehmen hier *Bachmairs* Tabelle zur Variablen „Arbeitsbelastung" der Schüler je nach Klassenstufe für verschiedene Fächer (*Bachmair* 1969, S. 166):

	Deutsch		Mathematik		Englisch		Musik		Erdkunde	
	$\bar{x}$	σ^2	$\bar{x}$	σ^2	$\bar{x}$	σ^2	$\bar{x}$	σ^2	$\bar{x}$	σ^2
Unterstufe	4,0	4,6	3,9	4,7	4,0	4,9	2,6	4,1	4,3	4,7
Mittelstufe	2,6	3,2	4,2	4,0	4,4	4,0	1,9	3,0	3,1	3,4
Oberstufe	2,8	2,9	4,8	4,3	4,7	3,4	2,0	2,1	3,2	3,3

Tab. 2: Arbeitsbelastung nach Klassenstufe und Schulfach (nach *Bachmair*)

Nur wenige Zeilen darunter bringt *Bachmair* eine Aufgliederung nach Geschlecht, aus der sich ergibt, daß sich Mädchen allgemein stärker belastet fühlen als Jungen. An ihr sieht man, daß *Bachmairs* Aufgliederung der Arbeitsbelastung nach Klassenstufe schlicht und einfach eine Vorspiegelung falscher Tatsachen ist. Die Geschlechterproportionen sind je nach Klassenstufe extrem unterschiedlich: zum Beispiel 4 Mädchen und 80 Jungen in der Mittelstufe.

Was *Bachmair* in seiner Aufgliederung nach „Klassenstufe" präsentiert, ist also nicht der „Klasseneffekt" – wie er ausdrücktlich behauptet – sondern primär ein Geschlechtseffekt.

Es kommt aber noch schlimmer: In einem Kapitel über die „Determinanten der Schülereinstellungen" untersucht er in einem ersten Abschnitt den „Einfluß des Alters", um entwicklungspsychologische Vorgänge aufzudecken (*Bachmair* 1969, S. 296).

Die Ergebnisse werden mit großem Enthusiasmus interpretiert; zum Beispiel: „Im typischen Pubertätsalter stehen die Schüler der Mittelstufe. Die daraus resultierenden Konflikte finden sich auch in den Einstellungen wieder" (S. 302). „Der Unmut nimmt schlagartig bei den Schülern der 8. und 9. Klassen (bei *Bachmair* die Mittelstufe) zu" . . . „Die Abneigung gegen die Schule nimmt in der Oberstufe, den Klassen 10 und 11, wieder leicht ab" (ebd. S. 297).

Kein Wort davon, daß die Mittelstufenschüler praktisch nur Jungen sind, während die Unterstufenschüler zu fast Zweidrittel aus Mädchen bestehen. Dabei bemerkt *Bachmair* zu Recht im nächsten Abschnitt über den Einfluß des Geschlechts auf die Schülereinstellungen: „Knaben und Mädchen unterscheiden sich in ihren schulischen Einstellungen" (ebd. S. 307).

Fassen wir zusammen:

Es gibt eine Reihe von grundlegenden erklärenden *Strukturvariablen*, die mit vielen geographiedidaktisch interessierenden Variablen – vorsichtig formuliert – irgendwie zusammenhängen: Dazu gehören z. B. das Geschlecht, die Altersstufe, die Zugehörigkeit zu einer Klassengemeinschaft, der Schultyp usw.. Ein Untersuchungsdesign muß diese Variablen auf alle Fälle berücksichtigen. Dies kann durch eine spezielle Form der *experimentellen* Anlage geschehen oder durch eine sorgfältige Stichprobenerstellung, bei der alle diese Strukturvariablen wenigstens durch genügend Fälle abgesichert sind. Bei einem experimentellen Design werden die Strukturvariablen als Faktoren jeweils konstant gehalten, so daß sich keine Verzerrungen bei der Analyse ergeben. Die meisten empirischen Untersuchungen zur Geographiedidaktik benutzen jedoch kein experimentelles Erhebungsdesign, sondern „survey"-ähnliche Erhebungen. Bei dieser Art von Datenerhebung ist es dann jedoch unbedingt notwendig, daß die Strukturvariablen explizit bei der Stichprobenerstellung und bei der Analyse berücksichtigt werden. Nur durch die (mehrdimensionale) Aufgliederung des Datensatzes anhand dieser Strukturvariablen ist es möglich, Interaktionseffekte

zwischen ihnen auszuschalten, die andernfalls alle Ergebnisse der Untersuchung verzerren würden.

(3) *Die unnötige Bevorzugung metrischer Analyseverfahren*

Die Bevorzugung metrischer Analyseverfahren in allen von uns durchgesehenen Arbeiten hat folgende Gründe:

Erstens sind die meisten statistischen Verfahren zur Analyse metrischer Daten seit langem bekannt und in Form der verschiedenen Statistik-Software-Pakete (wie SPSS, BMDP, OSIRIS usw.) leicht zugänglich.

Zweitens konnten mit solchen Verfahren unbestreitbare Verbesserungen in der empirischen Forschung erzielt werden. Um ein Beispiel zu erwähnen: Die Faktorenanalyse war die Basis zur Entwicklung umfangreicher Persönlichkeits- und Einstellungsinventarien und hat dadurch einen ganzen Zweig der Persönlichkeitspsychologie begründet. Ebenso ist die Intelligenzforschung bzw. die Testdiagnostik allgemein ohne die Impulse der Faktorenanalyse nicht denkbar – ja man kann sogar sagen, daß weite Gebiete der psychologischen Diagnostik erst durch sie entstanden sind.

Drittens waren statistische Verfahren für die Analyse *nichtmetrischer* Daten lange Zeit auf die Ebene *bivariater* Zusammenhänge beschränkt gewesen.

Erst seit gut zwei Jahrzehnten existieren statistische Modelle zur multiplen und multivariaten Zusammenhangsanalyse nichtmetrischer Daten, wie z. B. das GSK-Verfahren, das wir in dieser Arbeit vorstellen werden, oder die log-linearen Modelle. Die Verfahren wurden in den USA entwickelt und sind in der Bundesrepublik Deutschland erst vor wenigen Jahren erstmals rezipiert worden – allerdings nur von Methodenspezialisten, die sie auf mathematisch-statistischem Niveau diskutierten (z. B. *Bedall* 1974).

Ein vierter Grund für die Bevorzugung metrischer Variablen hängt mit dem vorhin diskutierten Problem der einseitigen Variablenauswahl zusammen. Nach unserem Eindruck besteht in der empirischen Forschung eine starke Tendenz zur primär psychologischen oder sozialpsychologischen Betrachtungsweise der Probleme. Dies gilt auch in den empirischen Geographiedidaktik. Ein gutes Beispiel ist wieder der schon erwähnte Aufsatz von *Leusmann* (*Leusmann* 1977): Hier wird die dimensionale Einstellungsstruktur von Schülern im Fach Erdkunde über ein semantisches Differential erfaßt und eingehend analysiert. Es geht dabei gar nicht darum, was die Schüler verschiedener Alterstufen konkret vom Fach Erdkunde halten, sondern es geht darum, ob ihr „Perzeptionsrahmen" für das Fach auf verschiedenen Klassenstufen konstant bleibt. Dies ist eine typisch kognitionspsychologische Analyseperspektive.

Da nach unserem Eindruck solche psychologischen Fragestellungen überwiegen, werden auch die Meßinstrumente verwendet, die primär in diesen Disziplinen entwickelt wurden, wie z. B. das semantische Differential, umfangreiche Statementbatterien oder ganze Tests zur Intelligenz oder

zum Wissen der Schüler. Diese Meßinstrumente produzieren typischerweise Daten auf metrischem Niveau, bzw. man unterstellt, daß die so gemessenen Variablen metrisches Niveau aufweisen. Die Art der verwendeten Erhebungsinstrumente legt also Analysetechniken auf metrischem Niveau oft nahe.

Wie wir gesehen haben, gibt es eine Reihe von verständlichen Ursachen dafür, daß nichtmetrische Verfahren bislang wenig beachtet wurden. Dennoch muß man die vorherrschenden Analysen metrischer Variablen als methodische Fehlentwicklung bezeichnen.

Denn erstens kommt es bei der Bevorzugung metrisch gemessener Variablen zu der schon genannten Blickverengung der Analyse. Viele sozialwissenschaftlich interessante Variablen – auch im Bezug auf das Unterrichtsgeschehen im Fach Erdkunde – sind nun einmal eindeutig nicht metrisch. Es handelt sich dabei um all jene Variablen, die durch Namen für verschiedenartige Dinge gekennzeichnet sind. Solche Nominalvariablen bestehen aus exklusiven Kategorien (deshalb auch Kategorialdaten genannt) wie „männlich" – „weiblich", „Hauptschule" – „Gymnasium", oder „Bundesland Hessen", „Niedersachsen" usw. Man kann diesen Kategorien nicht einfach Zahlen zuordnen, so daß die empirischen Relationen zwischen den Kategorien in den Relationen des natürlichen Zahlensystems abgebildet sind. Nominalvariablen, und dazu gehören vielfach auch die wichtigen Strukturvariablen, können deshalb prinzipiell nicht mit metrischen Verfahren analysiert werden.

Zweitens wird metrisches Datenniveau vielfach einfach unterstellt, um die vorhandenen metrischen Analysemethoden verwenden zu können, auch wenn sich jeder darüber im klaren ist, daß eigentlich allenfalls ordinale, wenn nicht gar nur nominale Meßinformation vorliegt. Das ist beispielsweise der Fall bei Ratingskalen oder Polaritätsprofilen, wenn nicht besondere Skalierungstechniken verwendet werden.

Es ist drittens höchst bedenklich, daß bei metrischen Variablen praktisch nie eine meßtheoretische Absicherung erfolgt. In der uns vorliegenden Literatur über empirische Untersuchungen zur Geographiedidaktik werden „metrische" Variablen ausschließlich mit den allereinfachsten Meßverfahren erhoben, nämlich mit einfachen „Ja-Nein"-Fragen – die zu Summen-Scores addiert werden – oder mit mehrstufigen Ratingskalen oder Polaritätsprofilen. Besondere Skalierungsmethoden (wie die Guttman-Skalierung, die Paarvergleichstechnik usw.), die speziell dafür entwickelt wurden, metrisches Niveau der Messung zu ermöglichen, wurden nirgends benutzt.

Zusammenfassend kann man festhalten, daß die Probleme metrischer Analysen heute in zunehmendem Maße gesehen werden. Immer häufiger werden statistische Analyseverfahren entwickelt, die nicht mehr ein metrisches Datenniveau voraussetzen. Die verschiedenen Techniken der nichtmetrischen multidimensionalen Skalierung sind dafür ein Beispiel. Sie benutzen nur den ordinalen Informationsanteil in den Meßwerten, d. h.

„Größer-Kleiner-Relationen", aber nicht die metrischen Abstände zwischen ihnen (*Shepard/Romney/Nerlove,* 1972).

Ein Mißverständnis besteht oft darin, daß metrisches Meßniveau bei den Variablen mit anspruchsvoller quantitativer Analyse schlechthin gleichgesetzt wird. Dem ist nicht so. Auch qualitative Daten (Nominaldaten und Ordinaldaten) können mittels komplizierter Rechenverfahren auf multiple und mulitvariate Zusammenhänge hin untersucht werden – wie wir in dieser Arbeit noch zeigen werden.

(4) *Fehlende Überprüfung von Meßsicherheit (Reliabilität) und Gültigkeit (Validität)*

Die Frage der Meßsicherheit, der sogenannten Reliabilität, stellt sich bei Variablen auf jedem Meßniveau: Bei metrischen Variablen genauso wie bei nichtmetrischen.

Es ist erstaunlich, daß Fragen der Reliabilität in der geographiedidaktischen Forschung der Bundesrepublik bislang praktisch überhaupt nicht Beachtung gefunden haben, denn vor mehreren Jahrzehnten war es unter anderem die empirische Unterrichtsforschung, die diese Probleme erstmals umfassend erarbeitet hat (vgl. *Cronbach* 1951). Damals war man vor allem mit dem Problem von umfassenden Testbatterien befaßt, mit denen Begabungs-, Leistungs- und Intelligenzunterschiede sicher gemessen werden sollten. Zu diesem Zweck wurde eine systematische Methodologie der Testkonstruktion und Testüberprüfung entwickelt.

Das bekannteste Resultat dieser frühen Bemühungen um die Meßsicherheit waren die teststatistischen Kennwerte wie „Itemschwierigkeit", „Trennschärfe", „Reliabilität" eines Items oder einer ganzen Skala, „Test-Retest-Reliabilität", „Homogenität" usw.

Diese Maße aus der klassischen Kennwertanalyse wurden in umfassender Weise von *Lienert* (*Lienert* 1969) dargestellt.

Vor allem in der experimentellen psychologischen Forschung wurden diese grundlegenden meßtheoretischen Verfahren weiterentwickelt und an spezielle Untersuchungsdesigns angepaßt. So existieren z. B. eine Vielzahl von Arbeiten über die Probleme der Meßsicherheit bei Versuchsanlagen mit Vor- und Nachtests (vgl. *Herbig* 1975 und *Klauer* 1972).

Versucht man nun diese, schon klassischen Grundlagen der Messung in geographiedidaktischen Untersuchungen aufzufinden, so erlebt man eine arge Enttäuschung. Zwar werden z. B. „lernzielorientierte Wissenstests" für geographische Inhalte häufig verwendet. Von einer teststatistischen Absicherung dieser Wissens-Tests kann jedoch nicht einmal in Ansätzen gesprochen werden. Man darf deshalb ohne Übertreibung von einem erheblichen Nachholbedarf der geographiedidaktischen Forschung bei meßtheoretischen Fragen sprechen.

Wir wollen diese, zugegeben harte Kritik an einem Beispiel belegen: *H. Nolzen* berichtet in einem Aufsatz über die „Entwicklung von Compu-

terprogrammen und Computertests und ihre Auswertung durch EDV" (*Nolzen* 1977). Dieser Beitrag schildert die Möglichkeiten des Computereinsatzes beim Unterricht in der Schule und an der Universität. Es ist darin viel die Rede von „Hard- und Software", von Dialogplätzen und Medienverbund, von Computersprachen und „Teachware" (S. 64). Ein Abschnitt ist einem sog. „Selbst-Test-Hilfe-Programm" gewidmet (S. 68) und ein anderer einer „Testbibliothek Geographie", die aus 14 solcher „Selbst-Test-Hilfe-Programme" besteht. Zur Testbibliothek bemerkt der Autor: „Über die Effizienz dieser Testprogramme im Rahmen des geographiedidaktischen Studienganges ist zu sagen, daß es seit etwa einem Jahr fast keinen Examenskandidaten gibt, der sich nicht am Computer getestet hat, sofern für seine Spezialgebiete Programme vorhanden sind" (S. 78 ebd.).

Nun mag es ja sehr erfreulich sein, wenn der Computer von den Studenten so intensiv genutzt wird. Man sollte dies aber doch wohl nicht als Kriterium für die Effizienz des Wissenstests ansehen. Wenn Begriffe noch irgend einen Sinn haben, dann muß sich die Qualität eines Tests danach bemessen, ob er gut „testet" und nicht daran, ob er bei den Schülern oder Studenten Anklang findet.

In *Noelzens* Aufsatz fehlt jede Diskussion darüber, ob die Studenten beim computergestützten Unterricht tatsächlich etwas lernen, ob sie *mehr* lernen als im normalen Unterricht, ob bestimmte Schülergruppen oder Altersstufen einen höheren Lerngewinn dabei erzielen, oder ob der Wissenszuwachs je nach Lerninhalten unterschiedlich hoch ausfällt. Diese Fragen lassen sich mit *Noelzens* Versuchsanordnung auch nicht klären: In seinem Computer-Testprogramm fehlt jegliche teststatistische Absicherung. Weder Trennschärfe noch Reliabilität der Testitems wurden überprüft.

Die Arbeit von *Noelzen* ist ein Beispiel dafür, wie man Lernvorgänge (im Bereich geographischen Wissens) gerade nicht wissenschaftlich aufklärt, sondern im Gegenteil, durch pseudoexakte Computerterminologie, mystifiziert. Sein Testprogramm entspricht in etwa dem des üblichen Fakten-„abfragens" durch den Lehrer. Weder existiert ein Wissens-Vortest, noch ist irgendwie abgesichert, daß mit den Testaufgaben auch tatsächlich ein lernzielrelevantes Wissen erfaßt wird und nicht z. B. die Fähigkeit geschickt zu „raten".

Ja, es ist nicht einmal auszuschließen, daß der Computer-Test in Wirklichkeit nur die Geschicklichkeit der Schüler mißt, mit dem Computerterminal umzugehen. Wenn Begriffe als Lösung eingegeben werden, genügt nämlich ein einziger falscher Buchstabe, damit die ganze Antwort vom Computer als „falsch" gewertet wird (vgl. S. 83 ebd.).

Die Gültigkeit einer Messung – die sog. Validität – baut auf ihrer Meßsicherheit (Reliabilität) auf.

Eine reliable Variable, also eine mit großer Meßsicherheit, mißt sicher und stabil, was sie messen soll, und nicht einmal dies und das andere Mal jenes. Eine valide Variable erfaßt darüberhinaus auch wirklich das Phänomen, das sie theoretisch repräsentieren soll.

„Größer-Kleiner-Relationen", aber nicht die metrischen Abstände zwischen ihnen (*Shepard/Romney/Nerlove,* 1972).

Ein Mißverständnis besteht oft darin, daß metrisches Meßniveau bei den Variablen mit anspruchsvoller quantitativer Analyse schlechthin gleichgesetzt wird. Dem ist nicht so. Auch qualitative Daten (Nominaldaten und Ordinaldaten) können mittels komplizierter Rechenverfahren auf multiple und mulitvariate Zusammenhänge hin untersucht werden – wie wir in dieser Arbeit noch zeigen werden.

(4) *Fehlende Überprüfung von Meßsicherheit (Reliabilität) und Gültigkeit (Validität)*

Die Frage der Meßsicherheit, der sogenannten Reliabilität, stellt sich bei Variablen auf jedem Meßniveau: Bei metrischen Variablen genauso wie bei nichtmetrischen.

Es ist erstaunlich, daß Fragen der Reliabilität in der geographiedidaktischen Forschung der Bundesrepublik bislang praktisch überhaupt nicht Beachtung gefunden haben, denn vor mehreren Jahrzehnten war es unter anderem die empirische Unterrichtsforschung, die diese Probleme erstmals umfassend erarbeitet hat (vgl. *Cronbach* 1951). Damals war man vor allem mit dem Problem von umfassenden Testbatterien befaßt, mit denen Begabungs-, Leistungs- und Intelligenzunterschiede sicher gemessen werden sollten. Zu diesem Zweck wurde eine systematische Methodologie der Testkonstruktion und Testüberprüfung entwickelt.

Das bekannteste Resultat dieser frühen Bemühungen um die Meßsicherheit waren die teststatistischen Kennwerte wie „Itemschwierigkeit", „Trennschärfe", „Reliabilität" eines Items oder einer ganzen Skala, „Test-Retest-Reliabilität", „Homogenität" usw.

Diese Maße aus der klassischen Kennwertanalyse wurden in umfassender Weise von *Lienert* (*Lienert* 1969) dargestellt.

Vor allem in der experimentellen psychologischen Forschung wurden diese grundlegenden meßtheoretischen Verfahren weiterentwickelt und an spezielle Untersuchungsdesigns angepaßt. So existieren z. B. eine Vielzahl von Arbeiten über die Probleme der Meßsicherheit bei Versuchsanlagen mit Vor- und Nachtests (vgl. *Herbig* 1975 und *Klauer* 1972).

Versucht man nun diese, schon klassischen Grundlagen der Messung in geographiedidaktischen Untersuchungen aufzufinden, so erlebt man eine arge Enttäuschung. Zwar werden z. B. „lernzielorientierte Wissenstests" für geographische Inhalte häufig verwendet. Von einer teststatistischen Absicherung dieser Wissens-Tests kann jedoch nicht einmal in Ansätzen gesprochen werden. Man darf deshalb ohne Übertreibung von einem erheblichen Nachholbedarf der geographiedidaktischen Forschung bei meßtheoretischen Fragen sprechen.

Wir wollen diese, zugegeben harte Kritik an einem Beispiel belegen: *H. Nolzen* berichtet in einem Aufsatz über die „Entwicklung von Compu-

terprogrammen und Computertests und ihre Auswertung durch EDV" (*Nolzen* 1977). Dieser Beitrag schildert die Möglichkeiten des Computereinsatzes beim Unterricht in der Schule und an der Universität. Es ist darin viel die Rede von „Hard- und Software", von Dialogplätzen und Medienverbund, von Computersprachen und „Teachware" (S. 64). Ein Abschnitt ist einem sog. „Selbst-Test-Hilfe-Programm" gewidmet (S. 68) und ein anderer einer „Testbibliothek Geographie", die aus 14 solcher „Selbst-Test-Hilfe-Programme" besteht. Zur Testbibliothek bemerkt der Autor: „Über die Effizienz dieser Testprogramme im Rahmen des geographiedidaktischen Studienganges ist zu sagen, daß es seit etwa einem Jahr fast keinen Examenskandidaten gibt, der sich nicht am Computer getestet hat, sofern für seine Spezialgebiete Programme vorhanden sind" (S. 78 ebd.).

Nun mag es ja sehr erfreulich sein, wenn der Computer von den Studenten so intensiv genutzt wird. Man sollte dies aber doch wohl nicht als Kriterium für die Effizienz des Wissenstests ansehen. Wenn Begriffe noch irgend einen Sinn haben, dann muß sich die Qualität eines Tests danach bemessen, ob er gut „testet" und nicht daran, ob er bei den Schülern oder Studenten Anklang findet.

In *Noelzens* Aufsatz fehlt jede Diskussion darüber, ob die Studenten beim computergestützten Unterricht tatsächlich etwas lernen, ob sie *mehr* lernen als im normalen Unterricht, ob bestimmte Schülergruppen oder Altersstufen einen höheren Lerngewinn dabei erzielen, oder ob der Wissenszuwachs je nach Lerninhalten unterschiedlich hoch ausfällt. Diese Fragen lassen sich mit *Noelzens* Versuchsanordnung auch nicht klären: In seinem Computer-Testprogramm fehlt jegliche teststatistische Absicherung. Weder Trennschärfe noch Reliabilität der Testitems wurden überprüft.

Die Arbeit von *Noelzen* ist ein Beispiel dafür, wie man Lernvorgänge (im Bereich geographischen Wissens) gerade nicht wissenschaftlich aufklärt, sondern im Gegenteil, durch pseudoexakte Computerterminologie, mystifiziert. Sein Testprogramm entspricht in etwa dem des üblichen Fakten-„abfragens" durch den Lehrer. Weder existiert ein Wissens-Vortest, noch ist irgendwie abgesichert, daß mit den Testaufgaben auch tatsächlich ein lernzielrelevantes Wissen erfaßt wird und nicht z. B. die Fähigkeit geschickt zu „raten".

Ja, es ist nicht einmal auszuschließen, daß der Computer-Test in Wirklichkeit nur die Geschicklichkeit der Schüler mißt, mit dem Computerterminal umzugehen. Wenn Begriffe als Lösung eingegeben werden, genügt nämlich ein einziger falscher Buchstabe, damit die ganze Antwort vom Computer als „falsch" gewertet wird (vgl. S. 83 ebd.).

Die Gültigkeit einer Messung – die sog. Validität – baut auf ihrer Meßsicherheit (Reliabilität) auf.

Eine reliable Variable, also eine mit großer Meßsicherheit, mißt sicher und stabil, was sie messen soll, und nicht einmal dies und das andere Mal jenes. Eine valide Variable erfaßt darüberhinaus auch wirklich das Phänomen, das sie theoretisch repräsentieren soll.

Ein Beispiel macht den Unterschied der beiden Konzepte deutlich: Der Lerneffekt einer Unterrichtseinheit soll durch einen Test gemessen werden. Es werden mehrere Aufgaben zusammengestellt, die den Stoff der Einheit abfragen sollen. Der Test ist dann reliabel, wenn – unter anderem – sichergestellt ist, daß die einzelnen Aufgaben weder zu schwer noch zu leicht sind; denn sonst würden entweder (fast) alle Schüler oder (fast) niemand den Test „bestehen", und er würde nichts mehr messen (Plafondeffekt).

Außerdem sollte der Test und jede seiner Aufgaben gute und schlechte Schüler trennen (hohe Trennschärfe). Und schließlich sollten sich bei einer Wiederholung des Tests bei den gleichen (oder vergleichbaren) Schülern ohne zwischenzeitliche Lernaktivität auch die gleichen Resultate ergeben (Retest-Reliabilität).

Die Validität dieses Tests betrifft dann die Frage, ob das, was hier gemessen wurde, tatsächlich auch der reine Lerneffekt einer Unterrichtseinheit war oder irgend etwas anderes, was z. B. parallel dazu auftrat. Es müßte überprüft werden, ob die Lernziele der Unterrichtseinheit tatsächlich durch die Testaufgaben abgedeckt sind (Lernzielvalidität). Es wäre aber auch notwendig zu untersuchen, ob der Unterricht wirklich den Lernzielen der Einheit entsprochen hat; sonst würde der Test etwas zu messen versuchen, was gar nicht vorhanden war. Und schließlich wäre die Frage zu klären, ob ein Wissenszuwachs wirklich durch die Unterrichtseinheit verursacht wurde, oder einfach durch natürliche Reifungsprozesse oder die Sensibilisierung für das Thema beim Ausfüllen der Testaufgaben.

Man sieht, die Frage der Validität betrifft letztlich auch die Qualität des Untersuchungsdesigns insgesamt, also z. B. die Notwendigkeit von Kontrollgruppen, von Vor- und Nachtests sowie von Replikationen der Untersuchung mit anderen Meßinstrumenten aber gleichen Variablen.

Manfred Herbig hat die Probleme der Validität speziell für Tests im (Schul-)Unterricht untersucht (vgl. *Herbig* 1975). Zur Validitätsproblematik bei Einstellungsskalen liegt eine zusammenfassende Arbeit von *Hanns Heinrich* vor (*Heinrich* 1974, siehe auch: *Thompson* 1979/80).

Bei keiner der von uns durchgesehenen empirischen Arbeiten zur Geographiedidaktik wurden Probleme der Validität näher untersucht. Eine Ausnahme ist allenfalls die methodische Arbeit von *H. Köck* über die „Problematik von Auswahlantworten in lernzielorientierten erdkundlichen Klassenarbeiten", wobei es allerdings eher um Fragen der Reliabilität als um solche der Validität geht (*Köck*, 1977).

(5) *Unvollständige Dokumentation der eingesetzten statistischen Analyseverfahren und fehlende Aufbereitung der Resultate*

Der Erörterung dieses Problems sei folgende Tabelle voranstellt. Sie stammt aus einem Aufsatz von *Leusmann* (*Leusmann*, 1978).

Dimension	Erklärende Variable	stand. Koeffizient	Fehler	T-Test	Signifikanz	Varianz
1. Allgemeine Bewertung	1. Dummy 2	-.388	.177	-4.80***	.001	.101
	2. Alter	-.155	.065	-1.57++	.119	.011
	3. Dummy 1	-.277	.163	-.329++	.002	.048
	4. pers. Beschäftigung	.143	.060	2.13+	.035	.020
	5. Dummy 4	-.136	.145	-1.90+	.060	.016
	6. Dummy 3	-.199	.175	-2.33+	.021	.024
	7. Benotung	.132	.062	1.88	.062	.016
	Konstante	mult.Korr. R^2		F-Wert	DP	P
	0.00	.433	.188	6.10	185	.001
2. Anforderung	1. pers. Beschäftigung	-.176	.051	-2.46*	.015	.030
	2. Dummy 5	.142	.112	1.94	.055	.019
	3. Alter	.081	.051	1.13	.262	.006
	0.00	.233	.054	3.63	189	.015
3. Allgemeine Fachstruktur	1. Medieneinsatz	-.238	.065	-3.31***	.002	.053
	2. Benotung	-.160	.051	-2.27+	.025	.025
	3. Dummy 5	-.121	.110	-1.71	.089	.014
	4. Urlaubsverhalten	-.075	.052	-1.05	.297	.005
	0.00	.305	.093	4.83	188	.001
4. Erklärungsvalenz	1. Erklärungsvalenz	.389	.069	5.45***	.001	.148
	2. Stoffanord-	-.135	.101	-1.90	.060	.018
	3. Dummy 4	-.119	.143	-1.67	.098	.014
	4. pers. Beschäftigung	.117	.063	1.64	.104	.013
	5. Dummy 3	.101	.145	1.42	.158	.010
	0.00	.438	.192	7.69	162	.001
5. Verarbeitungsvalenz	1. Dummy 4	-.202	.135	-2.70**	.008	.040
	2. Verarbeitungsvalenz	-.160	.074	-2.15*	.033	.026
	3. Dummy 3	.179	.139	2.37*	.019	.031
	4. Urlaubsverhalten	.107	.060	1.41	.160	.011
	0.00	.321	.103	4.69	163	.002

Tab. 3: Ergebnisse der multiplen Regression (nach *Leusmann*, 1978)

Die Tabelle zeigt die Ergebnisse von insgesamt 5 verschiedenen multiplen Regressionen. Doch was hier dabei im einzelnen gerechnet wurde, ist nicht zu erkennen. Über seine Methode schreibt *Leusmann* nicht viel mehr, als daß es sich um eine schrittweise multiple Regression handelt, bei der eine 1 %-ige Varianzaufklärung als Abbruchkriterium diente (vgl. seine Anmerkung 8, S. 136 ebd.).

Mindestens folgende Punkte bleiben unklar:

- Warum werden in die Regression Variablen aufgenommen, die keinen signifikanten Beitrag zur Varianzaufklärung leisten? In der Dimension 1 (Allgemeine Bewertung) ist die Variable 2 „Alter" in der Regressionsgleichung – obwohl ihr t-Wert mit P = 0,119 eindeutig insignifikant ist. Das gleiche gilt für die Variable 3 (Alter) bei der Regression zu der Dimension 3 (Anforderung), sowie für die Variable 4 bei der Regression zur Dimension 3 (Allgemeine Fachstruktur). Bei der Regression zur Dimension 4 (Erklärungsvalenz) sind gleich zwei Variablen insignifikant, nämlich Variable 4 und 5 (Dummy 3). Und auch bei der letzten Regressionsrechnung zur Dimension 5 (Verarbeitungsvalenz) ist die Variable 4 in ihren spezifischen Beitrag zur Varianzaufklärung insignifikant. „Insignifikanz" müßte eindeutig über das Abbruchkriterium von 1 %-iger Varianzaufklärung dominieren.
- Bei einer schrittweisen Regression (Methode „foreward") werden normalerweise die Variablen nach ihrer Wichtigkeit, d. h. Varianzaufklärung, in die Gleichung einbezogen, also zuerst jene Variable, die die abhängige Variable am besten aufklärt, dann die mit der zweitbesten Aufklärung, usw.
  Aus *Leusmanns* Tabelle geht diese sukzessive Variableneinbeziehung nicht hervor (vergl. die Variablen 2 und 3 bei der Dimension 1).
- Die Spalte „Varianz" ist unklar. Handelt es sich dabei um die spezifischen, zusätzlichen Varianzbeiträge der schrittweise einbezogenen Variablen *nach* Beendigung der gesamten Regressionsrechnung? Oder handelt es sich dabei um die einzelnen Schritte jeweils für sich, d. h. *vor* Berechnung der Gesamtgleichung. Im ersten Fall müßten sich die Varianzbeiträge zur gesamten Varianzaufklärung summieren – also mit der quadierten Multiplen Korrelation R^2 (Bestimmtheitsmaß) übereinstimmen. Im zweiten Fall würde es sich gewissermaßen um die Zwischenstufen der Regressionsrechnung handeln, und die Varianzaufklärung würde immer nur für die bis dahin eingeschlossenen Variablen gelten. Diese „Zwischenergebnisse", für die noch nicht alle Variablen enthaltende Regressionsgleichung, stimmen normalerweise nicht (!) mit den endgültigen Varianzbeiträgen überein, die sich aus der Gleichung mit allen Variablen ergeben.

Wir möchten nicht behaupten, daß man an der Richtigkeit der Regressionsrechnung von *Leusmann* wegen der obigen Unklarheiten grundsätzlich zweifeln muß. Aber bei multiplen (oder multivariaten) Verfahren darf man eine bessere Dokumentation der statistischen Vorgehensweise erwarten. Gerade bei der Regressionsrechnung gibt es mehrere verschiedene Varianten mit zum Teil erheblich unterschiedlichen Ergebnissen. Die Aufteilung der erklärten Varianz auf die einzelnen Variablen z. B. kann sich je nach dem Modus der Variableneinbeziehung sehr stark unterscheiden (siehe *Nie* 1975, S. 337).

Abgesehen von der mangelhaften technischen Dokumentation des Rechenvorganges präsentiert *Leusmann* die Ergebnisse ohne jede inhaltliche Aufbereitung und mit nur spärlicher theoretischer Interpretation.

Nun liegt aber der Sinn einer Regressionsrechnung ja eigentlich nicht so sehr in der Jagd nach signifikanten Effekten von unabhängigen Variablen und in der Erzielung einer möglichst großen gesamten Varianzaufklärung. Vielmehr geht es um die inhaltliche Frage, „ob", „wie stark" und letztlich „warum" eine abhängige Variable durch bestimmte unabhängige Variablen positiv oder negativ beeinflußt wird. Bei einer Regression sollten also die Regressionskoeffizienten – um die es eigentlich geht – samt ihren Vorzeichen genauestens interpretiert werden. Was bedeutet „inhaltlich" z. B. der Koeffizient von 0,143 aus der Tabelle für die Variable „persönliche Beschäftigung" zur Schätzung der abhängigen Variablen „Allgemeine Bewertung" (vgl. Tab. 3 (1) oben)?

Kein Wort verliert *Leusmann* auch über die inhaltliche Bedeutung des Wertes 0,00 für das konstante Glied, der auffallenderweise in allen fünf Regressionen erscheint. Üblicherweise ist die Konstante nämlich nicht Null!

In diesem besonderen Fall ist der Wert aber tatsächlich richtig, und zwar deshalb, weil *Leusmanns* abhängige Variablen „Faktorenwerte" sind, wie er an anderer Stelle kurz erwähnt. Faktorenwerte sind im allgemeinen standardisiert, d. h. sie haben eine Standardabweichung von 1 und einen Mittelwert von 0. Für standardisierte abhängige Variablen muß das konstante Glied in der Regression 0 sein, bei nicht standardisierten Variablen ergibt sich dagegen normalerweise eine von Null abweichende Konstante.

Dies sind nur einige der methodischen Besonderheiten, über die man als Leser der Arbeit im Unklaren gelassen wird, und über die man gern genaueres erfahren würde. Es ist nicht einzusehen, warum man sich mit methodischer Detektivarbeit die Informationen erst mühsam zusammensuchen muß. Dies gilt nicht nur für die Arbeit von *Leusmann*. Auch bei anderen geographiedidaktischen Untersuchungen, besonders bei solchen, die multivariate statistische Verfahren benutzen, ist die Dokumentation unzureichend und die Interpretation besteht häufig nur in der Zusammenstellung der Computerausdrucke zu Tabellen. Offenbar übernehmen viele Forscher mit den statistischen Methoden den Argumentationsstil der Mathematiker und Statistiker – so, als ob es um die formale Ableitung logisch-mathematischer Zusammenhänge ginge. Der Irrtum dieser Methodenspezialisten liegt in der Fetischisierung der statistischen Verfahren auf Kosten der sozialwissenschaftlichen Substanz. Bei einer sozialwissenschaftlichen Analyse darf man nie aus den Augen verlieren, wozu die Verfahren eigentlich dienen: Ihr Ziel ist es, einen Beitrag zur Aufklärung spezieller substanzwissenschaftlicher Probleme zu leisten. Letztlich müssen deshalb die Analyseergebnisse in der Form *sozialwissenschaftlicher Argumente* vorliegen, und nicht als Zahlenkolonnen in einem Computerausdruck.

1.4 Zusammenfassung und Gesamtbewertung

Wir haben in diesem ersten Kapitel versucht, einen Überblick über den Stand der empirischen Forschung zur Didaktik der Geographie zu gewinnen. Ausgangspunkt dazu war die Klage von Fachdidaktikern über den methodischen Rückstand ihres Faches.

Die Zusammenstellung vorliegender empirischer Arbeiten erbrachte folgendes Ergebnis:

(1) Die Arbeiten konzentrieren sich im wesentlichen auf 5 inhaltliche Bereiche: Unterrichtsevaluation, Lernweganalysen, Untersuchungen zu psycho-sozialen Rahmenbedingungen des Unterrichtsgeschehens und empirische Analysen von methodischen Problemen.

(2) Untersuchungen über Schülereinstellungen können in der Bundesrepublik auf eine über 75-jährige Forschungstradition zurückgreifen. Bereits 1905 wurde eine erste empirische Erhebung auf diesem Gebiet unternommen. Es folgten ein gutes Dutzend ähnlicher Arbeiten innerhalb von knapp zwei Jahrzehnten. Diese frühen Arbeiten übertreffen mit ihrer Datenbasis (bis zu 6000 Fälle) oft bei weitem die Stichprobenqualität neuerer Untersuchungen. Auch sonst sind sie auf einem überraschend hohen methodischen Stand.

(3) In allen fünf Bereichen scheint es seit einiger Zeit eine methodische Schwerpunktverschiebung zu geben: Während frühere Arbeiten durch größeren Aufwand bei der Datengewinnung gekennzeichnet sind, tendieren neuere Arbeiten zu exzessivem Gebrauch von komplexen, multivariaten Analyseverfahren oder aufwendigen statistischen oder mathematischen Modellen (z. B. Graphentheorie).

Im nächsten Abschnitt haben wir dann jenen Teil der Untersuchungen genauer analysiert, der für die vorliegende Arbeit besonders interessant ist – nämlich empirische Arbeiten über psycho-soziale Rahmenbedingungen des Erdkundeunterrichts. Diese Arbeiten haben vor allem folgende methodische Mängel:

(1) Häufig fiel ein krasses Auseinanderklaffen von hochkomplexen statistischen Analyseverfahren und der oft völlig unzureichenden Datenbasis der Untersuchungen auf.

(2) Grundlegende (erklärende) Variablen wie Geschlecht, Schulart, Klassenstufe, Klassenzugehörigkeit, Lehrereinfluß, Schulortgröße, sozioökonomischer Status der Eltern usw. wurden kaum untersucht. Diese Beschneidung eines sinnvollen Variablensatzes resultiert nicht zuletzt aus einer einseitig psychologischen Problemperspektive und aus den unzureichend differenzierten Stichproben.

(3) Ein weiterer Mangel der vorliegenden Arbeiten war die einseitige Bevorzugung *metrischer* Analyseverfahren. Dementsprechend wurden vor allem Variablen auf Zusammenhänge hin untersucht, bei denen

man (wenn auch oft mit zweifelhafter Berechtigung) metrischer Meßniveau unterstellen kann. Traditionelle und seit Jahrzehnten bewährte *nicht*-metrische Analyseverfahren, wie die multidimensionale Kreuztabellenanalyse, wurde in keiner der Arbeiten eingesetzt.

(4) Die klassische Testtheorie ist an beinahe allen empirischen Arbeiten zur Geographiedidaktik spurlos vorübergegangen. Standardüberprüfungen der Reliabilität, Trennschärfe, Schwierigkeit oder Homogenität von Fragenbatterien sind kaum zu finden. Der Skalierung von Variablen und auch der Indexbildung wurde so gut wie keine Aufmerksamkeit geschenkt.

(5) Die Präsentation von Ergebnissen aus empirischen Untersuchungen war teilweise unvollständig, manchmal auch obskur. Ein unnötiges „Fachchinesisch" beherrschte etliche neuere Arbeiten. Besonders bei der Darstellung multivariater Analysen fehlten häufig Angaben, die zur Beurteilung ihrer Qualität unabdingbar sind.

2. Problemstellung der vorliegenden Arbeit

Die methodische Kritik an einigen vorliegenden empirischen Arbeiten zur Didaktik der Geographie könnte den Eindruck erwecken, es ginge uns lediglich darum, die Arbeit bestimmter Wissenschaftler auf methodische Mängel und Fehler hin zu sezieren und damit zu disqualifizieren.
Nichts liegt uns ferner!

Die zur Kritik herangezogenen Arbeiten dienten lediglich als Beispiel für methodische Fehlenwicklungen, die nach unserer Auffassung die empirische Forschung in den Sozialwissenschaften überhaupt kennzeichnen. Das Ziel dieser Arbeit ist es, diese allgemeinen methodischen Fehlentwicklungen herauszuarbeiten und an konkreten Beispielen zu demonstrieren, wodurch man ihnen entgegenwirken könnte. Dabei orientieren wir uns an einem Verständnis von empirischer Forschung, das auf wissenschaftstheoretischen Überlegungen von *Karl Popper* beruht. Besonders dessen „Drei-Welten-Theorie" war Hintergrund unserer Überlegungen zum empirischen Forschungsprozeß (*Popper/Eccles*, 1977).

2.1 Die Komponenten eines empirischen Forschungsprozesses

Empirische Sozialforschung – und damit auch empirische Forschung zur Didaktik der Geographie – kann man am besten durch das Modell eines „informationsverarbeitenden Systems" veranschaulichen. Dieses informationsverarbeitende System hat mehrere Komponenten (siehe Abb. 1 auf der nächsten Seite):

(1) *Die „eigentliche" soziale Wirklichkeit*: (Wissenschaftstheoretiker nennen sie auch „Realität nullter Ordnung")

Gemeint ist damit die Fülle und Komplexität der sozialen Phänomene, die in ihrer Totalität weder dem beobachtenden Wissenschaftler, noch den in ihnen lebenden Menschen zugänglich sind. Gegner der empirischen Forschung meinen häufig diese soziale „Realität nullter Ordnung", wenn sie beklagen, daß die Empirie nicht in der Lage sei, das „Eigentliche", das „Wesen" der sozialen Phänomene zu erfassen. Wir wollen uns hier nicht in wissenschaftstheoretische Erörterungen einlassen. Deshalb sei dazu nur folgendes angemerkt: Die „eigentliche" soziale Wirklichkeit ist lediglich ein Konzept zur Markierung eines Grenz-Phänomens, also nichts anderes als ein theoretisches Hilfskonstrukt, das man benötigt, um die eigentlich interessanten „Realitäten" höherer Ordnung davon abgrenzen zu können. Diese höheren Realitäten sind z. B. die Alltagsrealität, die wissenschaftliche Realität oder auch spirituelle und religiöse Realitätserfahrungen.

(2) *Die „empirisch erfaßbare Realität"*

Den Wissenschaftler interessiert nur jener Ausschnitt aus der „eigentlichen" sozialen Wirklichkeit, über den er mit wissenschaftlich anerkannten, d. h. intersubjektiv geltenden Methoden und Techniken Informationen gewinnen kann. Dies ist seine empirische Basis; wenn man so will, eine wissenschaftliche „Realität erster Ordnung". Die intersubjektiven Methoden des Zugangs unterscheidet die empirische Wirklichkeit der Wissenschaftler zum Beispiel von der des ontologisch denkenden Philosophen, der das „wahre Wesen" der Dinge mit Mitteln erforscht, die allein seinem eigenen symbolischen Universum angehören. Dies unterscheidet Wissenschaftler aber auch von Alltagsmenschen, deren Mittel der Informationsgewinnung und Verarbeitung nicht universell und intersubjektiv sind, sondern subkulturell kanalisiert und subjektiv geprägt werden.

(3) *Die „gemessene" soziale Wirklichkeit*

Die gemessene soziale Wirklichkeit besteht aus jenen Informationen, die der Forscher bei einer bestimmten Untersuchung mit Hilfe seiner Techniken und Methoden aus der überhaupt wissenschaftlich erfaßbaren Wirklichkeit herausfiltert. Sie unterscheidet sich von der prinzipiell wissenschaftlich erfaßbaren Realität durch drei Elemente: nämlich durch eine Informationsreduktion, durch eine Informationsselektion und durch eine Informationsverzerrung. Dies hat folgende Gründe:

Erstens kann man nicht alles gleichzeitig untersuchen. Jede empirische Erhebung hat eine bestimmte Problemstellung, eine theoretisch begründete *Perspektive.*

Zweitens benutzt man ganz bestimmte *Meßinstrumente,* seien dies Fragebögen oder geschulte Beobachter, Intelligenztests oder Computerprogramme, die die Häufigkeiten bestimmter Ereignisse registrieren. Man kann nicht alle Meßinstrumente gleichzeitig benutzen, wenn man nicht die gemessene soziale Wirklichkeit dabei beeinträchtigen will.

Drittens tritt immer Informationsverlust bzw. Falschinformation durch *Meßfehler* im weitesten Sinn auf: Der Beobachter kann kurz einnicken, der Fragebogen kann Suggestivfragen enthalten, Schüler können sich bei einem Intelligenztest absichtlich dumm stellen.

Und viertens muß die Information bei jedem denkbaren Meßinstrument von einem Informationsträger in einen anderen überführt werden. Die dabei möglichen Fehler durch Informationsverlust haben wir bereits erwähnt. Hinzu kommt jedoch die Möglichkeit einer *systematischen Verzerrung.* Wenn man beispielsweise eine „Einstellung" in eine „Ratioskala" überträgt, fügt man der tatsächlich vorhandenen Information über die Einstellung höchstwahrscheinlich zusätzliche, fremde Informationen hinzu, die nur in der Konstruktion des Meßinstruments steckt.

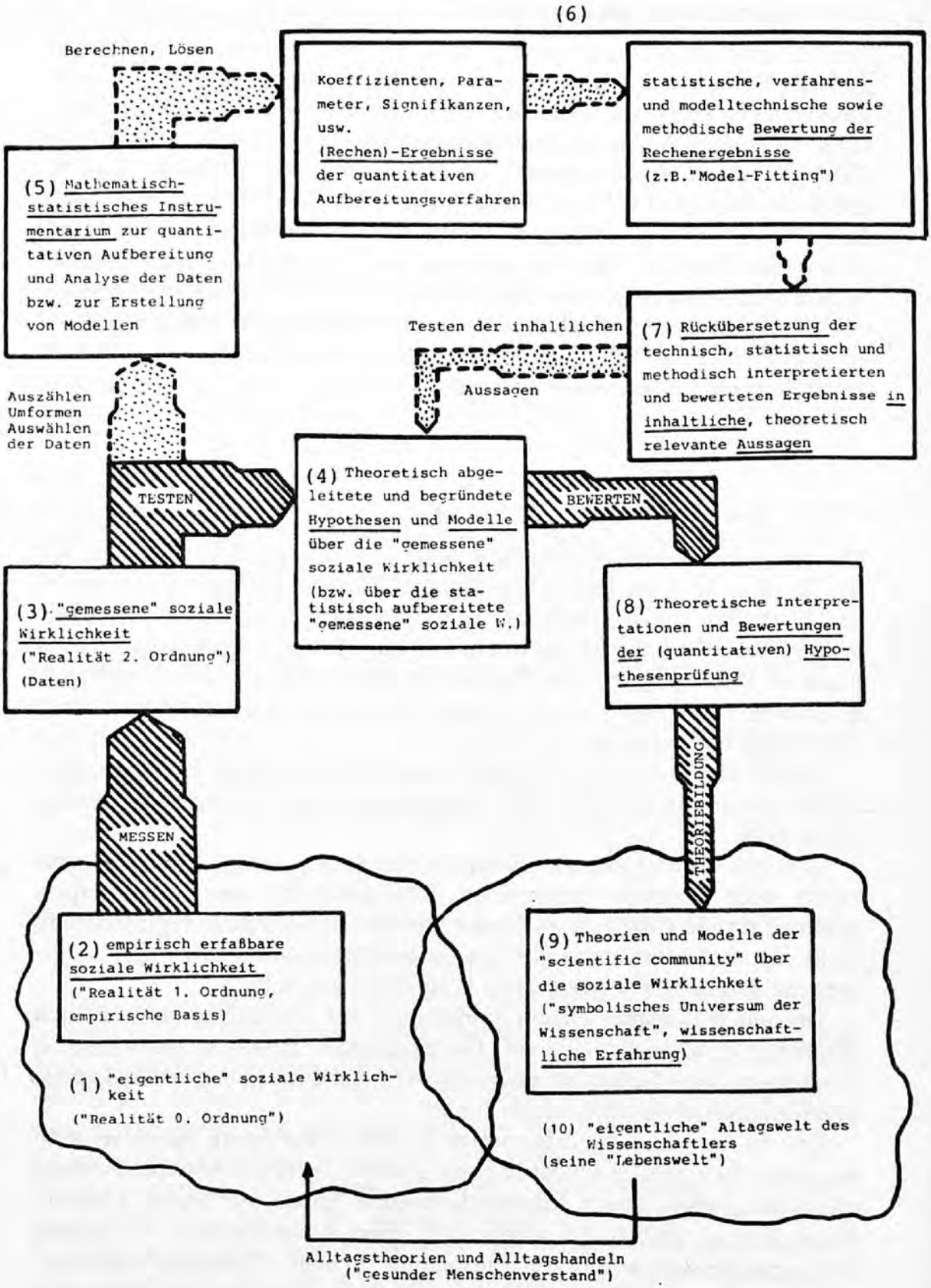


Abb. 1: Die Komponenten eines empirischen Forschungsprozesses

(4) Die *theoretisch begründeten Hypothesen und Modelle* über diese „gemessene" soziale Wirklichkeit

Oft wird beklagt, daß sich die Empiriker nicht mit den eigentlich relvanten Problemen befassen. Lassen wir einmal beiseite, welche Probleme wissenschaftlich relevant sind, eines ist sicher richtig: der empirische Forschungsprozess kann sich nur mit Problemen befassen, zu denen „Messungen" – gleich welcher Art – möglich sind. Es gibt sehr interessante Hypothesen, die man als wissenschaftlich unbrauchbar bezeichnen muß, weil man bisher keine Mittel hat, um Informationen darüber zu gewinnen, ob sie richtig oder falsch sind. Damit ist aber nichts über die sonstige Bedeutung dieser Probleme gesagt. Auf viele kluge Fragen hat die Wissenschaft einfach keine Antwort und wird vielleicht auch niemals eine Antwort versuchen, einfach deshalb, weil über die Richtigkeit dieser Antworten nicht entschieden werden kann.

Leider beantworten einige Sozialwissenschaftler Fragen (d. h. Hypothesen) dadurch, daß sie viele Worte darüber verlieren, warum eine bestimmte Sache theoretisch so sein muß und nicht anders: Sie entwerfen *verbale* Modelle der sozialen Realität, die sie aus ihren Theorien ableiten. Nur geben sie keine Ratschläge, wie man zur Überprüfung dieser verbalen Modelle intersubjektiv akzeptable empirische Informationen gewinnen kann.

Hypothesen und Modelle ergeben sich *nicht* aus den Daten, sondern aus der theoretischen Erfahrung der Wissenschaftler. Anhand von Hypothesen oder theoretischen Modellen werden die vorliegenden Daten gesichtet, sortiert und aufbereitet. Erst dann kann man mit Hilfe der Daten die Hypothesen testen und die Modelle auf Angemessenheit hin überprüfen.

Daten allein beweisen gar nichts. Erst wenn ein Forscher
- eine bestimmte Problemstellung verfolgt,
- sie in ein angemessenes Untersuchungsdesign umsetzt,
- mit reliablen und validen Meßinstrumenten die notwendigen Daten für die Variablen des Designs erhebt
- und diese für bestimmte Hypothesen oder Modelle aufbereitet,

gewinnen Daten eine Beweiskraft. Hypothesen und Modelle sind gewissermaßen die Instrumente, welche die Daten zum Sprechen bringen. Diese können jedoch immer nur auf jene Fragen antworten, die ihnen durch die Hypothesen gestellt werden. Ein und dieselben Daten können ganz verschiedene „Antworten" geben, je nachdem welche „Fragen" an sie mittels der Hypothesen und der Modelle gestellt werden. Eine empirische Forschung ohne Theorie gibt es nicht, weil empirische Daten ihre Bedeutung erst durch theoretische Konstrukte erhalten.

(5) Die *Aufbereitung der Daten* für die Hypothesen und Modelle

Die Datenaufbereitung kann verschiedenste Techniken umfassen; angefangen von einer einfachen Häufigkeitsauszählung bis hin zu einer multiplen Regression oder einem Log-linearen Modell.

Manche Hypothesen kann man gewissermaßen fast unmittelbar mit den vorliegenden Daten testen. Bei anderen Hypothesen und besonders bei komplizierten theoretischen Modellen müssen die Daten durch quantitative Verfahren erst bearbeitet werden. Bei jeder Datenaufbereitung kommt es zu einer weiteren Informationsreduktion; auch bei einer schlichten Kreuztabelle, denn es ist gerade der Sinn der Datenaufbereitung, nur jene Informationen aus den Daten zu entnehmen, die für die Problemstellung relevant sind.

Sobald kompliziertere quantitative Verfahren der Datenaufbereitung notwendig werden, um die Problemstellung der Untersuchung zu beantworten, genügt es nicht, wenn man die Ergebnisse der Analysen einfach mit den theoretischen Hypothesen konfrontiert, um diese beurteilen zu können. Bei mulitplen und multivariaten statistischen Verfahren liegen die Ergebnisse nämlich in Form von Korrelationskoeffizienten, Regressionsgewichten, Koordinaten, Ladungen, Eigenwerten oder ähnlichem vor. Diese Zahlenwerte enthalten die in den Daten ursprünglich vorhandene relevante Information (oder Teile davon) in extrem verdichteter Form. Deshalb müssen die Ergebnisse dieser Rechenverfahren erst interpretiert werden.

(6) Die *(Rechen-)Ergebnisse der* verschiedenen (statistischen) Aufbereitungsverfahren

Die Logik der multivariaten und multiplen statistischen Datenaufbereitungsverfahren erfordert es, ihre Ergebnisse einer *internen technischen Bewertung und Interpretation* zu unterwerfen, ehe man mit ihnen die eigentliche theoretische Fragestellung der Untersuchung beantworten kann. Diese *technische* Interpretation der Ergebnisse ist je nach Verfahren sehr unterschiedlich. Bei einer Faktorenanalyse muß beispeilsweise geklärt werden, wieviele Faktoren gezogen werden sollen. Es muß untersucht werden, ob die Kommunalitäten der einbezogenen Items angemessen sind, oder ob sie nahelegen, das eine oder andere Item auszuschließen. Schließlich ist zu entscheiden, nach welchem Rotationsverfahren die Ladungsmatrix rotiert wird, und ob das Ladungsmuster eine Einfachstruktur der Faktoren ergibt oder nicht. Bei komplizierten Faktorenanalysen muß außerdem festgelegt werden, ob und wie die Anfangskommunalitäten geschätzt werden.

Wir können hier nur einige Aspekte dieses Problems andeuten. Es sollte aber deutlich geworden sein, daß statistische „Aufbereitungsverfahren" niemals einen quasi automatischen Ablauf haben. Es gibt nicht „die" Faktorenanalyse, oder „die" multiple Regression, sondern jeweils mehrere Dutzend verschiedene Varianten. Welche Varianten man wählt, und wie man die Ergebnisse bewertet, hängt ab von der jeweiligen Problemstellung, von den Daten – und nicht zuletzt von der eigenen Kenntnis der Verfahren.

Es fällt vielleicht auf, daß wir oben von statistischen „Aufbereitungsverfahren" sprachen. Damit sollte auf den *pragmatischen* Aspekt aller sta-

tistischen Verfahren und Modelle aufmerksam gemacht werden. Es besteht nämlich bisweilen die Tendenz, solche Verfahren in ihrer Bedeutung zu überschätzen und ihnen gewissermaßen ein Eigenleben zuzuschreiben. Gewisse mathematische Gleichungssysteme werden dann als (mathematische) Modelle der sozialen Wirklichkeit betrachtet, wobei die *mathematischen Eigenschaften* dieser Gleichungssysteme stellvertretend für soziale Zusammenhänge analysiert werden. Diese Auffassung von quantitativer Analyse teilen wir nicht. Wir glauben vielmehr, daß es letztlich allein auf die Information ankommt, die aus der „empirischen" Wirklichkeit (Realität erster Ordnung) extrahiert wurde, und nicht auf die formal-logischen Eigenschaften der mathematischen oder statistischen Modelle.

(7) *Die Umsetzung der* (technisch interpretierten und bewährten) *Rechenergebnisse in inhaltliche Aussagen.*

Es ist wohl inzwischen allgemein bekannt, daß beispielsweise ein höchst signifikanter Korrelationskoeffizient zwischen zwei Variablen nicht mehr ist, als eine schlichte Rechengröße. Erst die inhaltliche Interpretation dieses Rechenergebnisses kann daraus z. B. einen (Kausal-)Zusammenhang ableiten.

Diese Umsetzung der statistischen Rechenergebnisse in inhaltliche Aussagen ist das Gegenstück zum Meßvorgang am Beginn der empirischen Untersuchung. Genauso, wie dort – durch den Vorgang der Messung von Variablen – die sozialen Phänomene in ein Zahlensystem übersetzt werden, müssen am Ende der quantitativen Analyse die Rechenergebnisse wieder zurückübersetzt werden in inhaltliche Beschreibungen einer sozialen Wirklichkeit. Der ganze Vorgang der quantitativen Analyse darf niemals Selbstzweck werden. Er dient lediglich der Umformung, Strukturierung und Verdichtung von Informationen über die soziale Wirklichkeit mit dem Zweck, der so gewonnenen Beschreibung eine größere analytische Tiefe zu geben.

Die Ablehnung quantitativer Forschung, wie sie nicht selten zu beobachten ist, hat eine Ursache darin, daß häufig nur Rechenergebnisse publiziert werden, statt allgemeinverständlicher, inhaltlicher Beschreibungen der untersuchten Realität.

2.2 Drei Thesen über methodische Fehlentwicklungen

Anhand der im vorigen Abschnitt 2.1 skizzierten Komponenten der empirischen Forschung wollen wir nun drei Thesen entwicklen, aus denen sich dann die Ziele der vorliegenden Arbeit ergeben:

(1) These 1: Die Integration der einzelnen Komponenten einer empirischen Untersuchung ist das Hauptproblem uns vorliegender empirischer Arbeiten.

Die Qualität einer empirischen Arbeit ergibt sich nicht daraus, daß einzelne Komponenten des Forschungsprozesses auf methodisch hohem Niveau durchgeführt werden. Qualität entsteht vielmehr, wenn alle Schritte der Untersuchung in einem ausgewogenen Verhältnis zueinander stehen. Auch hier gilt der Satz: „Jede Kette ist nur so stark, wie ihr schwächstes Glied".

- So ist beispielweise das *Design* einer Untersuchung von der theoretischen Problemstellung abhängig. Es gibt kein Design, das sich für alle Fragestellungen eignet. Durch das Untersuchungsdesign wird festgelegt, welche Variablen die Untersuchungsvariablen sind, welche Variablen extern bleiben aber kontrolliert werden sollen, und welche Variablen extern bleiben aber als zufällige Fehler zu betrachten sind („randomized errors") (vgl. *Kish* 1970). Vergißt man, für die Problemstellung wichtige Variablen zu kontrollieren oder als Untersuchungsvariablen einzuführen, kann dies nachträglich nicht mehr gut gemacht werden.
- Das Design bestimmt Art und Umfang der *Datenerhebung*: Experimentelle Designs mit wenigen kontrollierten Variablen benötigen relativ wenige Fälle, aber große Sorgfalt bei der Randomisierung der Gruppen. Surveyähnliche Erhebungen benötigen im allgemeinen viele Fälle; nicht nur zur Realisierung einer repräsentativen Stichprobenauswahl, sondern auch damit später bei der Analyse nach wichtigen Strukturvariablen gruppiert werden kann, ohne daß die Gruppen zu klein werden.
- Die Datenerhebung muß auf die verwendeten *Meßinstrumente* abgestimmt sein. Es ist z. B. unmöglich, eine repräsentative Stichprobe von Schülern mit einem dreistündigen IQ-Test verbinden zu wollen.
  Welche Meßinstrumente verwendet werden, hängt natürlich auch ab von den Variablen des Untersuchungsdesigns und von der Problemperspektive (z. B. eher psychologisch oder eher soziologisch). Die Auswahl der Meßinstrumente hat seinerseits erhebliche Folgen für die statistischen Rechenverfahren, die später eingesetzt werden sollen. Nicht nur das Datenniveau spielt dabei eine Rolle. Will man später für eine Item-Batterie eine Faktorenanalyse rechnen, wäre es z. B. unsinnig, möglichst heterogene Items zu formulieren, da diese nur schlecht zu einzelnen Dimensionen zusammengefaßt werden können.
- Die *Hypothesenbildung* basiert u. a. auf dem Untersuchungsdesign, auf den zur Verfügung stehenden statistischen Analyseverfahren, auf der Art und Anzahl gleichzeitig betrachteter Variablen und auf dem Umfang und der Qualität der Stichprobe.
- Die *Aufbereitung der Daten* zur Hypothesenprüfung hängt oft von so profanen Dingen wie einer ordentlichen Kodierung oder korrekt ausgefüllten Fragebögen ab. Es gibt Untersuchungen, die allein wegen zu

vieler „missing values" unbrauchbar sind. Dieses Problem wird häufig unterschätzt. Es ist deshalb so gravierend, weil sich bei multiplen oder multivariaten Verfahren die „missing values" der einzelnen Variablen im allgemeinen überschneiden, so daß schließlich nur ganz wenige Fälle übrigbleiben, die Daten auf allen (oder zumindest vielen) Variablen aufweisen.

- Die Durchführung komplizierterer statistischer *Aufbereitungsverfahren* setzt voraus, daß die Erhebung in der Fallzahl und der Struktur der Stichprobe gewissen Anforderungen genügt. Man kann nicht jedem Datensatz ein multivariates Analyseverfahren überstülpen. Ganz abgesehen davon, muß das Untersuchungsdesign in seiner Problemstellung bereits auf die Verwendung multivariater Verfahren ausgerichtet sein: Will man z. B. eine bestimmte abhängige Variable durch zehn Einstellungsvariablen „erklären", und hat man vor, eine multiple Regression mit diesen Einstellungsvariablen zu rechnen, dann kann es wegen möglicher Multikolinearität zu unsinnigen Ergebnissen kommen, wenn zu ähnliche Einstellungsvariablen ausgewählt wurden – auch wenn sonst alles einwandfrei durchgeführt wurde.
- Die *inhaltliche Füllung der Resultate* bei komplizierten statistischen Auswertungen ist Anlaß unzähliger Mißverständnisse. Die Jagd auf „Signifikanz" ist dafür ein Beispiel. Die Signifikanz hängt ab von der Stichprobengröße. Man kann sie bei sonst gleichen Verhältnissen, durch Vergrößerung der Stichproben rein technisch erhöhen. Wie die Qualität der Rechenergebnisse zu interpretieren ist, ergibt sich primär aus der Problemstellung der Untersuchung und daraus, welches Vertrauen man in seine Meßinstrumente haben darf, wie man also die Reliabilität und Validität der Daten einschätzt.

Dies ist nur eine kleine Auswahl der vielfältigen Verflechtungen zwischen den einzelnen Teilen einer empirischen Untersuchung. Sie sollen belegen, daß es unmöglich ist, eine Komponente für sich allein zu optimieren. Der *häufigste* Verstoß gegen eine ausgewogene empirische Untersuchung ist die Diskrepanz zwischen der Stichprobe und den sonstigen Komponenten der Untersuchung. Während die Variablen mit größtem Aufwand operationaliert werden (vgl. *Bachmair* 1969) und während die kompliziertesten statistischen Analyseverfahren Verwendung finden (vgl. *Leusmann* 1977), ist die Datenbasis häufig wegen einer zu kleinen und schlecht strukturierten Stichprobe praktisch auf vorwissenschaftlichem Niveau. Weitere Verstöße sind das Auseinanderklaffen zwischen Dokumentation und Auswertungsverfahren und das unterschiedliche Niveau bei der Variablenmessung und der Analyse.

(2) These 2: Es besteht die Gefahr, daß statistische Analyseverfahren und (mathematische) Modelle zum Selbstzweck werden.

Zugegebenermaßen existiert diese Gefahr in der empirischen Forschung zur Didaktik der Geographie erst in Ansätzen. Gewisse Tendenzen zeigen sich jedoch deutlich in den Arbeiten von *Chr. Leusmann.*

Man kann gegen eine Überbewertung der Modelle und Analyseverfahren auf sehr verschiedenen Ebenen argumentieren: Wir wollen ein historisches Argument vorlegen.

Im Jahre 1904 begann die Entwicklung der Faktorenanalyse mit dem berühmten Aufsatz „General Intelligence – Objectively Determined and Measured" von *Charles Spearman* (*Spearman* 1904). Sein Verdienst war es, eine psychologische Theorie der Intelligenz zu entwickeln, die sich nahtlos mit der Analyse von Korrelationsmatrizen verbinden ließ. Er ging davon aus, daß Intelligenz ein genereller „Faktor" sei, der hinter den einzelnen spezifischen Intelligenzleistungen (= spezifische Faktoren) steckt. Diese inhaltliche Theorie ließ sich gut mit den formal-statistischen Überlegungen von *Karl Pearson* über die Hauptachsen („principal axes") bei Matrizen aus Korrelationskoeffinzienten verbinden (*Pearson* 1901). Man nahm an, daß die erste Hauptachse (heute würde man sagen: die Hauptdimension) einer Korrelationsmatrix von Intelligenzaufgaben die „generelle Intelligenz" repräsentiert. In der Folgezeit gab es buchstäblich eine Revolution in der psychologischen Intelligenzdiagnostik, die eine unmittelbare Folge der Faktorenanalyse war.

Nach einigen Jahrzehnten intensiver Arbeit mit den verschiedenen Varianten der Faktorenanalyse bei der Erstellung und Überprüfung immer umfangreicherer Testbatterien passierte folgendes: Das statistische Verfahren wurde teilweise zum Selbstzweck. Immer mehr Testpsychologen glaubten damals, die Logik des Verfahren würde die grundlegenden Strukturen und Gesetze der Intelligenz und der Persönlichkeit abbilden. Die Theoriebildung nahm ihre Inspiration nicht mehr aus den Informationen, die sie aus der empirischen Wirklichkeit gewann, sondern aus den formalen Eigenschaften des statistischen Verfahrens. Diese Eigenschaften der verschiedenen Rechenvarianten wurden mit psychologischen Deutungen versehen: So gab es lange Auseinandersetzungen darüber, ob Intelligenz ein „multiple factor"- oder ein „single factor"-Phänomen sei. Dies war ein direktes Ergebnis zweier verschiedener Lösungsansätze für das Hauptachsenproblem der Faktorenanalyse. *H. Harman* schrieb über diese Zeit: „. . . factor analysis was perceived as a kind of mystique among many psychologist." (*Harman* 1976, S. 5). Der Irrtum dieser Forscher damals war, das „Werkzeug" mit dem zu bearbeitenden „Gegenstand" zu verwechseln. Man verlor damals aus den Augen, daß es letztlich um die Information aus der „Realität erster Ordnung" (d. h. um Informationen aus der empirischen Wirklichkeit) geht und nicht um Eigenschaften eines formalen mathematischen Kalküls. *Harman* beschreibt dieses Problem in seinem modernen Lehr-

buch der Faktorenanalyse so: „The methodology of this text – „exploratory" factor analysis – may be useful in formulating theories in the behavioral and social sciences, but the „analytical tools" (including factor analysis) should not be confused with the „science". As an exploratory tool (among others), factor analysis can be used to verify or modify theories through new experiments and new data subjected to fresh analyses for purposes of clarifying or polishing previous formulations. By contrast „confirmatory" factor analysis may be used to check or test a preconceived or given hypothesis about the structure of empirical data." (*Harman* 1976, S. 5).

Heute wird selbstverständlich niemand mehr auf die Idee kommen, in eine Faktorenanalyse mehr hineinzudeuten, als in ihr steckt. Dennoch besteht auch heute noch die Gefahr, daß statistische Analyseverfahren und mathematische Modelle zum Selbstzweck werden. In der Soziologie gibt es bereits bestimmte Strömungen, die man als „Methodischen Essentialismus" bezeichnen könnte. Vertreter dieser Richtung haben den empirischen Forschungsprozeß praktisch auf eine *einzige* Komponente reduziert: Sie interessieren sich nur noch für die formalen Eigenschaften von mathematischen Modellen der Wirklichkeit. Beispielhaft dafür sind die Arbeiten zur mathematischen „Katastrophentheorie", die ausdrücklich auch soziale Phänomene beschreiben wollen (vgl. *Waddington* 1977, *Jiobu/Lundgren* 1978).

(3) These 3: Bei empirischen Untersuchungen sollten verstärkt die Möglichkeiten der nicht-metrischen Zusammenhangsanalyse ausgeschöpft werden.

Wenn man sich mit dem empirischen Forschungsprozeß in einer ganzheitlichen Sicht auseinandersetzt, so wie wir dies in dieser Arbeit versuchen, dann fällt ein Mangel metrischer Analyseverfahren sofort ins Auge: Diese Verfahren verlagern einen Großteil aller Probleme in eine einzige Komponente des Forschungsprozesses, nämlich in die Datengewinnung. Wie unter einem Brennglas konzentrieren sich dabei die Schwierigkeiten auf den Meßvorgang, d. h. auf die Übertragung von Information aus der empirischen Realität („Realität 1. Ordnung") in eine numerische Realität (oder in ein numerisches „Relativ" – wie man auch sagt). Das „empirische Relativ" muß dabei auf das „numerische Relativ" so abgebildet werden, daß jedem empirischen Element eine Zahl und jeder Beziehung zwischen empirischen Elementen eine Beziehung zwischen den jeweiligen Zahlen entspricht (vgl. *Kriz* 1981, S. 36). Der kritische Punkt dabei ist, daß beim *metrischen* Meßvorgang die Relationen zwischen den Zahlen sehr restriktiv vorgegeben sind. Die empirische Information muß sich beim metrischen Meßvorgang in das System der natürlichen Zahlen übertragen lassen. Rechenoperationen, die im natürlichen Zahlensystem mathematisch zulässig sind (wie Addition

und Mulitplikation), müssen in der empirischen Realität eine Entsprechung haben.

Das dies nicht so ohne weiteres gegeben ist, dürfte ohne weiteres einleuchten. Eine naive „soziale Physik" ist deshalb von Anfang an zum Scheitern verurteilt. Aus diesem Grund ist es kein Wunder, daß die Übertragung von Informationen aus der empirischen Wirklichkeit in ein metrisches „numerisches Relativ" das zentrale Forschungsproblem in der Methodenentwicklung der Sozialwissenschaften von Anbeginn an war. Es begann Ende des 19. Jahrhunderts, als die „Psychophysiker" die Beziehung zwischen physikalischen Intensitäten (z. B. Lautstärke gemessen in Dezibel) und psychischen Empfindungen (z. B. die empfunde Lautheit) zu ergründen versuchten (vergl. *Fechner* 1860). Seit den zwanziger Jahren unseres Jahrhunderts wurden eine Vielzahl von *Skalierungstechniken* entwickelt, die Ratio- oder wenigstens Intervallskalen produzieren sollten (wie z. B. die Paarvergleichstechnik, das Verfahren der gleichen Abstände, usw.).

Es besteht kein Zweifel darüber, daß diese verschiedenen Skalierungstechniken, bis hin zur modernen multidimensionalen Skalierung, ganz entscheidend für die methodische Weiterentwicklung der Sozialwissenschaften waren und noch sind.

Worauf wir aber aufmerksam machen wollen ist die Tatsache der einseitigen Schwerpunktverlagerung auf einen einzigen Aspekt im Forschungsprozeß, zu dem diese Verfahren geführt haben. Diese Verfahren implizieren nämlich größtenteils einen enormen technischen, zeitlichen, statistisch-mathematischen und theoretischen Aufwand.

Bereits beim Vorgang der Messung muß hier ein ganzes Arsenal von theoretisch-methodischen und mathematisch-statistischen Konzepten eingesetzt werden, um die erforderlichen Skaleneigenschaften zu realisieren. In der Forschungspraxis fehlt nicht nur häufig das notwendige Wissen für den Einsatz solcher Meßverfahren, sondern oft sind die Verfahren typische „Labormethoden", die für einen praktischen Einsatz zu aufwendig sind oder sich schlicht nicht durchführen lassen (wie z. B. die Technik des vollständigen Paarvergleichs bei größeren Erhebungen).

Wir haben bisher nur von der Meßproblematik gesprochen. Aber auch die Stichproben oder Versuchsanlagen müssen bei metrischen Daten oft extrem hohen Anforderungen genügen. Erinnert sei an das Problem der multivariaten Nominalverteilung, die bei allen metrischen Analyseverfahren für Variablenzusammenhänge als Voraussetzung erfüllt sein müßte.

Besteht man auf metrischen Daten für die eigentliche Analyse, dann wird der gesamte Forschungsprozeß zwangsläufig bereits bei seinem Beginn mit mathematisch-statistischen Konzepten überfrachtet und erleidet dadurch erhebliche Einschränkungen (in bezug auf den Gegenstand, die praktische Durchführung usw.).

In der Praxis sieht dieses Problem natürlich anders aus: Fast niemand kümmert sich wirklich um die verschiedenen Skalierungsverfahren und Meßmodelle. Wenn solche Verfahren benutzt werden, dann handelt es sich

bezeichnender Weise meist um rein methodische Arbeiten, die die Funktionsweise dieser Verfahren demonstrieren, ohne daß irgend etwas anderes dabei ernsthaft untersucht wird. Es ist verständlich, wenn sich kommerzielle Sozialforschungsinstitute mit dieser Meßproblematik nicht abgeben. Aber auch empirische Untersuchungen mit wissenschaftlichen Anspruch benutzen meistens nur die allereinfachsten Statementfragen, bei denen nach „multiple- choice-Methode" mehrere Antwortalternativen (nach Intensität abgestuft) vorgegeben sind.

Der Versuch, Daten auf metrischem Meßniveau zu erhalten, führt also entweder zu einer starken mathematisch-statistischen „Überfrachtung" der Datengewinnung durch sehr komplizierte Skalierungs- und Meßverfahren oder zu einer ständigen methodischen Unaufrichtigkeit, die darin besteht, daß einfach metrisches Datenniveau angenommen wird.

2.3 Die Konsequenzen aus unseren methodischen Vorüberlegungen für die vorliegende Arbeit

Ensprechend den oben entwickelten Thesen über methodische Fehlentwicklungen in der empirischen Sozialforschung (und auch in den von uns durchgesehenen empirischen Arbeiten zur Geographiedidaktik), ergeben sich für uns folgende Konsequenzen:

(1) Es soll versucht werden, *konkrete Verbesserungsvorschläge für möglichst viele Komponenten* des empirischen Forschungsprozesses zu entwickeln. Dabei sollten jene Komponenten besonders berücksichtigt werden, die in den vorliegenden geographiedidaktischen Untersuchungen vernachlässigt wurden. Diese methodischen Verbesserungsvorschläge müssen für die Forschungspraxis praktikabel sein, und sie sollten anhand empirischer Analysen geographiedidaktischer Problemstellungen dargestellt werden.

(2) Ein wichtiger Bereich, in dem methodische Verbesserungen nötig und möglich sind, betrifft die *Datenaufbereitung* und *Datenaggregation*. Die vorliegende Arbeit soll zeigen, wie aus Einzelerhebungen größere Datensätze erstellt werden können, und welche datentechnischen und methodischen Problem dabei zu beachten sind.

(3) Ein zweiter, bisher vernachlässigter Bereich ist die Reliabilität und Dimensionalität bei der Variablenmessung. Es soll aufgezeigt werden, daß die *Reliabilität und Dimensionalität* von Variablen auch nachträglich noch überprüft und verbessert werden kann – und zwar mit relativ einfachen Routineverfahren.

(4) Die statistische Analyse von Variablenzusammenhängen ist eine weitere Komponente im Forschungsprozeß, bei der in geographiedidaktischen Untersuchungen erhebliche methodische Mängel bestehen. Die vorliegende Arbeit wird an mehreren Beispielen belegen, daß *bivariate Ana-*

lysen nur zur groben Beschreibung des Datensatzes geeignet sind und nicht zur Überprüfung von (kausalen) Zusammenhängen.

(5) Die Diskussion der bivariaten Analyseverfahren zeigt die Notwendigkeit von Mehrvariablenanalysen auf. Nur multiple oder multivariate Zusammenhangsanalysen eröffnen die Möglichkeit zur (statistischen) Ausschaltung von Scheinzusammenhängen. Gleichzeitig muß jedoch berücksichtigt werden, daß viele Variablen (gerade bei didaktischen Untersuchungen) *nicht*metrisches Meßniveau aufweisen. Die vorliegende Arbeit wird deshalb verschiedene Techniken zur Zusammenhangsanalyse bei nichtmetrischen Variablen darstellen.

(6) Die einfachste Art von nichtmetrischen Verfahren zur Mehrvariablenanalyse sind *mehrdimensionale Kreuztabellen*. Es soll gezeigt werden, daß solche Kreuztabellen aus Partialtabellen aufgebaut sind. Zusammenhangskoeffizienten für diese Partialtabellen können in Analogie zum partiellen Korrelationskoeffizienten (im metrischen Fall) interpretiert werden.

(7) Ein weiteres Ziel der Arbeit ist es, das Vorurteil zu widerlegen, man könne mit nichtmetrischen Verfahren nur relativ einfache Problemzusammenhänge untersuchen. Dazu wird ein spezielles Verfahren zur Konstruktion mehrdimensionaler Kontigenztabellen vorgestellt, das auf der Basis einer *sukzessiven Chi-Quadratanalyse* beruht. Dieses Verfahren ist vergleichbar mit der (metrischen) schrittweisen multiplen Regression.

(8) Da mehrdimensionale Kreuztabellen oft so kompliziert sind, daß sie sich kaum interpretieren lassen, benötigt man ein Verfahren, welches die grundlegenden Strukturen einer solchen mehrdimensionalen Tabelle – ähnlich wie bei einer (metrischen) multiplen Regression – durch wenige Parameter repräsentiert. Mit dem *GSK-Modell* soll ein solches Verfahren vorgestellt werden.

(9) Die letzte Komponente im Forschungsprozeß, die nach unserer Einschätzung bisher vernachlässigt wurde, ist die Umsetzung von Rechenergebnissen aus statistischen Analyseverfahren in allgemeinverständliche, inhaltliche Aussagen. In der vorliegenden Arbeit soll deshalb versucht werden, die Ergebnisse der nichtmetrischen Mehrvariablenanalysen in speziellen, *graphischen Darstellungen* zu präsentieren. Es handelt sich dabei um eine Art Baumdiagramm und um ein Flußdiagramm. Beide Darstellungsformen erleichtern die visuelle Erfassung multipler Variablenzusammenhänge.

3. Die Datenbasis der Arbeit

3.1 Die RCFP-Erhebungen:

„In den Jahren 1976 bis 1978 führte das raumwissenschaftliche Curriculum-Forschungsprojekt des Zentralverbandes der Deutschen Geographen unter Leitung des RCFP-Lenkungsausschusses und durch die wissenschaftliche Vorbereitung, Begleitung und Auswertung des RCFP-Forschungsstabes die ‚Evaluation, Revision und Implementation der von RCFP entwickelten Unterrichtseinheiten' durch." (*Fürstenberg/Jungfer* 1980)

Es wurden dabei insgesamt 9 Unterrichtseinheiten evalutiert, die in den Jahren 1973 bis 1976 entwickelt worden waren. Näheres zur Entwicklungsphase der Unterrichtseinheiten findet sich bei *Engel* 1978 und bei RCFP-Lenkungsausschuß 1978. Es handelt sich dabei im einzelnen um folgende Unterrichtseinheiten:[1]

- FLUG: „Im Flughafenstreit dreht sich der Wind" – Ein geographisches Unterrichtsprojekt für die Klassen 8–10 über die Flughafenplanung München II aus der Reihe „Standortprobleme der Verkehrsinfrastruktur" (Pilotprojekt)
- TABI: „Tabi Egbe will nicht Bauer werden" – Einführung in die Entwicklungsproblematik (Kamerun) für Klassen 5–6
- RHEIN: „Tatort Rhein" – Eine geographische Unterrichtseinheit zum Curriculm „Umweltschutz: Wasser" für die Klassen 8–10
- GELT: „Der Geltinger Bucht soll geholfen werden" – Ein geographisches Unterrichtsprojekt über Wege und Probleme der „Entwicklung von Küstenräumen" für die Klassen 6–8
- BODEN: „Bodenzerstörung und Bodenerhaltung" – Eine geographische Unterrichtseinheit mit regionalen Beispielen aus den USA, der UdSSR und Mitteleuropa für die Klassen 7–8
- BRAND: „Brand in Tannenweiler" – Eine geographische Unterrichtseinheit zur Frage nach dem besten Standort von Feuerwehrstationen aus der Reihe „Infrastrukturplanung- auf der Suche nach dem besten Standort für zentrale öffentliche Einrichtungen" für die Klassen 7–8
- INDIOS: „Indios in Peru – Menschen am Rande der Gesellschaft" – Eine Unterrichtseinheit für die Klassen 7–9 aus der Reihe „Entwicklungsprobleme der Dritten Welt"
- GAST: „Gastarbeiterkinder in einer deutschen Großstadt" – Ein geographisches Unterrichtsprojekt zum Theme „Probleme sozialer Randgruppen in großstädtischen Ballungsräumen" für die Sekundarstufe II

- MOBI: „Innerstädtische Mobilität" – Eine Unterrichtseinheit über Probleme des Wohnungswechsels und der Wohnungsversorgung in einer Stadtregion für die Sekundarstufe I (9.–10. Schuljahr)

Diese RCFP-Unterrichtseinheiten wurden während der Evaluationsphase von 1975 bis 1978 bundesweit erprobt. Dabei konnten anhand mehrerer Fragebögen auch Variablen erhoben werden, denen eine über die einzelne Evaluation hinausgehende, allgemeine Bedeutung zukommt. Diese Variablen betreffen

- die Unterrichtsvoraussetzungen (Schulorganisation, Lehrer, Schülererwartungen)
- die Unterrichtsprozesse während der Erprobung (z. B. Zeitaufwand) und
- die Unterrichtsergebnisse (Leistungstest, Einstellungen der Schüler und Lehrer zur Unterrichtseinheit).

Für die neun aufgeführten Unterrichtseinheiten wurden Erhebungen durchgeführt, denn die Einheit FLUG wurde mit zwei getrennten Erhebungen (unterschiedliche Variablensätze!) evaluiert.

Vom RCFP wurden also 10 selbständige Datensätze übernommen. Bei diesen Datensätzen handelte es sich um die EDV-aufbereiteten Angaben der Erprobungsschüler zu den beiden Schülerfragebögen 1 (vor der Erprobung) und 2 (nach der Erprobung).

Wie erwähnt wurden bei den Evaluationen auch die Lehrer befragt. („Lehrerfragebogen 1"). Leider waren die EDV-aufbereiteten Daten zur Lehrerbefragung nicht mehr verfügbar. Trotz intensiver Bemühungen konnten zunächst weder die Lochkarten noch die ursprünglichen Fragebogen aufgefunden werden. Erst Ende 1979 tauchte ein Ordner mit ausgefüllten Lehrerfragebögen im Aktennachlaß des RCFP-Forschungsstabes auf. Dieser Ordner enthielt jedoch nur die Erprobungslehrer von sechs Projekten, nämlich von FLUG, BODEN, INDIOS, GAST und MOBI. Die restlichen Lehrerfragebögen blieben verschollen.

3.2 Die Aggregation zweier Gesamt-Datensätze für die Schüler- und Lehrerdaten

Diese Datensätze des Raumwissenschaftlichen Curriculum Forschungsprojektes wurden zu zwei Gesamtdatensätzen (SCHÜLER und LEHRER-SCHÜLER) zusammengeführt.

Dies war mit erheblichen datentechnischen und methodischen Problemen verbunden, da bei den empirischen Erhebungen des RCFP zu wenig Rücksicht auf die Vergleichbarkeit der Variablensätze und der Kodepläne gelegt wurde. Beispielsweise dauerte es fast ein halbes Jahr, bis allein die Datensätze zu den Schülerfragebögen vereinheitlicht, überprüft und zusammengeführt werden konnten. Da es ein Ziel der vorliegenden Arbeit ist, methodi-

sche Verbesserungen für *alle* Stufen des empirischen Forschungsprojektes vorzuschlagen, sollen in diesem Abschnitt einige der dabei häufig auftretenden Probleme angesprochen werden.

(1) Erstellung des Gesamt-Datensatzes „SCHÜLER":

Die Zusammenführung der erwähnten 10 einzelnen Datensätze aus den Schülerfragebögen des RCFP war deswegen so schwierig, weil bei den Kodeplänen folgenreiche Fehler gemacht wurden. Die schwerwiegendsten Mängel waren die folgenden:

- Es gab bei den 10 Erhebungen keine einheitliche *Nummerierung* der „Standard"-Variablen. Beispielsweise wurde die Variable „Alter des Schülers" beim Projekt INDIOS und EGBE als sechste Variable aufgenommen, sonst als fünfte Variable.
- Auch die *Anordnung* der sonstigen Variablenkomplexe in den Kodeplänen war sehr uneinheitlich. Bei allen Projekten (außer bei EGBE und Indios) war eine Batterie von Einstellungsitems zum Erdkundeunterricht enthalten: Bei den beiden Erhebungen zur Einheit FLUG (FLUG 1 und FLUG 2) wurde diese Batterie als Variable V028 bis V049 kodiert, bei allen anderen Projekten als Variable V008 bis V029. Ähnlich uneinheitlich waren auch die erhobenen Polaritätsprofile kodiert worden. Damit lagen 10 Datensätze vor, die in ihrer Struktur völlig uneinheitlich waren.
- Der gravierenste Mangel bei der Variablenkodirung betraf aber die uneinheitliche Kodierung in den *Ausprägungen* bei ein und derselben Variablen. Ein Beispiel: Die Frage „Hat Ihnen die Arbeit am Projekt Spaß gemacht?" wurde bei den Schülern bei fast allen Erprobungen gestellt. Die Anworten darauf wurden vom RCFP-Forschungsstab auf vier (!) verschiedene Weisen kodiert; wobei sowohl die Variablennummer unterschiedlich war als auch – was viel schlimmer ist – die Variablenausprägungen:

	Variablen-Nummer		kodierte Ausprägungen
Beim Projekt:	RHEIN	V088 – V095	1, 2, 3, 4, 5
	BODEN	V076 – V083	1, 2
	FLUG 1	V076 – V081	0, 1
	FLUG 2	V076 – V081	0, 1, 2, 3, 4

Tab. 4: Unterschiedliche Kodierung der Variablenausprägungen bei den gleichen Variablen durch das RCFP

– Der vierte datentechnische Mangel betraf die völlig uneinheitliche Kodierung der *„Missing-Values"*.
Die Kodierung von fehlenden Werten erfolgt normalerweise durch Zuweisung einer einheitlichen Kodenummer (z. B. 99) oder einfach durch Freilassen (*Blank*) der entsprechenden Datenstelle. Bei den RCFP-Datensätzen wurde die Kodierung der „Missing-Values" nicht nur zwischen den einzelnen Projekten ständig gewechselt, sondern auch innerhalb eines Projektes.

Vielfach wird die Auffassung vertreten, die oben angesprochenen datentechnischen Probleme seien von sekundärer Bedeutung und eher eine Angelegenheit des Programmierers. Wir möchten an dieser Stelle mit allem Nachdruck dieser Auffassung entgegentreten. Erstens sind uns mehrere (z. T. sehr umfangreiche) empirische Projekte bekannt, die ausschließlich wegen datentechnischer Unzulänglichkeiten gescheitert sind.

Zweitens werden bei einer Verharmlosung der datentechnischen Probleme auch die inhaltlichen Auswirkungen solcher Fehler übersehen. Dies läßt sich an einem Beispiel aus den RCFP-Erhebungen demonstrieren: Die unterschiedlichen Kodierung der Ausprägungen der Variablen *„Spass"* an den Erprobungseinheiten führte dazu, daß man bei der Zusammenfassung der Datensätze die ursprünglich rating-skalierten Antworten (Ausprägungen: 1 bis 5) nur als „Ja-Nein"-Antwort (also dichotom skaliert) verwenden konnte, da zwei Projekte nur mit 1 und 2 kodiert worden waren. Man mußte sich also gewissermaßen auf den „kleinsten gemeinsamen Nenner" beschränken und vorhandene Information verschenken.

Die eigentliche Zusammenführung der RCFP-Datensätze erfolgte in mehreren Stufen:

1. Stufe: *Auswahl identischer Variablen:*
Aus den Einheiten mußten jene Variablen ausgewählt werden, die bei allen 9 Erprobungen erhoben wurden. Dies führte zum Ausschluß der Projekte „EGBE" und „INDIOS", da sie so gut wie keine, mit den anderen Projekten gemeinsame Variablen aufwiesen (außer grundlegenden Variablen wie Alter, Geschlecht usw.). Es ergab sich ein Satz aus 90 Variablen (s. 3.3. ebd.), die für die restlichen Projekte identisch waren.

2. Stufe: *Umkodierung der Ausprägungen je Variable,* so daß alle verbliebenen Datensätze identische Kodierung hatten:
Bei einem Projekt waren die Kodes der „Missing-Values" zulässige, inhaltliche Variablenwerte, bei einem anderen fehlten sie. Darüber hinaus mußten die zulässigen Kodierungen mit gleichen Zahlenwerten versehen werden, was teilweise dazu führte, daß Variablenausprägungen zusammengefaßt werden mußten.

3. Stufe: *Umgruppierung der Variablen,* so daß jede der verbleibenden Datensätze eine identische Variablenreihenfolge aufwieß.

4. Stufe: *Zusammenfassung der Datensätze* zu einem Gesamt-Datensatz „SCHÜLER".

Die verbliebenen acht strukturgleichen Datensätze (zwei aus Projekt FLUG) wurden auf Magnetplatte zu einem Datensatz „zusammengespielt" und auf Magnetband gesichert.

Die gesamte Prozedur der Aggregierung des Gesamtdatensatzes „SCHÜLER" aus den RCFP-Erhebungen wird durch folgende Graphik veranschaulicht:[2] (siehe Abb. 2)

(2) Erstellung des Gesamt-Datensatzes „LEHRER-SCHÜLER":

Die Lehrerfragebögen des RCFP (Lehrerfragebogen 1) waren, wie schon erwähnt z. T. nicht mehr auffindbar. Aus dem Restbestand der Fragebögen wurde in äußerst mühevoller Detailarbeit ein Datensatz erstellt, bei dem zu jedem Lehrer die Schüler zugeordnet wurden, die an seinem Erprobungsunterricht teilgenommen hatten. Insgesamt umfaßte dieser Datensatz 116 Lehrer mit ihren 4 558 Schülern.

Um die Lehrervariablen (z. B. ihre Reformbereitschaft, ihr Geschlecht, ihr Alter usw.) mit den Schülervariablen (z. B. Spass an der Erprobung, Interesse am Fach Erdkunde) in Beziehung setzen zu können, war es nötig, die Lehrerdaten und die Schülerdaten in *einem* Datensatz zu vereinen. Nun kommen aber auf jeden Lehrer mehrere Schüler seiner Klasse, ja häufig die Schüler mehrerer Erprobungsklassen und teilweise sogar die Schüler aus den Erprobungsklassen mehrerer Projekte. Es besteht also keine 1 zu 1 Relation in der Datenstruktur. Um dieses Problem zu lösen, wurden mittels eines eigens erstellten Pascal-Programms die Daten eines jeden Lehrers so oft mal „vervielfältigt", wie von ihm Schüler vorhanden waren. Dadurch entstand ein weiterer Datensatz für die Lehrer, der die *gleiche Fallzahl* aufwies wie bei den Schülern. Diese beiden Datensätze wurden nun so aneinandergekoppelt, daß ein neuer Datensatz „LEHRER-SCHÜLER" entstand.

Dies ist nur ein grober Überblick über die sehr aufwendige Zusammenführung von Schüler- und Lehrerdaten. Es wurde relativ viel Zeit in die Erstellung dieses Datensatzes investiert, weil man zunächst davon ausgehen konnte, daß viele Einstellungen der Schüler, ihr Spaß an den RCFP-Einheiten oder ihr Lernerfolg beim Erprobungsunterricht verknüpft sind mit der Persönlichkeit und dem Unterrichtsstil ihres Lehrers. Leider wurden unsere Erwartungen gründlich enttäuscht: Die Analysen in diesem Datensatz „LEHRER-SCHÜLER" führten buchstäblich zu nichts. Dies lag vor allem daran, daß die RCFP-Lehrer offenbar eine ganz und gar untypische Lehrerauswahl darstellten. Außerdem verzerrten die Lehrer mit Mehrfacherprobungen alle Zusammenhangsanalysen, da sie ja wesentlich mehr Schüler hatten. Infolgedessen schlugen ihre Einstellungen, Unterrichtsbeurteilungen usw. überproportional „zu Buch".

Aus diesen Gründen wurde für die vorliegende Arbeit auf den gemischten Datensatz „LEHRER-SCHÜLER" ganz verzichtet. Alle folgenden Ka-

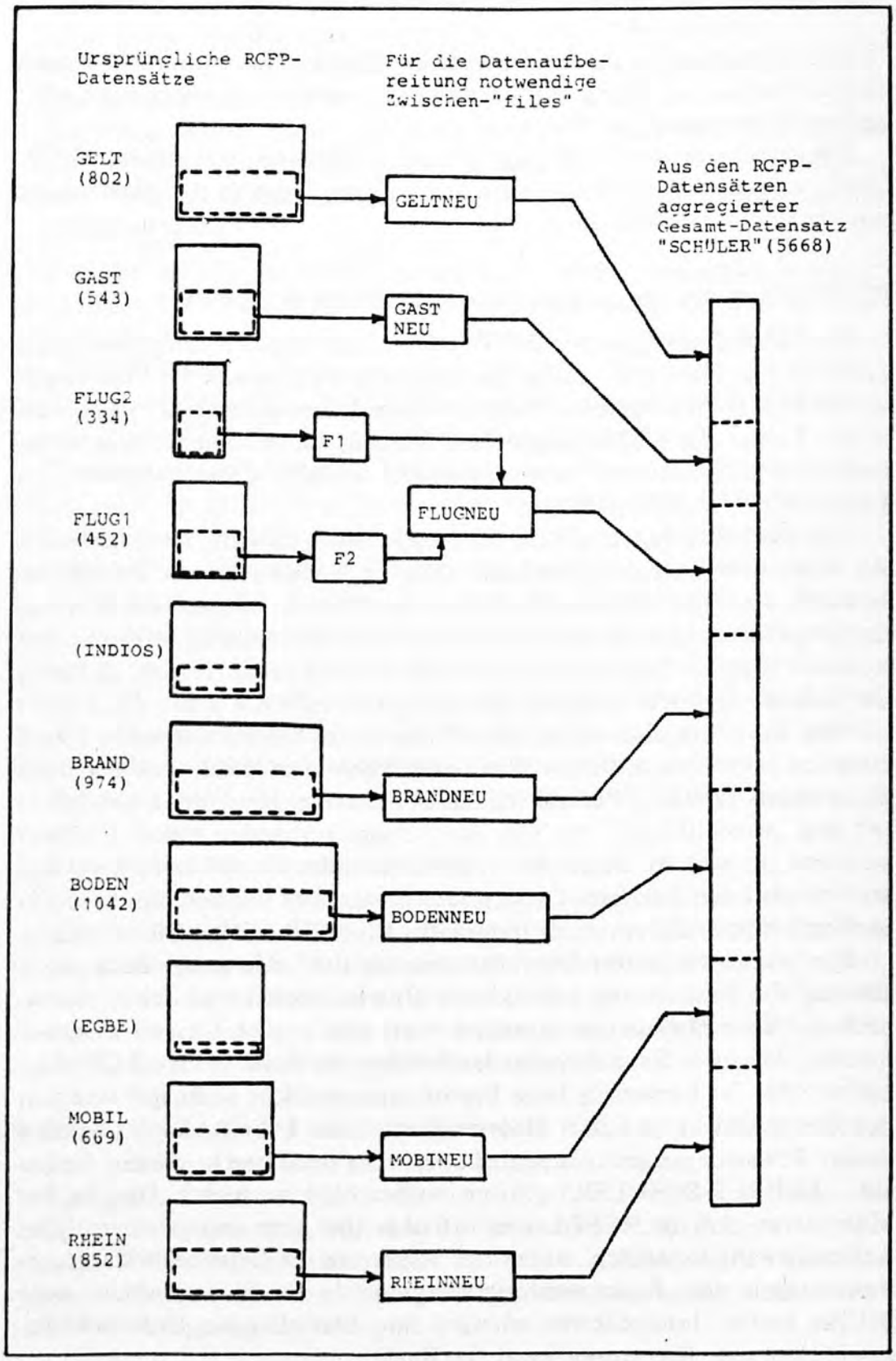


Abb. 2: Erstellung des Gesamtdatensatzes „SCHÜLER"

pitel beruhen nur auf den Daten aus dem Gesamtdatensatz „SCHÜLER".

Wir haben die Erstellung des Datensatzes „LEHRER-SCHÜLER" deshalb aber erwähnt, weil auch hier deutlich wird, daß datentechnische Mängel den Erfolg einer Untersuchung schwer beeinträchtigen könne. Gerade bei Erhebungen auf verschiedenem Aggregationsniveau (Lehrer-Schüler), also bei Mehrebenenanalysen, müssen bestimmte datentechnische Grundregeln unbedingt befolgt werden. Eine der wichtigsten Regeln lautet:

Alle zusammengehörigen „Elemente" – also hier die Lehrer mit all ihren Schülern – sollten eine identische Identifikationsnummer erhalten. Bei der RCFP-Erhebung hatten sie *verschiedene* Projekt-Nummern, mit denen gleichzeitig das jeweilige Projekt, das Bundesland der Erprobung, die Klasse und jeder einzelne Schüler erfaßt wurde. Weil man hier ein paar „Spalten" bei der Kodierung einer eigenen Identifikationsnummer sparen wollte, mußten hunderte von Arbeitsstunden aufgewendet werden, um per Hand die zusammengehörenden Projektnummern eines Lehrers zu identifizieren. Bei einer eigenen Identifikationsnummer für jeden Lehrer, die für jeden seiner Schüler identisch ist, wäre die Zusammenführung der beiden Variablensätze ein Kinderspiel gewesen. Außerdem hätte man dann sofort erkannt, daß nur sehr wenige Lehrerfragebögen tatsächlich brauchbar waren. Jeder der doppelt oder sogar mehrfach vorliegenden Fragebögen eines Lehrers mit Mehrfacherprobungen hatte nämlich nach der RCFP-Verkodung eine *andere* Projektnummer, so daß zunächst die Zahl der Erprobungslehrer wesentlich höher eingeschätzt wurde als sie tatsächlich war.

3.3 Die Variablen (Datensatz: SCHÜLER)

Der Datensatz SCHÜLER umfaßte 90 Variablen, wobei der größte Teil der Variablen aus zwei identischen Polaritätsprofilen mit je 20 Items und aus einer Einstellungsbatterie mit je 22 Statements bestand.

Folgende Variablen stammen aus dem Vortest:

Nr.	Variable	Ausprägungen/Bemerkungen
V001	Kartennummer	1–4
V002	RCFP-Projekt-Nummer	–
V003	Klassenstufe	(1) Klasse 5, (2) Klasse 6, (3) Klasse 7, (4) Klasse 8, (5) Klasse 9, (6) Klasse 10, (7) Klasse 11, (8) Klasse 12
V004	Schultyp	(1) Hauptschule, (2) Gesamtschule, (3) Realschule (4) Berufsschule, (5) Gymnasium

V005	Alter	(1) 11 Jahre, (2) 12 Jahre, (3) 13 Jahre, (4) 14 Jahre, (5) 15 Jahre, (6) 16 Jahre, (7) 17 Jahre, (8) 18 Jahre, (9) 19 Jahre
V006	Geschlecht	(1) männlich, (2) weiblich
V007	Note in Erdkunde	(1)–(6)
V008	Einstellungsstatements zum Fach Erdkunde-	
	„Einstellungs-	(1) stimmt, (2) weiß nicht, (3) stimmt nicht
V029	batterie“[3]	
V030	Polaritätsprofil zum	(1)–(5)
V049	Fach Erdkunde[3]	
V050	„Ist das Thema der RCFP-Einheit	
	für später	wurde vor der Erprobung erhoben
	wichtig?	(1) sehr wichtig – (5) völlig überflüssig
V051	Rollenspielteilnahme	(1) Ja, (2) Nein
V052	Vortragserfahrung	(1) Ja, (2) Nein
V053	Gruppenarbeit	(1) fast täglich – (5) nie

Folgende Variablen stammen aus dem Nachtest:

V054	Interesse an der RCFP-Einheit	(1) hochinteressant – (5) sehr langweilig
V055	Spaß an den einzelnen Teilen der	
V062	RCFP-Einheiten	(1) Viel Spaß – (5) kein Spaß (SPASS)[4]
V063	Nutzen der einzelnen Teile der	
V070	RCFP-Einheiten	(1) nützlich – (5) unnütz (NUTZEN)[4]
V071	Polaritätsprofil zur	
V090	RCFP-Einheit.[3]	(1)–(5)

3.4 Methodische Bewertung der Datensätze aus den RCFP-Erhebungen in Bezug auf ihre Repräsentativität

Der Datensatz „SCHÜLER" hat sicher einen hervorstechenden „Pluspunkt": er beruht auf der umfangreichsten empirischen Erhebung, die jemals in der Bundesrepublik Deutschland zum Bereich Erdkundeunterricht durchgeführt wurde. Er umfaßt – nach dem geschilderten Ausschluß zweier Projekte – 5 668 Schüler aller Schultypen, fast aller Klassenstufen und aus allen Bundesländern. Der kleinere Datensatz „LEHRER-SCHÜLER" enthält immerhin Informationen über 116 Lehrer aus der ganzen Bundesrepublik, zusammen mit ihren 4 558 Schülern; das sind insgesamt 178 Schulklassen, für die der jeweilige Erprobungslehrer noch eindeutig identifiziert werden konnte.

Wenn man bedenkt, daß neuere „empirische" Untersuchungen zum Erdkundeunterricht bislang nicht selten auf nur wenigen Dutzend Schülern eines Schultyps beruhen, so sind die vorliegenden Daten zweifellos eine wesentlich fundiertere empirische Grundlage als alles, was bisher zum Thema vorliegt.

(1) Ist Repräsentativität gegeben?

Trotz des gewaltigen Umfangs der Erhebungen sind die Daten aber leider weder repräsentativ für die Lehrer noch für die Schüler in der Bundesrepublik Deutschland.

Die Ausführungen von *Fürstenberg* und *Jungfer* zu diesem Thema sind mehr als widersprüchlich. In der zusammenfassenden Veröffentlichung zu den RCFP-Evaluationen schrieben sie:

„Die Zusammensetzung der 7 237 Erprobungsschüler wies keine außergewöhnlichen Merkmale auf. Die Ähnlichkeit der Stichproben aus den verschiedenen Untersuchungen läßt darauf schließen, daß die Schüler *repräsentativ für ihre jeweiligen Grundgesamtheiten zusammengesetzt waren*." (Hervorhebung von mir.)

Wenige Sätze vorher schreiben die Autoren: „Auffallendes Kennzeichen in allen Erprobungsklassen war der starke Anteil der Schulart Gymnasium (zwei Drittel der Klassen), während der Anteil der Hauptschulen bei nur 30%–20% lag." Und: „Die RCFP-Erprobungslehrer waren jedoch auch in anderer Hinsicht nicht repräsentativ für ihren Berufsstand: Es waren Lehrer mit weniger Schulpraxis (. . .) aber hoher Motivation (. . .). Insbesondere aber unterscheiden sie sich von Kollegen in ihrer Haltung gegenüber Reformzielen, wie sie das RCFP anstrebt (sie beurteilen diese wesentlich positiver)." (alles: *Fürstenberg/Jungfer* 1979, S. 10). Dem ist eigentlich nichts hinzuzufügen. Die RCFP-Erhebungen sind ganz eindeutig weder repräsentativ für die Lehrer noch für die Schüler, wobei die Lehrer mit Sicherheit sehr viel untypischer für ihre Grundgesamtheit sind als die Schüler.

Bedeutet dies, daß unsere folgenden Analysen mit dem Datensatz SCHÜLER wertlos sind?

(2) Ist Repräsentativität für unsere Analysen notwendig?

„Repräsentativität" ist kein Wert an sich, sondern nur ein methodischer Kunstgriff für eine bestimmte Art sozialwissenschaftlicher Untersuchung – den „Survey". Ein „Survey" ist eine Erhebung, die ein möglichst typisches Abbild der fraglichen Population hinsichtlich der interessierenden Merkmale liefern soll. Ziel eines Surveys ist die *Beschreibung* der Population und weniger die Analyse von Bedingungszusammenhängen. Die *Analyse* von Variablenzusammenhängen dagegen kommt im Prinzip vollkommen ohne Repräsentativität aus. Sie basiert letztlich auf dem experimentellen Design. Hier wird die Größe und Zusammensetzung der Stichprobe ausschließlich durch die Anforderungen des statistischen Beweisverfahrens bestimmt. Ein Beispiel: Angenommen, es soll untersucht werden, ob Gymnasiasten interessierter am Fach Erdkunde sind als Grundschüler oder Berufsschüler: Zur Klärung dieser Frage genügt es im Prinzip, wenn man drei gleich große (Zufalls-)Stichproben von Grund-, Berufs- und Gymnasialschülern zieht und mittels t-Test (oder Varianzanalyse) auf Unterschiede im „Interesse am Fach Erdkunde" überprüft. Ob es „in Wirklichkeit" doppelt so viele Grundschüler wie Gymnasiasten gibt, ist – bei dieser Fragestellung – vollkommen irrelevant.

Unsere Arbeit ist an der Aufdeckung von Bedingungszusammenhängen interessiert, und nicht an der Beschreibung des „typischen Schülers" hinsichtlich seiner Einstellungen. Es interessiert also nicht die Frage „was denken „die" Schüler vom Fach Erdkunde?", sondern „bestehen Zusammenhänge zwischen bestimmten Einstellungen von Schülern und anderen Variablen?". Zu diesem Punkt wäre noch viel zu sagen, da das Suchen nach Bedingungszusammenhängen (statt der Beschreibung einer Population) natürlich nicht das Problem fehlender Repräsentativität automatisch erledigt. Es wird nur verlagert auf einen eventuell folgenden, nächsten Untersuchungsschritt: Selbst wenn zwischen zwei (oder mehr) Merkmalen ein Zusammenhang nachgewiesen ist, bleibt die Frage, ob diese Merkmale in der Grundgesamtheit häufig oder selten vorkommen; d. h. ob der bewiesene Zusammenhang wichtig und typisch ist oder nicht. Dazu muß man aber wissen, wie häufig diese Merkmale in der Population auftreten – was aber nur mit einem repräsentativen Survey möglich ist. Zusammenfassend kann man festhalten, daß für die Art unserer Fragestellungen eine (globale) Repräsentativität (noch!) nicht notwendige Voraussetzung ist.

(3) Wäre Repräsentativität bei den RCFP-Untersuchungen überhaupt herstellbar gewesen?

Bei ganz wenigen empirischen Erhebungen besteht eine sog. echte, globale Repräsentativität. D. h. in den seltensten Fällen repräsentiert die Stichprobe

alle in Frage kommenden Merkmale der zugrunde liegenden Grundgesamtheit. Dazu müßte nämlich eine reine Zufallsauswahl unter den Elementen der Grundgesamtheit durchgeführt werden. Dies können sich normalerweise nicht einmal kommerzielle Sozialforschungsinstitute leisten, abgesehen davon, daß es oft auch technisch gar nicht möglich ist. Üblicherweise wird deshalb ein einfacheres Verfahren benutzt, die geschichtete Zufallsstichprobe: Hier werden die Elemente der Stichprobe von vornherein nur nach wenigen Merkmalen repräsentativ ausgewählt. Man schichtet beispielsweise nach „Geschlecht", „Alter", „Stadt-Land" usw.. Nur hinsichtlich dieser Merkmale ist dann die Repräsentativität gewährleistet.

Es wäre aber eine überzogene Forderung, wollte man verlangen, daß die RCFP-Stichprobe repräsentativ für den Schultyp, die Alterszusammensetzung der Schüler, das Geschlechtsverhältnis der Schüler, die Verteilung auf die Bundesländer, auf Alter und Schulerfahrung der Erprobungslehrer usw. sein soll. Würde man nach obigen 6 Merkmalen eine geschichtete Stichprobe ziehen wollen, so müßte man aus Statistiken die Proportionen der jeweiligen Gruppen ermitteln und Quoten für die Zufallsauswahl vorgeben. Bereits bei 6 Merkmalen ergäbe dies einen ungeheuren Aufwand.

Aber nicht nur dieser Aufwand verbietet einen strengen Repräsentativitätsanspruch. Die Frage ist auch, ob Repräsentativität hinsichtlich der Schülerpopulation mit Repräsentativität hinsichtlich der Lehrerpopulation überhaupt zu verbinden ist. Man sieht, daß schon vom rein stichprobentechnischen her gesehen realistischerweise nur hinsichtlich ganz weniger Merkmale eine Repräsentativität *gesichert* werden kann.

Noch größere Schwierigkeiten ergeben sich aus inhaltlich-theoretischen Gründen: Die einzelnen Erprobungseinheiten des RCFP sind jeweils mit ihren Lernzielen und Unterrichtsmethoden auf bestimmte Klassenstufen hin zugeschnitten. Damit ist klar, daß die Stichprobe der Erprobungsschüler nicht für alle Schüler schlechthin repräsentativ sein kann, sondern nur für jene Schülerpopulationen, die der Zielgruppe der RCFP-Unterrichtseinheiten entspricht. Da aber nicht alle Altersgruppen und Schultypen gleichmäßig von den verschiedenen Projekten abgedeckt sind, ist es *zwangsläufig*, daß unser Datensatz nicht repräsentativ für die Schülerpopulation im allgemeinen ist.

Eine ähnliche zwangsläufige Diskrepanz besteht zwischen den Erprobungslehrern des RCFP und der gesamten Lehrerschaft. Man kann nicht erwarten, daß beispielsweise reformfeindliche Lehrer in jener Proportion unter den Erprobungslehrern vertreten sind, der ihrem Anteil in der Lehrerschaft entspricht. Lehrer, die Reformmodellen wie dem RCFP kritisch gegenüberstehen, werden kaum bereit sein, ausgerechnet bei deren Erprobung mitzuwirken. Man kann also aus vielfältigen Gründen dem RCFP-Forschungsstab keinen Vorwurf für die zweifelhafte Repräsentativität der Erhebung machen. Immerhin aber sind im Datensatz „Schüler" alle

Schultypen, fast alle Klassenstufen, alle Bundesländer und natürlich beide Geschlechter vertreten.

Damit übertrifft dieser Datensatz bei weitem den Stand bisheriger Erhebungen zum selben Thema, auch wenn er nicht die theoretisch mögliche Stichprobenqualität hat.

4. Einführung in die Reliabilitäts- und Dimensionalitäts-Analyse

Die Überprüfung und Verbesserung der Meßsicherheit – der sog. Reliabilität – ist ein wesentlicher Schritt in jeder empirischen Untersuchung.

Sie beginnt mit der Frage, wie man die Informationen aus der empirisch gegebenen Wirklichkeit in „Daten" übersetzen will. Man muß dazu ein bestimmtes Meßinstrument verwenden; sei es eine einzelne, schlichte Frage, eine Ratingskala, eine Fragenbatterie, die zu einem Index zusammengefaßt wurde, ein Test aus Einzelaufgaben oder auch ein kompliziertes probabilistisches Meßmodell wie das von Rasch. Je nachdem, mit welchem Verfahren die Daten erhoben werden, ergeben sich dabei ganz unterschiedliche Probleme bei der Meßsicherheit.

In den folgenden drei Kapiteln werden wir uns mit einigen Aspekten dieser Meßproblematik befassen.

Dabei konzentrieren wir uns auf Verfahren, die sich zur Überprüfung und Verbesserung der Meßqualität von Statement-Batterien und von Polaritätsprofilen eignen, wie sie bei den RCFP-Erhebungen eingesetzt wurden. Auf die sog. Test-Retest-Reliabilitätsanalysen (Verfahren, die auf einer Widerholungsbefragung basieren) mußten wir leider verzichten, da sie bei der RCFP-Untersuchung nachträglich nicht mehr durchgeführt werden konnten.

Außerdem sollten die benutzten Verfahren keine hochkomplizierten „Labormethoden" sein, sondern relativ einfache Techniken, die in der Forschungspraxis sinnvoll einzusetzten sind.

Auf der Grundlage dieser Voraussetzungen waren die klassische Kennwertanalyse und die Faktorenanalyse Verfahren, die uns für eine Reliabilitätsüberprüfung und -verbesserung geeignet erschienen.

4.1 Methodischer Gewinn der Reabilitätsanalyse

Um den methodischen Gewinn derartiger Reabilitätsuntersuchungen verdeutlichen zu können, ist es nötig, das Grundproblem der Meßproblematik vorweg zu erörtern.

(1) *Das Grundproblem der Meßproblematik:*

Beginnen wir mit einer einfachen Skizze: (Abb. 3)

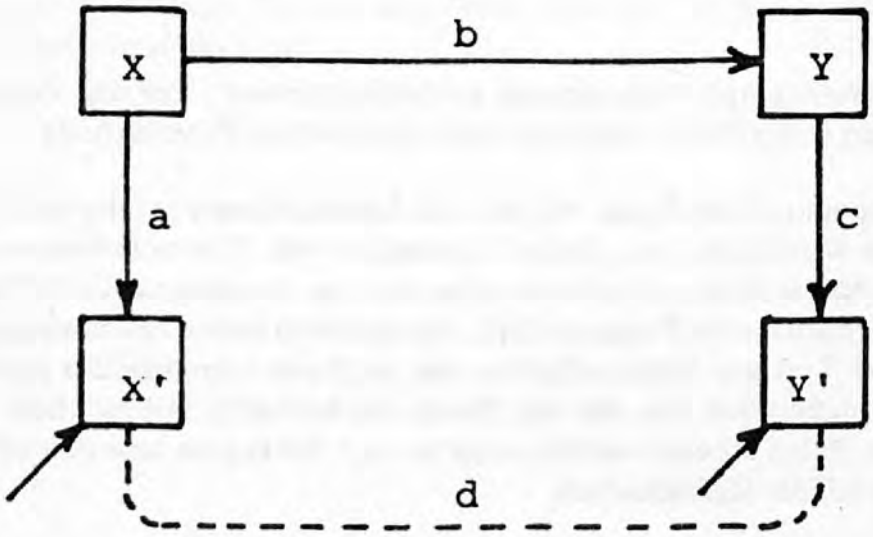


Abb. 3: Grundproblem der Meßproblematik

Die Skizze stellt das Grundproblem der Meßproblematik dar. X und Y sind zwei Variablen, b ist ein theoretisch angenommener (kausaler) Zusammenhang zwischen diesen beiden Variablen. X, b und Y veranschaulichen eine Hypothese: Die unabhängige Variable X, der angenommene Zusammenhang b und die abhängige Variable Y bilden ein *theoretisches Konstrukt*, das mit empirischen Daten überprüft werden soll. Dazu müssen die Variablen X und Y *mit Daten aus der empirischen Wirklichkeit verknüpft* werden. Man sagt: Zu den theoretischen Konstrukten X und Y müssen empirische „Referenten" X' und Y' gefunden werden. Diese empirischen Referenten sind nichts anderes als die durch eine bestimmte Operationalisierung gewonnenen Daten. Die Daten kann man auch als „Indikatoren" für das theoretische Konstrukt, also die Variablen, bezeichnen.

Das Problem der Reliabilität sind nun die Beziehungen a und c. d. h. die Frage, ob die Indikatoren X' und Y' in einem konstanten, stabilen Verhältnis zu X und Y stehen. Die Beziehungen a und c sind deshalb so wichtig, weil der Hypothesenzusammenhang b zwischen X und Y ja nur indirekt überprüft wird: Es wird untersucht, ob zwischen den empirischen Indikatoren X' und Y' ein (statistischer) Zusammenhang d besteht. Aus der Existenz von d wird auf den Hypothesenzusammenhang b geschlossen. Dieser Schluß von der Beziehung zwischen den empirischen Indikatoren auf das theoretische Konstrukt ist nur dann gerechtfertigt, wenn die Indikatoren mit einem stabilen Verhältnis zu den (theoretischen) Variablen stehen. Wir werden dies gleich an einem inhaltlichen Beispiel erläutern. Zunächst soll

aber das Grundmodell der Meßproblematik noch etwas ausgeweitet werden, damit es der üblichen Forschungssituation besser entspricht. Dazu dient die folgende Abbildung 4.

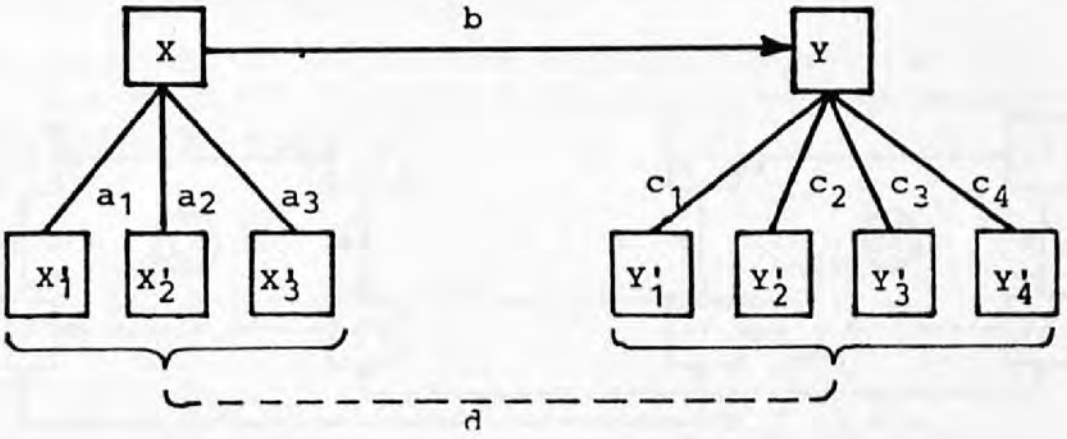


Abb. 4: Erweitertes Grundmodell der Meßproblematik

Zur Erfassung einer Variablen bedient man sich oft *mehrerer* Indikatoren. Nehmen wir an, X sei das „Interesse am Fach Erdkunde": Dann könnte man im Prinzip zwar die „Anzahl der Meldungen eines Schülers in einer Erdkundestunde" als einzigen Indikator X' für das Interesse dieses Schülers nehmen. Normalerweise wird man aber – aus theoretischen Erwägungen – mehrere Indikatoren benutzen, z. B. Einstellungsstatements, die sich auf verschiedene Aspekte des Interesses beziehen. X'_1, X'_2 und X'_3 wären in diesem Fall dann z. B. die Items einer Einstellungsbatterie zum Erfassen von „Interesse".

Nehmen wir weiter an, Y sei die abhänige Variable „Lernerfolg bei der RCFP-Unterrichtseinheit", dann wird man natürlich diese Variable auch nicht durch eine einzige Wissenfrage operationalisieren, sondern durch einen Test aus mehreren Aufgaben. Die Hypothese könnte bei diesem Beispiel lauten: „Je größer das Interesse, desto höher der Lernerfolg bei der RCFP-Einheit". Dieser theoretisch postulierte Zusammenhang b wird überprüft durch den Zusammenhang d zwischen den Indikatorgruppen. Nur wenn die Stabilität von a_1, a_2, a_3 bzw. c_1, c_2, c_3 und c_4 gewährleistet ist, kann man von d auf b schließen. Es ist nützlich, wenn man sich die Zusammenhänge a= (a_1, a_2, a_3), b, c= (c_1, c_2, c_3, c_4) und d als Korrelationszusammenhänge vorstellt. Dann wären bei einem vollständig reliablen Meßvorgang die Korrelationskoeffizienten zu a und c gleich 1,00 (vergl. dazu: *Blalock* 1971, S. 301)

An diesem erweiterten Grundmodell des Meßvorganges wollen wir nun einige Aspekte der Reliabilität aufzeigen.

(2) *Test – Retest – Reliabilität:*

Die Beziehungen a bzw. c sind u. a. dann reliabel, wenn sie zeitlich konstant sind. Dies versucht man durch die sog. Test – Retest – Reliabilitätsüberprüfung zu klären. Im einfachsten Fall (keine Mehrfachindikatoren) sieht die Reliabilitätsüberprüfung durch Testwiederholung so aus:

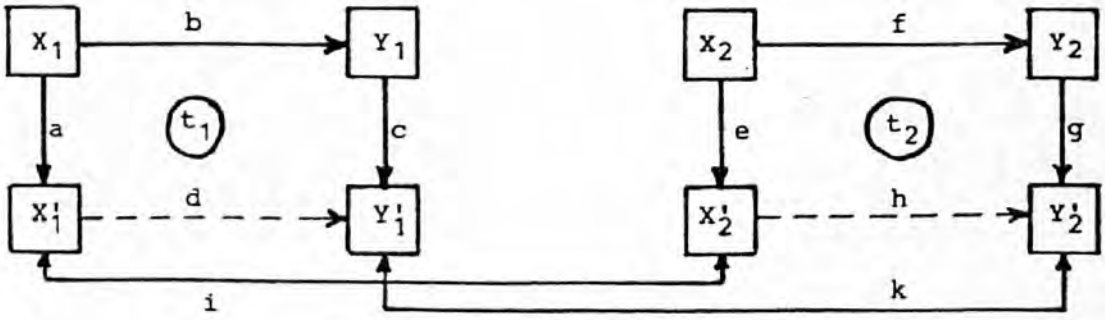


Abb. 5: Modell der Test-Retest-Reliabilität

Bei der Test-Retest Reliabilität geht es um die Beziehungen i und k.

Angenommen, wir hätten ursprünglich maximale Meßsicherheit bei a und c. Zeitliche Stabilität der Messung würde dann bedeuten, daß auch e und g maximal genau messen. Wir überprüfen dies, indem wir die Indikatorenwerte X'_1 und X'_2 bzw. Y'_1 und Y'_2 vergleichen, z. B. indem wir eine Korrelation rechnen. Bis auf zufällige Fehler müßte diese Korrelation 1,00 sein.

Da wir die Test-Retest Reliabilität an unserem RCFP-Datensatz nicht mehr überprüfen können, wollen wir dieses Verfahren nicht mehr weiter verfolgen.

Reliabilitätsanalysen nach der Methode der Testwiederholung sind dann unvermeidbar, wenn pro Variable nur jeweils – wie in unserem ersten Grundmodell – ein Indikator vorliegt.

(3) *Paralleltest-Reliabilität:*

Da wir aber oft mehrere Indikatoren zur Erfassung einer Variablen haben (Grundmodell 2) können wir häufig die sog. Paralleltest-Methoden verwenden.

Bei der Paralleltestmethode werden zwei vergleichbare Indikatorensätze (Paralleltests) bei der Messung benutzt, wobei die Versuchspersonen zufällig auf die Parallelformen aufgeteilt werden. Die Reliabilität der beiden Indikatorensätze ergibt sich daraus, wie eng die beiden Parallelformen miteinander korrelieren.

Bei der Reliabilitätsüberprüfung durch Paralleltests besteht die Schwierigkeit, zwei Indikatorensätze zu entwickeln, die wirklich inhaltlich vergleichbar sind.

(4) *Splitt-half-Reliabilität/Homogenität:*

Die am weitesten entwickelte Form der Reliabilitätsanalyse (vgl. *Schanz* 1973, S. 47) sind die „Split-Half"-Methode und die Homogenitätsanalyse (vgl. *Bohrnstedt/Borgatta* 1980, S. 141).

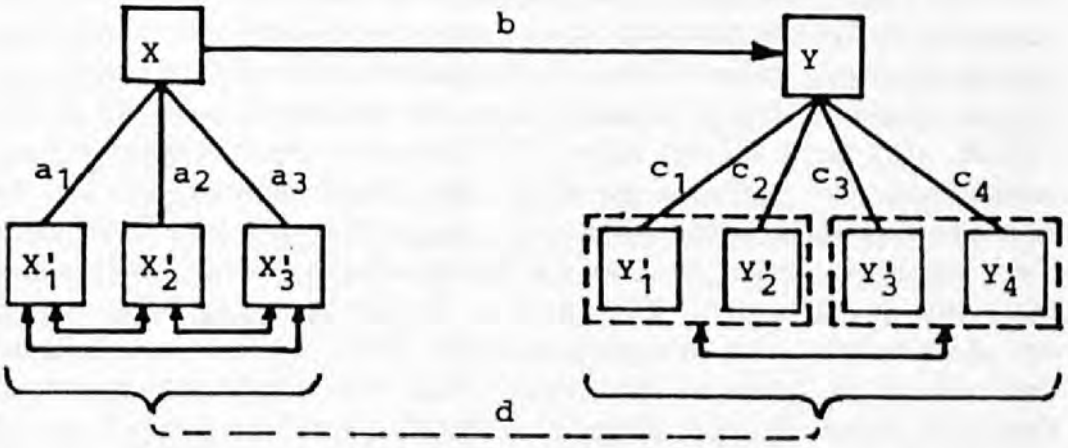


Abb. 6: Modell der Split-Half-Reliabilität und der Homogenitätsanalyse

Die „Split-half-Methode" wird veranschaulicht durch die Indikatoren Y'_1, Y'_2, Y'_3 und Y'_4. Stellen wir uns vor, dies wären vier Items einer Einstellungsbatterie. Bei der „Split-half-Methode" würden zwei gleich große Gruppen aus diesen Items gebildet (z. B. Y'_1 und Y'_2 bzw. Y'_3 und Y'_4; genauso gut könnte man aber auch Y'_1 und Y'_3 bzw. Y'_2 und Y'_4 zusammenfassen). Die Korrelation zwischen den beiden „Hälften" wäre ein Maß für die Reliabilität der ganzen Einstellungsbatterie.

Bei der Homogenitätsanalyse ist die Aufteilung des Indikatorensatzes (X'_1, X'_2, X'_3) bis an die möglichen Grenzen gesteigert worden: Jeder Indikator wird jeweils mit dem Rest der Indikatoren verglichen. Dadurch erhält man Reliabilitätskoeffizienten für jeden einzelnen Indikator (Item) und kann dadurch „schlechte" Indikatoren (nachträglich) entfernen.

Einen großen Wert haben Reliabilitätsanalysen bei Meßanordnungen mit *Mehrfach*indikatoren, d. h. bei Fragenbatterien, Polaritätsprofilen und (Wissens-)Tests. Hier sind Homogenitätsanalysen möglich, bei denen man einzelne Indikatoren oder Items als relativ schlecht identifizieren und somit auch nachträglich noch aussondern kann. Homogenitätsanalysen sind deshalb das ideale Instrument, um aus einem größeren „Itempool" (zu einem bestimmten Problem) schlechte Items auszusondern.

4.2 Die Faktorenanalyse als ein Verfahren zur Verbesserung der Meßqualität

Die Reliabilität von „Indikatoren" wäre viel leichter zu überprüfen und zu verbessern, wenn das Problem der Dimensionalität nicht wäre. Dieses Problem soll nun erläutert werden.

(1) Die Dimensionalität von Indikatoren

Wie wir festgestellt haben, werden sehr häufig Mehrfachindikatoren zur Messung einer Variablen herangezogen. Homogenitätsanalysen können sicherstellen, daß die einzelnen Indikatoren „ähnlich gut", eben homogen, messen. Bei einer solchen Homogenitätsanalyse werden jedoch leider auch Einzelindikatoren (Items) ausgeschlossen, die eigentlich „reliabel" messen würden, aber nicht auf der selben Meßdimension liegen wie die anderen Indikatoren. Was verstehen wir dabei unter Meßdimension? Denken wir dazu an eine Itembatterie zur Messung von „Interesse am Fach Erdkunde". Die Variable „Interesse" ist dabei ein theoretisches Konstrukt, das mehrere Einzelphänomene umfaßt: Zum Interesse der Schüler gehört beispielsweise die „Bereitschaft" sich mit geographischen Themen (auch außerhalb des Unterrichts) zu befassen, ihr „Fleiß" und ihre „Aufmerksamkeit" im Unterricht selbst oder der „Ehrgeiz" der Schüler, gute Noten zu bekommen. Das Interesse am Fach Erdkunde ist selbstverständlich auch (unbewußt) verbunden mit der Einschätzung des jeweiligen Erdkundelehrers, oder mit strategischen Überlegungen der Schüler darüber, in welchem Fach man leichter als in anderen Fächern gute Noten bekommen könnte. Dies und anderes mehr schwingt im Bewußtsein der Schüler mit, wenn man sie nach ihrem Interesse am Fach Erdkunde frägt. Damit wird deutlich, daß die Variable mehrere Dimensionen hat.

Diese Diemensionalität entsteht

- durch die *Variable selbst*, also die Art, wie das theoretische Konstrukt definiert ist,
- durch die *Sprache*, in der Indikatoren formuliert werden
- und durch die *„Perspektive"* der jeweiligen Versuchsperson.

Ist beispielsweise die *Abstraktionsebene* einer Variablen sehr hoch, dann umfaßt sie zwangsläufig mehr Einzelaspekte eines empirischen Phänomens als bei einer sehr konkreten Definiton. „Lebensqualität" wäre z. B. eine Variable, die zwangsläufig Indikatoren aus verschiedenen Lebensbereichen erfordert. Diese Variable hätte (ceteris paribus) mehr Dimensionen als z. B. die Variable „Arbeitszufriedenheit".

Die *Sprache* zwingt uns mit ihren Strukturen Erfassungsschemata auf, die uns oft gar nicht bewußt werden: Enthalten die Indikatoren affektiv oder politische vorbesetzte Begriffe (wie z. B. „Schulstreß", „Leistungs-

druck", usw.) dürfen wir uns nicht wundern, wenn unsere Variablen die konventionellen Stereotypen aus der öffentlichen Diskussion widerspiegeln. Problematischer ist jedoch, daß das Sprachniveau einen Einfluß darauf hat, wie differenziert eine „Sache" gesehen werden kann.

Die wohl wichtigste Ursache für die Dimensionalität einer Variablen liegt in der *Perspektivität* aller Erfahrung: „Interesse" am Fach Erdkunde ist etwas, das beispielsweise ein „Klassenprimus" unter einer ganz anderen Perspektive beurteilen wird, als ein schlechter Schüler. Ihre unterschiedliche Perspektive ergibt sich aus dem jeweiligen Bezugsrahmen, in den die beiden Schüler die Indikatoren der Variablen „Interesse" stellen werden.

Perspektivität ist eine Folge des unterschiedlichen Standorts: Beleuchtet man mit einem Scheinwerfer einen Gegenstand aus unterschiedlichen Positionen, dann erhellt der Lichtkegel immer neue Seiten dieses Gegenstandes. Unser Bewußtsein ist wie dieser Lichtkegel[1]. Es hebt jene Aspekte eines Phänomens hervor, die für uns aus unserem augenblicklichen Standort wichtig sind.

Das Interesse eines Schülers am Fach Erdkunde kann deshalb ganz unterschiedliche Dimensionen umfassen, wenn er seinen Standort (evtl. auch nur „im Geiste") wechselt und die Sache – wie der Volksmund sagt – in einem „anderen Licht" sieht.

(2) Die Aufdeckung von Dimensionen

Eine Reliabilitätsanalyse ist nur schlecht in der Lage eine solche eventuell vorhandene Mehrdimensionalität einer Variablen aufzudecken. *Jeffrey K. Smith* hat an einem empirischen Beispiel nachgewiesen, daß ein hoher Reliabilitätskoeffizient α bei einer Homogenitätsanalyse keineswegs automatisch Eindimensionalität des Indikatorensatzes bedeutet (*Smith* 1980, S. 885 ff.). Er schreibt:

"... They (die Daten) clearly illustrate that a high coefficient alpha ist a necessary but *not* sufficient condition or unidimensionality ..." „It must be realized that a high coefficient alpha does not preclude the possibility of two or more related, but distinct dimensions, operating in that test." (*Smith* 1980, S. 887).

Eine klassische Reliabilitätsanalyse – beispielsweise in Form einer Homogenitätsuntersuchung – reicht für eine Überprüfung der Meßqualität nicht aus. Sie muß unbedingt durch ein spezielles Verfahren der Dimensionsanalyse ergänzt werden.

Die Frage der Dimensionalität wird leider häufig unterschätzt. Eine ganze Reihe von Skalierungsverfahren unterstellt schlicht und einfach Eindimensionalität des Itemsatzes. Dazu gehört beispielsweise die sog. Guttman-Skalierung und die höchst aufwendige Skalierung nach dem probabilistischen Meßmodell von *Rasch*. Gerade das letzte Verfahren birgt die Gefahr in sich, daß ein empirisch gegebenes Antwortverhalten künstlich in *eine*

Meßdimension hineingezwängt wird, da die Abweichungen von der idealen (eindimensionalen) Skala teilweise als „zufällige" Meßfehler interpretiert werden können. In seinem oben zitierten Aufsatz hat *Smith* ausdrücklich auf dieses Problem aufmerksam gemacht. Er konnte an dem empirischen Beispiel zeigen, daß die „Anpassungs-Statistik" nach dem Rasch-Modell Eindimensionalität anzeigt, obwohl die Items ganz offensichtlich *zwei* Dimensionen einer Variablen abdecken. (*Smith*, 1980, S. 888)

Es gibt verschiedene Verfahren, um die Meßdimensionen eines Indikatorenansatzes aufzudecken, wobei jedes Verfahren auf theoretischen Prämissen darüber beruht, wie der Perzeptionsraum der Versuchsperson am besten abgebildet werden könnte. Das verbreitetste Verfahren neben der multidimensionalen Skalierung ist sicher die Hauptkomponentenanalyse, die einfachste Form einer Faktorenanalyse.

(3) Zusammenhänge zwischen Faktorenanalyse und Reliabilitätsanalyse

Die Faktorenanalyse bietet sich nicht nur deswegen zur Dimensionsaufdeckung eines Indikatorensatzes an, weil sie (mit SPSS) leicht durchführbar ist. Es bestehen vielmehr mathematische Zusammenhänge zwischen den Reliabilitätskoeffizienten zur Messung der Homogenität und der Faktorenanalyse; genauer gesagt: den Eigenwerten, auf denen sie beruht.

V. L. Greene und *E. G. Carmines* haben diese Zusammenhänge mathematisch bewiesen (*Greene/Carmines* 1980, S. 160ff.).

Wir ersparen es uns, die formale Ableitung der Zusammenhänge hier zu referieren. Inhaltlich gesehen zeigen die Autoren, daß aus den Ladungen der ersten Hauptkomponente bei der Faktorisierung ein spezieller alpha-Koeffizient berechnet werden kann (genannt „θ"), der die obere Grenze für den Homogenitätskoeffizienten darstellt. Daraus ergibt sich, daß der übliche Koeffizient α die Untergrenze der Reliabilität eines Indikatorensatzes markiert. Um die Reliabilität des Itemsatzes zu erhöhen, müssen die Items ausgesondert werden, die *nicht* auf der jeweiligen Meßdimension liegen. Das geschieht bei einer Hauptkomponentenanlyse dadurch, daß die einzelnen Items pro Faktor „Gewichte" (die sog. Ladungen) bekommen. Diese Gewichte spiegeln wider, welche Bedeutung jedes einzelne Item für diese eine Meßdimension hat – bei einer Ladung von (nahe) 0 ist sein Gewicht gering und das Item wird praktisch, als nicht zur Meßdimension gehörig, ausgeschlossen. Diese so gewichteten Items haben – nach *Greene* und *Carmines* – dann eine maximale Reliabilität: „... maximizing led us to a principal component analysis of the items..." (*Greene/Carmines* 1980, S. 171).

Die Durchführung einer Hauptkomponentenanalyse deckt also nicht nur auf, ob der Indikatorensatz mehrdimensional ist. Vielmehr führt der Gewichtungsprozeß durch die Ladungen je Dimension automatisch zu einer Erhöhung der Reliabilität. Man kann sich diesen Zusammenhang so

vorstellen: Während bei der normalen Berechnung des Koeffizienten α alle Indikatoren mit dem Gewicht „1" in die Berechnung eingehen, bewirkt die Hauptkomponentenanalyse, daß jene Indikatoren durch die Ladungen heruntergewichtet werden, die nicht so optimal auf der jeweiligen Meßdimension liegen.[2]

4.3 Zusammenfassung und Schlußfolgerungen

Reliabilitätsanalysen sind notwendig, um von den gemessenen „Indikatoren" auf die theoretischen Variablen schließen zu können. Ein Aspekt der Reliabilität betrifft die Frage der Homogenität eines Satzes aus mehreren Indikatoren. Die Homogenitätsfrage läßt sich nicht klären ohne Berücksichtigung einer möglichen Mehrdimensionalität des Indikatorensatzes.

Wir werden im folgenden mit RCFP-Daten eine „klassische" Kennwertanalyse durchführen. Stellt sich dabei der Verdacht ein, die jeweiligen Itemsätze seien mehrdimensional, so wird eine Hauptkomponentenanalyse durchgeführt. Auch dieser Analyseschritt soll im folgenden dargestellt werden. Bei beiden Verfahren wollen wir im Auge behalten, daß sie nur *ein* Schritt im Forschungsprozeß sind. Sie dürfen nicht durch übertriebene methodische Ansprüche überfrachtet werden. Der Gewinn an Meßsicherheit muß in einem vertretbaren Verhältnis zum eingesetzten Methodenaufwand stehen.

5. Die Meßqualität der Einstellungsbatterie zum Fach Erdkunde allgemein

Die Einstellungsbatterie des RCFP zum Erdkundeunterricht bestand aus 22 Einzelitems, die drei Einstellungsdimensionen abdecken sollten. Die drei Dimensionen waren:

- Interesse am Fach Erdkunde (Items I1 bis I8),
- Wichtigkeit des Faches Erdkunde (Items W1 bis W8).
- Schwierigkeit des Faches Erdkunde (Items S1 bis S5).

Die Antwortmöglichkeiten waren: (1) Stimmt, (2) Weiß nicht, (3) Stimmt nicht. Die Reihenfolge der Items im Fragebogen entsprach nicht den drei intendierten Einstellungsdimensionen. In der folgenden Tabelle wurden sie jedoch nach diesen Dimensionen geordnet:

5.1 Reliabilitätsanalyse der Einstellungsbatterie

Nach einer Reihe von vorbereitenden Arbeiten (sinngemäß richtige Umpolung der Statements, Eliminierung der „Weiß-Nicht"-Kategoie) wurden für die drei „Dimensionen" der Einstellungsbatterien folgende Kennwerte aus der klassischen Testtheorie berechnet:

- die Schwierigkeit der Items,
- die Trennschärfe der Items,
- die quadrierte multiple Item-Korrelation
- und der Reliabilitätskoeffizient α nach *Cronbach.*

Es muß betont werden, daß die obengenannten drei „Dimensionen" nur die vom RCFP theoretisch angenommenen drei Einstellungsaspekte „Wichtigkeit", „Interesse" und „Schwierigkeit" sind. Ob diese drei Gruppen von Einstellungsstatements auch tatsächlich Dimensionen im meßtheoretischen Sinn darstellen, wird noch – durch eine Faktorenanlyse – zu klären sein.

Der Berechnung der Kennwerte lagen jedoch zunächst diese „theoretisch" intendierten Dimensionen „Interesse", „Wichtigkeit" und „Schwierigkeit" zugrunde. Sie wurden als Subskalen aufgefaßt, so daß die Kennwerte die Qualität der Items für diese Subskalen angeben. Da es verschiedene Varianten bei der Berechnung der Kennwerte gibt, wollen wir unsere Vorgehensweise im Detail darstellen. Wir haben uns bei unseren Berechnungen vor allem an die Empfehlungen von *Lienert* (1969), von *Dieterich* (1977), von *Specht* (o. J.), von *Bergler* (1975) und von *Silverstein* (1980) gehalten.

081	I1	Andere Unterrichtsfächer interessieren mich mehr als gerade Erdkunde (EK).	1	③	2
009	I2	Ich interessiere mich ganz allgemein für EK.	1	3	2
013	I3	Der Unterrichtsstoff in EK ist eigentlich ziemlich trocken.	1	③	2
017	I4	Der Ek-Unterricht hat mich angeregt, über einige seiner Themen weiter nachzudenken.	1	3	2
020	I5	Der EK-Unterricht macht mir Spaß	1	3	2
021	I6	EK langweilt mich häufig	1	③	2
023	I7	Wenn ich die Unterrichtsfächer frei wählen könnte, würde	1	③	2

081	I1	Andere Unterrichtsfächer interessieren mich mehr als gerade Erdkunde (EK).	1	③	2
009	I2	Ich interessiere mich ganz allgemein für EK.	1	3	2
013	I3	Der Unterrichtsstoff in EK ist eigentlich ziemlich trocken.	1	③	2
017	I4	Der Ek-Unterricht hat mich angeregt, über einige seiner Themen weiter nachzudenken.	1	3	2
020	I5	Der EK-Unterricht macht mir Spaß	1	3	2
021	I6	EK langweilt mich häufig	1	③	2
023	I7	Wenn ich die Unterrichtsfächer frei wählen könnte, würde ich EK nicht nehmen.	1	③	2
026	I8	EK ist mein Lieblingsfach	1	3	2
010	W1	Das was ich im EK-Unterricht erfahre, kann ich auch außerhalb der Schule gut gebrauchen.	1	3	2
014	W2	Der EK-Unterricht regt mich an, über gesellschaftliche Probleme nachzudenken.	1	3	2
015	W3	Ich finde EK als Schulfach ziemlich überflüssig.	1	③	2
018	W4	Die meisten Schulfächer sind wichtiger als EK.	1	③	2
022	W5	EK ist zwar recht interessant, aber im Vergleich zu anderen Fächern ziemlich unwichtig.	1	③	2
027	W6	Einer der in EK viel weiß, kommt später im Leben besser voran	1	3	2
028	W7	Das meiste, was wir in EK lernen, vergißt man nach den Prüfungen doch gleich wieder.	1	③	2
029	W8	Erdkundliches Wissen braucht man bei vielen Gelegenheiten im Leben.	1	3	2
011	S1	Der Unterrichtsstoff in EK ist leichter als in den meisten anderen Fächern.	1	③	2
012	S2	Um in EK mitzukommen, muß man sich anstrengen.	1	3	2
016	S3	Manches, was in EK behandelt wird, ist schwierig zu verstehen.	1	3	2
019	S4	EK ist ein reines Lernfach.	1	③	2
024	S5	Ek ist ein Fach, in dem man sich leicht eine gute Note holen kann.	1	③	2
025	S6	Wenn ich in EK mal nicht aufpasse, finde ich nur schwer wieder Anschluß. (Die Items mit umrandeten Kodes wurden umgepolt.)	1	3	2

Tab. 5: Die Einstellungsbatterie des RCFP zum Fach Erdkunde allgemein (Die Nummern vor den Items sind die Variablennummern im Gesamtdatensatz SCHÜLER bzw. in den Datensätzen des RCFP. Die Ziffern nach den Items sind die Kodierungen)

(1) Der Schwierigkeitsindex „S":

Der *Schwierigkeitsindex* der Items ist einfach der Prozentsatz an sinngemäß „richtigen" Antworten zu einem Item, dividiert durch 100. Bei Item I1 („Andere Unterrichtsfächer interessieren mich mehr als gerade Erdkunde") beispielweise wäre die sinngemäß „richtige" Antwort eine *Ablehnung*, und der Schwierigkeitsindex wäre der Anteil der Schüler, die dieses Item ablehnen. Bei Item I2 („Ich interessiere mich ganz allgemein für Erdkunde") wäre der Schwierigkeitsindex demnach entsprechend der Anteil der Schüler, die diesem Item *zustimmen*. Der Begriff „Schwierigkeit" mag in diesem Zusammenhang merkwürdig erscheinen, da es um Einstellungsitems und nicht um Wissensfragen geht. Der Begriff ist jedoch ein feststehender „terminus technicus" der Itemanalyse.

Interessant für die Itemanalyse sind nur extreme „Ausreißer" bei den Schwierigkeitsindizes: Wenn alle oder extrem viele Versuchspersonen einem Item zustimmen, dann ist das Item entweder zu positiv formuliert oder die gemessene Einstellung ist so allgemein verbreitet, daß es keinen Sinn hat, sie überhaupt zu messen. Sinngemäß gilt das gleiche bei extrem niedrigen Schwierigkeitsindizes. Indizes von mehr als 0,9 und weniger als 0,1 sollen bei uns als extrem gelten. Der Schwierigkeitsindex sollte also möglichst im Mittelbereich liegen. Extrem „leichte" oder extrem „schwere" Items sind, item-analytisch gesehen, unbrauchbar.

(2) Die Trennschärfe „T_{korr}" und „$T_{korr/norm}$":

Die Trennschärfe der Items ist nach Lienert gleich „dem Korrelationskoeffizienten zwischen Aufgabenantwort – der richtigen oder falschen – und dem Rohwert" (*Lienert* 1969, S. 93). In unserem Fall bedeutet Rohwert gleich dem Summenwert (oder Summen-Score) auf der jeweiligen „theoretisch" intendierten Subskala. Der Trennschärfeindex gibt an, wie gut ein Item im Sinn der zugehörigen Subskala die Versuchspersonen trennt. Ein Item mit hoher Trennschärfe mißt das Gleiche wie die zugehörige Skala. Bei der Interesse-Skala z. B. wird ein bestimmtes Item mit *hoher* Trennschärfe von „interessierten" Schülern überwiegend bejaht, von „desinterissierten" Schülern dagegen überwiegend verneint.

Um dieses Problem nochmals von einer etwas anderen Seite zu beleuchten, muß man sich die Beziehung zwischen Schwierigkeit und Trennschärfe klar machen: Eine Fragestellung, die fast alle Schüler gleich beantworten, muß automatisch eine geringere Trennschärfe aufweisen. Sind nämlich alle anderen Faktoren gleich, erreicht ein Item optimale Trennschärfe bei einem Schwierigkeitsindex von 0,5. Man kann sich diesen Zusammenhang verdeutlichen, wenn man sich vorstellt, daß ein Item zu 100% identisch beantwortet wurde (S = 1,00). Dann kann dieses Item natürlich nicht mehr zwischen jenen Schülern trennen, die einen hohen oder einen niedrigen „Wert" auf der zugehörigen Gesamtskala haben. Seine Trennschärfe wäre infolgedessen gleich Null.

Zur rechnerischen Ermittlung der Trennschärfe benutzte man früher den sog. „punktbiserialen Korrelationskoeffizienten", da dieser „per Hand" leichter zu berechnen ist als der heute allgemein übliche „Produkt-Moment-Koeffizient". Beide Koeffizienten sind aber algebraisch identisch, weswegen der punkbiseriale Korrelationskoeffizient im Zeitalter der elektronischen Datenverarbeitung nur noch historische Bedeutung hat.

Möglich wäre es allerdings, die Trennschärfe durch den sog. „biserialen Korrelationskoeffizienten" zu erfassen. Er berücksichtigt das unterschiedliche Meßniveau von Item und Skala besser. Allerdings ist er nicht auf ± 1 normiert, weswegen die meisten Autoren von seiner Verwendung abraten.

Ein weiteres Problem ist die Korrektur des Korrelationskoeffizienten. Verwendet man die normale Produkt-Moment-Korrelation, erhält man immer einen etwas zu hohen Trennschärfekoeffizienten. Der Grund ist folgender: Bei der Berechnung der Summenscores pro Subskala werden normalerweise *alle* Items berücksichtigt. Berechnet man die Korrelation zwischen einem einzelnen Item und diesem Summenscore, so korreliert man das Item mit einem Wert, zu dem es selbst beigetragen hat. Man bezeichnet diesen Sachverhalt als „part-whole-correlation". Die Gesamtkorrelation enthält also einen kleinen Anteil an Scheinkorrelation, weil man das Item zu einen kleinen Teil gleichsam *mit sich selbst* korreliert.

Dieser kleine Fehler läßt sich ganz einfach korrigieren: Bei der Auszählung der Summenscores muß jeweils immer gerade jenes Item unberücksichtigt bleiben, für das gerade die Trennschärfe berechnet wird.

Ein letztes Problem ist die Abhängigkeit der Trennschärfe von der Schwierigkeit des Items. Besonders wenn die Schwierigkeiten stärker von 0,5 abweichen, sollte dies bei der Berechnung der Trennschärfe berücksichtigt werden. Dies geschieht nach der von Lienert empfohlenen Formel (*Lienert* 1969, S. 153), die den üblichen Trennschärfekoeffizienten an der maximal möglichen Trennschärfe je Schwierigkeitsstufe des Items normiert:

$$T_{korr/norm} = \frac{T_{korr}}{2 \cdot s \cdot (1-s)}$$

$T_{korr/nor}$: korrigierter und normierter Trennschärfekoeffizient

s: Standardabweichung

(3) Das Bestimmtheitsmaß "R^2":

Die *quadrierte multiple Item-Korrelation* wird auch als sog. Bestimmtheitsmaß bezeichnet. Sie gibt unmittelbar an, welcher Anteil an der Varianz eines Items durch die anderen Items der betreffenden Skala erklärt werden kann. Oder anders ausgedrückt: der R^2-Wert (mal Hundert) ist der Prozentsatz an gemeinsamer Varianz (gemeinsam mit der Skala) pro Item.

(4) Die Reliabilität „α":

Der Reliabilitätskoeffizient (*Cronbachs* α) ist ein Reliabilitätsmaß, das auf der Varianzanalyse beruht. Er vergleicht die Varianz der Items mit der Varianz der Gesamtskala:

Bei SPSS (vergl. Nie u. a. 1975) wird er nach folgender Formel berechnet:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\frac{s_i^2}{k}}{S_T^2}\right) = \frac{k}{k-1}\left(1 - \frac{\bar{s}_i^2}{S_T^2}\right)$$

wobei: k: Anzahl der Items

$\frac{s_i^2}{k} = \bar{s}^2$: durchschnittliche Item-Varianz

S_T^2: Varianz der Skala, dividiert durch Anzahl der Items, also: durchschnittliche Skalen-Varianz

Der α-Koeffizient ist ein Maß für die Konsistenz oder Homogenität der Skala (vergl.: *Novick/Lewis* 1967).

In unseren Tabellen 6 bis 8 wurde zunächst ein α-Wert für die ganze Sub-Skala berechnet. Zusätzlich wurde bei jeder Skala nacheinander jeweils ein Item entfernt und für die restlichen ein α-Wert berechnet. Dadurch ist aus den Tabellen zu entnehmen, wie die Reliabilität der Skala ausfallen würde, wenn das jeweilige Item *entfernt* werden würde. Steigt der α-Wert, wenn man ein bestimmtes Item eliminiert, ist dieses Item schlecht. Fällt der α-Wert, dann war die Eliminierung dieses Items ein Fehler, und man sollte es besser in der Skala lassen.

Die insgesamte Reliabilität (α (Skala)) bezieht sich auf die Skala, wenn *keine* Items ausgeschlossen werden.

Die oben beschriebenen Itemkennwerte wurden nun für die Einstellungsbatterie zum Fach Erdkunde allgemein berechnet, und zwar jeweils für die drei ursprünglichen Subskalen, wie sie vom RCFP konzipiert wurden.

5.1.1 Die Itemkennwerte für die Subskala „Interesse"

Die Itemwerte der Skala ergeben folgendes Bild:

Die Schwierigkeitsindizes der Interesse-Items bewegen sich im allgemeinen, wie es erwünscht ist, im mittleren Bereich (0,5 bis 0,6). Nur das Item I8 („Erdkunde ist mein Lieblingsfach") ist zu „schwer", d. h. nur sehr wenige Schüler konnten ihm – im Sinn der Interesse-Skala – zustimmen.

Subskala: INTERESSE ($\alpha_{(gesamt)}$ = 0,82)

			S	T_{korr}	$T_{korr/norm}$	R^2	α
I1	8	Andere Unterrichtsfächer interessieren mich mehr als gerade Erdkunde.	0,29	0,49	0,54	0,32	0,80
I2	9	Ich interessiere mich ganz allgemein für Erdkunde.	0,57	0,41	(0,41)	(0,19)	(0,81)
I3	13	Der Unterrichtsstoff in Erdkunde ist eigentlich ziemlich trocken.	0,59	0,58	0,59	0,40	0,79
I4	17	Der Erdkundeunterricht hat mich angeregt, über einige seiner Themen weiter nachzudenken.	0,59	0,41	(0,42)	(0,18)	(0,81)
I5	20	Der Erdkundeunterricht macht mir Spaß.	0,57	0,74	0,75	0,64	0,76
I6	21	Erdkunde langweilt mich häufig.	0,62	0,67	0,69	0,58	0,77
I7	23	Wenn ich die Unterrichtsfächer frei wählen könnte, würde ich Erdkunde nicht nehmen.	0,56	0,59	0,59	0,37	0,79
I8	26	Erdkunde ist mein Lieblingsfach.	0,11	0,39	0,62	(0,24)	(0,81)
		Durchschnitt	0,49		0,58		

Tab. 6: Die Itemkennwerte für die Subskala „Interesse"

Die korrigierten und normierten Trennschärfeindizes der Items sind insgesamt gut bis sehr gut. Am deutlichsten trennt Item I5 („Der Erdkundeunterricht macht mir Spaß") zwischen interessierten und nicht interessierten Schülern. Am wenigsten trennt Item I2 („Ich interessiere mich ganz allgemein für Erdkunde"). Item I2 ist offensichtlich zu unspezifisch formuliert: Auch einige an sich desinteressierte Schüler konnten ihm zustimmen.

Die Items der Interesse-Skala haben sehr unterschiedliche gemeinsame Varianz: Bei Item I4 („Der Erdkundeunterricht hat mich angeregt über einige seiner Themen weiter nachzudenken") beträgt sie nur 18%, d. h. nur 18% der Varianz dieses Items kann aus den anderen Items der Skala erklärt werden. Dieses Item mißt offenbar einen sehr spezifischen Aspekt im Interesse der Schüler. Item I5 („Der Erdkundeunterricht macht mir Spaß.") hat die größte gemeinsame Varianz und wird zu 64% aus den anderen Interesse-Items erklärt. Insgesamt gesehen kann man aus den R^2-Werten ersehen, daß die Skala (mit Ausnahme von I4 und I2) relativ viel gemeinsame Varianz besitzt. Die Items messen also relativ ähnliche Sachverhalte.

Die Homogenitätsanalyse anhand der α-Werte ergibt fast das gleiche Resultat wie die Analyse der R^2-Werte. Schlechte Items sind: I2, I4 und I8. Item I8 hat jedoch eine relativ gute Trennschärfe, wenn man seine hohe „Schwierigkeit" berücksichtigt. Den besten Wert liefert I5. Insgesamt gesehen ist die Skala „Interesse" eine sehr homogene Skala (α = 0,82).

5.1.2 Die Itemkennwerte der Subskala „Wichtigkeit"

Die Schwierigkeitsindizes für die Items der Wichtigkeits-Skala schwanken zwischen 0,24 für Item W4 („Die meisten Schulfächer sind wichtiger als Erdkunde") und 0,89 für Item W3 („Ich finde Erdkunde als Schulfach ziemlich überflüssig"). Item W4 ist ziemlich schwer, d. h. viele Schüler mußten (!) zustimmen, daß die meisten Schulfächer wichtiger als Erdkunde sind. (Eine „richtige" Antwort im Sinn der Skala wäre – wegen der umgekehrten Polung des Items – ein *Ablehnung* des Items.) Insgesamt gesehen sind aber die Schwierigkeitsindizes in Ordnung. (siehe Tab.)

Die Trennschärfe der Items variiert zwischen 0,3 für Item W2 („Der Erdkundeunterricht regt mich an, über gesellschaftliche Probleme nachzudenken") und 0,53 für Item W3 („Ich finde Erdkunde als Schulfach ziemlich überflüssig"). Neben dem Item W2 gibt es noch zwei weitere Items die eindeutig unbefriedigende Trennschärfe ausweisen: Es handelt sich um W6 („Einer der in Erdkunde viel weiß, kommt später im Leben besser voran") und um W7 („Das meiste, was wir in Erdkunde lernen, vergißt man nach den Prüfungen doch gleich wieder"). Als Grund für die mangelhafte Tennschärfe der drei Items W2, W6 und W7 läßt sich bereits jetzt vermuten, daß nicht eine unspezifische Formulierung dafür verantwortlich ist, sondern vermutlich eher die Heterogenität der angesprochenen Sachverhalte.

Subskala: WICHTIGKEIT ($\alpha_{(gesamt)}$ = 0,65)

			S	T_{korr}	$T_{korr/norm}$	R^2	α
W1	10	Das was ich im Erdkundeunterricht erfahre, kann ich auch außerhalb der Schule gut gebrauchen.	0,76	0,38	0,44	0,20	0,61
W2	14	Der Erdkundeunterricht regt mich an, über gesellschaftliche Probleme nachzudenken.	0,45	0,30	(0,30)	(0,09)	(0,63)
W3	15	Ich finde Erdkunde als Schulfach ziemlich überflüssig.	0,89	0,33	0,53	(0,14)	(0,63)
W4	18	Die meisten Schulfächer sind wichtiger als Erdkunde.	0,24	0,33	0,39	0,17	0,62
W5	22	Erdkunde ist zwar recht interessant, aber im Vergleich zu anderen Fächern ziemlich unwichtig.	0,56	0,39	0,39	0,21	0,61
W6	27	Einer, der in Erdkunde viel weiß, kommt später im Leben besser voran.	0,26	0,31	(0,35)	(0,11)	(0,63)
W7	28	Das meiste, was wir in Erdkunde lernen, vergißt man nach den Prüfungen doch gleich wieder.	0,57	0,36	(0,36)	(0,14)	0,61
W8	29	Erdkundliches Wissen braucht man bei vielen Gelegenheiten im Leben.	0,74	0,42	0,48	0,23	0,60
		Durchschnitt	0,56		0,41		

Tab. 7: Die Itemkennwerte für die Subskala „Wichtigkeit"

Diese Heterogenität zeigt sich deutlich, wenn man die gemeinsame Varianz (R^2) der Items betrachtet: Bei keinem Item wird mehr als 23 % seiner Varianz durch die anderen Items der Skala erklärt. Jedes Item mißt also zu mindestens 75 % seiner Varianz Schülereinstellungen, die nichts mit den restlichen Aspekten der Einstellungsskala zu tun haben. Im Vergleich mit der Interessen-Skala ist also die Wichtigkeits-Skala eindeutig zu heterogen. Zu viele verschiedene Aspekte wurden vermischt: Die Bedeutung als Schulfach, die Verwertbarkeit von Erdkunde im späteren Leben, der Einfluß des Faches auf die politische Bewußtseinsbildung (W2) und die Einprägsamkeit erdkundlichen Wissens (W7).

Die Homogenitätsanalyse bestätigt auch hier voll diese Ergebnisse: Zunächst kann man feststellen, daß die Homogenität der Gesamt-Skala mit einem α-Wert von 0,65 deutlich unter der Homogenität der Interesse-Skala liegt. Untersucht man, welche Items dabei besonders negativ auffallen, dann zeigt sich, daß W2 und W6 die Homogenität der Wichtigkeits-Skala besonders drücken – also jene Items, die uns schon wegen ihrer geringen Trennschärfe und ihrer niedrigen gemeinsamen Varianz aufgefallen waren. Auch W3 („Ich finde Erdkunde als Schulfach ziemlich überflüssig") verschlechtert die Homogenität der Skala – vermutlich wegen seiner geringen Schwierigeit von 0,89 (bei relativ hoher Trennschärfe von 0,56).

5.1.3 Die Itemkennwerte für die Subskala „Schwierigkeit"

Subskala: SCHWIERIGKEIT ($_{(gesamt)}$ = 0,71)

			S	T_{korr}	$T_{korr/norm}$	R^2	
S1	11	Der Unterrichtsstoff in Erdkunde ist leichter als in den meisten anderen Fächern.	0,46	0,51	0,51	0,28	0,65
S2	12	Um in Erdkunde mitzukommen, muß man sich schon anstrengen.	0,37	0,54	0,56	0,30	0,64
S3	16	Manches was in Erdkunde behandelt wird, ist schwierig zu verstehen.	0,56	0,38	(0,38)	(0,17)	(0,69)
S4	19	Erdkunde ist ein reines Lernfach ohne große Schwierigkeiten.	0,50	0,39	(0,39)	(0,18)	(0,69)
S5	24	Erdkunde ist ein Fach, in dem man sich leicht eine gute Note holen kann.	0,40	0,46	0,47	0,23	0,67
S6	25	Wenn ich in Erdkunde mal nicht aufpasse, finde ich nur schwer wieder Anschluß.	0,30	0,39	0,43	(0,18)	(0,68)
		Durchschnitt			0,46		

Tab. 8: Die Itemkennwerte für die Subskala „Schwierigkeit"

Die Schwierigkeitsindizes der Schwierigkeits-Skala liegen sehr schön im Mittelbereich zwischen 0,30 und 0,56. Man kann deshalb feststellen, daß bei dieser Subskala die Schwierigkeitskennwerte fast optimal sind.

Die Trennschärfe der Items ist auch befriedigend. Relativ hoch ist sie bei Item S2 („Um in Erdkunde mitzukommen, muß man sich schon anstrengen."). Die Items S3 und S4 fallen dagegen in ihrer Trennschärfe deutlich ab.

Ein befriedigendes Gesamtbild ergibt sich auch bei der Analyse von R^2 und α. Das heißt: Die Skala „Schwierigkeit" hat zwar auch nur eine gemeinsame Varianz von etwas über 20 %. Die Multiple Korrelation der Items beträgt im Durchschnitt 0,22. Trotzdem ergibt sich insgesamt ein durchaus befriedigender α-Wert von 0,71. Damit ist die Homogenität der Skala „Schwierigkeit" wesentlich besser als die der „Wichtigkeits-Skala, erreicht aber nicht die Qualität der „Interessen"-Skala.

5.1.4 Ergebnisse der klassischen Kennwerteanalyse

Auf dem Hintergrund der klassischen Testtheorie läßt sich in Bezug auf die Subskalen der Einstellungsbatterie zum Erdkundeunterricht folgendes sagen:

- Die verschiedenen Itemkennwerte ergeben bei allen Subskalen ein in sich konsistentes Bild; d. h. sie stützen sich fast durchweg gegenseitig und erhöhen damit die Gültigkeit der Ergebnisse:
- Die Qualität der Subskalen insgesamt ist deutlich verschieden. Die beste Qualität besitzt die „Interesse"-Skala, die schlechteste die „Wichtigkeits"-Skala. Die „Schwierigkeits"-Skala hat mittlere Qualität.
- Die Kennwerte identifizieren ziemlich eindeutig folgende „schlechte" Items: I2, I4 bei der „Interesse"-Skala; W2, W3, W6 und W7 bei der „Wichtigkeits"-Skala; S3, S4, S6 bei der „Schwierigkeits"-Skala. Das Items I2 („Ich interessiere mich ganz allgemein für Erdkunde") ist ganz offensichtlich zu schwammig formuliert. Das Item I4 („Der Erdkundeunterricht hat mich angeregt, über einige seiner Themen weiter nachzudenken") ist zwar inhaltlich klar formuliert, nur hat es nicht nur mit „Interesse" zu tun. Es paßt eher in die Wichtigkeits-Subskala, wo bereits ähnlich formulierte Items enthalten sind. Die Items W2, W3, W6 und W7 (vergl. Tabelle 7) sind ebenfalls zwar inhaltlich eindeutig, aber zu heterogen in ihrer Bedeutung.
  S3 („Manches, was in Erdkunde behandelt wird, ist schwierig zu verstehen") ist dagegen eindeutig schlecht formuliert: das kritische Wort ist „manches". Es läßt zu viel Deutungsspielraum. Das Items S6 („Wenn ich in Erdkunde mal nicht aufpasse, finde ich nur schwer wieder Anschluß") dürfte vermutlich deshalb schlecht abschneiden, weil „schwer

Anschluß finden" etwas mit „schlechten Noten" und „durchfallen" zu tun hat. D. h. es ist im Bewußtsein der Schüler negativ vorbesetzt. Der Schüler, der hier zustimmt, muß sich gewissermaßen selbst für ziemlich dumm erklären.

- Aus der Item-Analyse ergibt sich inhaltlich, daß nur die Items der Interessen-Skala wirklich in etwa das gleiche erfassen. Die Items der Wichtigkeitsskala sind – wie sich empirisch erwiesen hat – im Grunde lediglich Einzelfragen, bzw. Statements. Die Intention des RCFP-Forschungsstabes, nämlich entweder nur „Schwierigkeits"- oder „Wichtigkeits"-Einschätzungen der Schüler zu erheben, kann durch diese Subskala nicht realisiert werden.

- Der Grund für die ungenügende Qualität der Subskala „Wichtigkeit" kann im folgenden liegen:
  Entweder sind die Items tatsächlich alle zu heterogen formuliert. Dann wäre eine Bildung einer Subskala prinzipiell unmöglich. Oder die Items sind nur falsch zu einer Subskala zusammengefaßt.
  Letzteres würde bedeuten, daß die Items zwar wenige zugrundeliegende Dimensionen oder Grundstrukturen aufweisen, diese Dimensionen aber nicht mit der geplanten Subskala zusammenfallen.
  Ob überhaupt wenige Grunddimensionen durch die Items aufgespannt werden, und wie diese Dimensionen liegen (z. B. quer zu den gedachten Subskalen), kann durch die Analyse der Itemkennwerte nicht beantwortet werden.
  Dazu ist eine Hauptkomponentenanalyse der Items nötig.

5.2 Dimensionsanalyse (Faktorenanalyse) der Einstellungsbatterie zum Fach Erdkunde allgemein

Die oben durchgeführte Itemanalyse anhand einiger Kennwerte der klassischen Testtheorie hat einen wesentlichen Nachteil: Sie unterstellt, daß die Items einer Subskala auf *einer* Meßdimension liegen.

Eindimensionalität der Subskalen ist jedoch eine Eigenschaft, die man nicht ohne weiteres voraussetzen darf. Nur weil für den Forscher bestimmte Items inhaltlich zusammengehören, d. h. auf einer Dimension liegen, bedeutet dies noch lange nicht, daß auch die Versuchspersonen es so sehen. Um es vorwegzunehmen: In unserem Fall hängen, empirisch gesehen, die Items zum Teil anders zusammen als theoretisch gedacht, d. h. es ergeben sich Dimensionen, die teilweise quer zu den geplanten Subskalen liegen.

5.2.1 Die Dimensionalität der ursprünglich vom RCFP theoretisch angenommenen Subskalen der Batterie

In einem ersten Schritt wurde durch Hauptkomponentenanalysen[1] überprüft, ob die Subskalen „Interesse", „Wichtigkeit" und „Schwierigkeit" jeweils eindimensional sind (siehe Tabelle 9).

Die Hauptkomponentenanalyse der *„Interessen"-Skala* ergab *einen* signifikanten Faktor (Eigenwert 2>1). Er erklärt ca. 62 % der Gesamtvarianz. Die Skala „Interesse" enthält also knapp 40 % „fremde" Varianz, d. h. Varianz, die nicht aus der einen intendierten Meßdimension herrührt.

Man kann sagen, daß die Subskala „Interesse" einigermaßen eindimensional mißt, obwohl die Fehlervarianz beträchtlich ist. In Bezug auf die einzelnen Items bestätigen sich die Ergebnisse der Itemanalyse vollkommen: Die beiden Items mit den schwächsten Kennwerten – nämlich I2 und I4 – haben auch bei der Hauptkomponentenanalyse die niedrigste Varianzaufklärung (Kommunalitäten von 0,34 bzw. 0,37). Damit ist nachgewiesen, daß die Ursache für die schlechte Qualität dieser Items darin liegt, daß sie nicht optimal auf der einen Hauptmeßdimension liegen.

Bei der *„Wichtigkeits"-Skala* ergibt sich durch die Hauptkomponentenanalyse ein wesentlich komplexeres Bild.

Die Skala ist eindeutig *zwei*dimensional. Die beiden signifikanten Faktoren erklären hier aber nur genausoviel Varianz wie der eine (!) Faktor bei der Interessen-Skala, nämlich 62,5 % (wobei beide Skalen gleich viele Items haben). Um es vereinfacht zu sagen: Die Skala mißt (als „Skala") nicht nur nicht besonders gut, sondern sie mißt auch zwei verschiedene Dinge.

Die Faktorisierung der *„Schwierigkeits"-Skala* erbringt dagegen wieder ein recht einfaches und klares Ergebnis:

Sie ist zwar eine eindimensionale Skala; d. h. es ergibt sich nur *ein* signifikanter Eigenwert, der 53,7 % der Varianz aufklärt; aber man muß sich vergegenwärtigen, daß die Skala nur 6 Items hat (die beiden anderen hatten je 8 Items), und daß beinahe die Hälfte der Varianz unerklärt bleibt, wenn man sie als eindimensionale Meßskala verwendet. Nur die Interessenskala kann man also guten Gewissens als eindimensional betrachten. Die Schwierigkeitsskala liefert zwar auch nur einen Faktor, dieser erklärt jedoch relativ wenig, so daß es fraglich ist, ob die Skala nach dieser Meßdimension charakterisiert werden kann. Die Wichtigkeitsskala ist zweifellos zweidimensional.

Die Analysen der jeweiligen Subskalen legen den Verdacht nahe, daß sich die Items der Gesamtbatterie weit besser zusammenfassen lassen als dies mit den drei Subskalen geschehen ist. Diese Aufdifferenzierung der Items in optimal eindimensionale Subskalen geschieht durch eine Faktorisierung der *Gesamtbatterie*.

INTERESSE			WICHTIGKEIT			SCHWIERIGKEIT		
Eigenwerte/Varianzaufklärung::			Eigenwerte/Varianzaufklärung:			Eigenwerte/Varianzaufklärung:		
Faktor	Eigenwerte	Varianzaufkl.	Faktor	Eigenwerte	Varianzaufkl.	Faktor	Eigenwerte	Varianzaufkl.
1	4,99	62,4 %	1	3,90	48,7 %	1	3,22	53,7 %
2	0,81	10,1 %	2	1,10	13,8 %	2	0,97	16,3 %
3	0,80	10,0 %	3	0,77	9,6 %	3	0,55	9,2 %
4	0,57	7,2 %	4	0,69	8,7 %	4	0,50	8,3 %
5	0,35	4,5 %	5	0,54	6,8 %	5	0,45	7,4 %
6	0,27	3,3 %	6	0,42	5,3 %	6	0,30	5,0 %
7	0,17	2,1 %	7	0,31	3,9 %			
8	0,04	0,4 %	8	0,26	3,3 %			
VAR	Ladungen auf Faktor 1:		VAR	Ladungen auf den 2 signifikanten (rotierten) Faktoren		VAR	Ladungen auf Faktor 1:	
				Faktor 1:	Faktor 2:			
008	0,79		010	0,79*	0,15	011	0,80	
009	0,59		014	0,45	0,35	012	0,84	
013	0,79		015	0,55	0,57	016	0,67	
017	0,61		018	0,07	0,90*	019	0,66	
020	0,97		022	0,27	0,83*	024	0,74	
021	0,88		027	0,73*	0,10	025	0,68	
023	0,83		028	0,59	0,37			
026	0,80		029	0,86*	0,21			
VAR	endgültige Kommunalitäten		VAR	endgültige Kommunalitäten		VAR	endgültige Kommunalitäten	
008	0,62		010	0,65		011	0,64	
009	0,34		014	0,34		012	0,70	
013	0,62		015	0,63		016	0,45	
017	0,37		018	0,81		019	0,44	
020	0,93		022	0,77		024	0,54	
021	0,78		027	0,54		025	0,47	
023	0,68		028	0,49				
026	0,64		029	0,78				

Tab. 9: Ergebnisse der Faktorenanalysen zu den ursprünglichen Subskalen „Interesse", „Wichtigkeit" und „Schwierigkeit"

5.2.1 Die Dimensionalität der ursprünglich vom RCFP theoretisch angenommenen Subskalen der Batterie

In einem ersten Schritt wurde durch Hauptkomponentenanalysen[1] überprüft, ob die Subskalen „Interesse", „Wichtigkeit" und „Schwierigkeit" jeweils eindimensional sind (siehe Tabelle 9).

Die Hauptkomponentenanalyse der *„Interessen"-Skala* ergab *einen* signifikanten Faktor (Eigenwert 2>1). Er erklärt ca. 62 % der Gesamtvarianz. Die Skala „Interesse" enthält also knapp 40 % „fremde" Varianz, d. h. Varianz, die nicht aus der einen intendierten Meßdimension herrührt.

Man kann sagen, daß die Subskala „Interesse" einigermaßen eindimensional mißt, obwohl die Fehlervarianz beträchtlich ist. In Bezug auf die einzelnen Items bestätigen sich die Ergebnisse der Itemanalyse vollkommen: Die beiden Items mit den schwächsten Kennwerten – nämlich I2 und I4 – haben auch bei der Hauptkomponentenanalyse die niedrigste Varianzaufklärung (Kommunalitäten von 0,34 bzw. 0,37). Damit ist nachgewiesen, daß die Ursache für die schlechte Qualität dieser Items darin liegt, daß sie nicht optimal auf der einen Hauptmeßdimension liegen.

Bei der *„Wichtigkeits"-Skala* ergibt sich durch die Hauptkomponentenanalyse ein wesentlich komplexeres Bild.

Die Skala ist eindeutig *zwei*dimensional. Die beiden signifikanten Faktoren erklären hier aber nur genausoviel Varianz wie der eine (!) Faktor bei der Interessen-Skala, nämlich 62,5 % (wobei beide Skalen gleich viele Items haben). Um es vereinfacht zu sagen: Die Skala mißt (als „Skala") nicht nur nicht besonders gut, sondern sie mißt auch zwei verschiedene Dinge.

Die Faktorisierung der *„Schwierigkeits"-Skala* erbringt dagegen wieder ein recht einfaches und klares Ergebnis:

Sie ist zwar eine eindimensionale Skala; d. h. es ergibt sich nur *ein* signifikanter Eigenwert, der 53,7 % der Varianz aufklärt; aber man muß sich vergegenwärtigen, daß die Skala nur 6 Items hat (die beiden anderen hatten je 8 Items), und daß beinahe die Hälfte der Varianz unerklärt bleibt, wenn man sie als eindimensionale Meßskala verwendet. Nur die Interessenskala kann man also guten Gewissens als eindimensional betrachten. Die Schwierigkeitsskala liefert zwar auch nur einen Faktor, dieser erklärt jedoch relativ wenig, so daß es fraglich ist, ob die Skala nach dieser Meßdimension charakterisiert werden kann. Die Wichtigkeitsskala ist zweifellos zweidimensional.

Die Analysen der jeweiligen Subskalen legen den Verdacht nahe, daß sich die Items der Gesamtbatterie weit besser zusammenfassen lassen als dies mit den drei Subskalen geschehen ist. Diese Aufdifferenzierung der Items in optimal eindimensionale Subskalen geschieht durch eine Faktorisierung der *Gesamtbatterie*.

INTERESSE			WICHTIGKEIT			SCHWIERIGKEIT		
Eigenwerte/Varianzaufklärung::			Eigenwerte/Varianzaufklärung:			Eigenwerte/Varianzaufklärung:		
Faktor	Eigenwerte	Varianzaufkl.	Faktor	Eigenwerte	Varianzaufkl.	Faktor	Eigenwerte	Varianzaufkl.
1	4,99	62,4 %	1	3,90	48,7 %	1	3,22	53,7 %
2	0,81	10,1 %	2	1,10	13,8 %	2	0,97	16,3 %
3	0,80	10,0 %	3	0,77	9,6 %	3	0,55	9,2 %
4	0,57	7,2 %	4	0,69	8,7 %	4	0,50	8,3 %
5	0,35	4,5 %	5	0,54	6,8 %	5	0,45	7,4 %
6	0,27	3,3 %	6	0,42	5,3 %	6	0,30	5,0 %
7	0,17	2,1 %	7	0,31	3,9 %			
8	0,04	0,4 %	8	0,26	3,3 %			
VAR	Ladungen auf Faktor 1:		VAR	Ladungen auf den 2 signifikanten (rotierten) Faktoren		VAR	Ladungen auf Faktor 1:	
				Faktor 1:	Faktor 2:			
008	0,79		010	0,79*	0,15	011	0,80	
009	0,59		014	0,45	0,35	012	0,84	
013	0,79		015	0,55	0,57	016	0,67	
017	0,61		018	0,07	0,90*	019	0,66	
020	0,97		022	0,27	0,83*	024	0,74	
021	0,88		027	0,73*	0,10	025	0,68	
023	0,83		028	0,59	0,37			
026	0,80		029	0,86*	0,21			
VAR	endgültige Kommunalitäten		VAR	endgültige Kommunalitäten		VAR	endgültige Kommunalitäten	
008	0,62		010	0,65		011	0,64	
009	0,34		014	0,34		012	0,70	
013	0,62		015	0,63		016	0,45	
017	0,37		018	0,81		019	0,44	
020	0,93		022	0,77		024	0,54	
021	0,78		027	0,54		025	0,47	
023	0,68		028	0,49				
026	0,64		029	0,78				

Tab. 9: Ergebnisse der Faktorenanalysen zu den ursprünglichen Subskalen „Interesse", „Wichtigkeit" und „Schwierigkeit"

5.2.2 Die Dimensionalität der Gesamtbatterie zum Fach Erdkunde allgemein

Nicht auf drei, sondern mindestens auf fünf Dimensionen liegen die 22 Items der Einstellungsbatterie zum Erdkundeunterricht. Fünf signifikante Faktoren (nach dem Eigenwertkriterium) erklären 68% der Gesamtvarianz der Statementbatterie. Die wichtigsten drei Hauptkomponenten erklären nur 57%. Mit anderen Worten: Würde man auf theoretisch intendierten *drei* Meßdimensionen bestehen, könnte man mehr als 40% der Varianz der Items nicht erklären. Ein Meßinstrument, daß zu gut 40% seiner Varianz nicht das mißt, was man eigentlich messen möchte, kann nicht als besonders gut eingestuft werden.

Auch bei dieser 5-Faktoren-Lösung fielen – wie schon bei den Reliabilitätsanalysen – die Items I2 und W7 als besonders schwach auf (niedrige Kommunalitäten). Sie wurden bei einer zweiten Runde der Faktorenanalyse ausgeschlossen.
Damit ergaben sich folgende fünf Dimensionen:

Dimension 1: INTERESSE

VAR			
008	I1	Andere Unterrichtsfächer interessieren mich mehr als gerade Erdkunde.	0,83
013	I3	Der Unterrichtsstoff in Erdkunde ist eigentlich ziemlich trocken.	0,70
020	*I5	Der Erdkundeunterricht macht mir Spaß.	0,84
021	I6	Erdkunde langweilt mich häufig.	0,81
023	I7	Wenn ich die Unterrichtsfächer frei wählen könnte, würde ich Erdkunde nicht nehmen.	0,67
026	*I8	Erdkunde ist mein Lieblingsfach.	0,82

*wurde umgepolt

Tab. 10: Dimension INTERESSE aus der Faktorenanalyse der Gesamtbatterie zum Fach Erdkunde allgemein

Diese erste Dimension entspricht der ursprünglichen Interessen-Skala des RCFP – mit Ausnahme des ausgeschlossenen Items 12 und des Items I4 („Der Unterricht hat mich angeregt, über einige seiner Themen weiter nachzudenken"). Item I4 hat, inhaltlich gesehen, tatsächlich nicht unbedingt etwas mit bloßem Interesse zu tun. Wir werden sehen, daß dieses Item erwartungsgemäß der Dimension „ANREGUNG" zugeordnet wird.

Die Qualität dieses ersten Faktors der Hauptkomponentenanalyse ist hervorragend: Die Ladungen liegen im Durchschnitt bei etwa 0,8, was einer Varianzaufklärung von 64% entspricht. Auch inhaltlich gesehen erscheint diese Meßdimension als in sich konsistent und aussagekräftig.

Dimension 2: SCHWIERIGKEIT

VAR	*wurde umgepolt	
011	S1 Der Unterrichtsstoff in Erdkunde ist leichter als in den meisten anderen Fächern.	0,77
012	*S2 Um in Erdkunde mitzukommen muß man sich anstrengen.	0,82
016	*S3 Manches, was in Erdkunde behandelt wird, ist Schwierig zu verstehen.	0,64
019	S4 Erdkunde ist ein reines Lernfach.	0,69
024	S5 Erdkunde ist ein Fach, in dem man sich leicht eine gute Note holen kann.	0,73
025	*S6 Wenn ich in Erdkunde mal nicht aufpasse, finde ich nur schwer wieder Anschluß.	0,66

Tab. 11: Dimension SCHWIERIGKEIT aus der Faktorenanalyse der Gesamtbatterie

Die ursprüngliche Schwierigkeitsskala kann perfekt durch die Dimensionsanalyse reproduziert werden. Wie wir bereits gesehen haben (in Abschnitt 5.2.1), war ja die ursprüngliche Schwierigkeitsskala auch eindimensional. Dies bestätigt sich hier nochmals. Die Schwierigkeits-Dimension ist in ihrer Meßqualität geringfügig schlechter als die Interessedimension. Die Ladungen erreichen im Durchschnitt Werte um die 0,7. Dies entspricht einer Varianzaufklärung von 50%.

Dimension 3: ANREGUNG

VAR	*wurde umgepolt	
014	*W2 Der Erdkundeunterricht regt mich an, über gesellschaftliche Probleme nachzudenken.	0,79
017	*I4 Der Erdkundeunterricht hat mich angeregt, über einige seiner Themen weiter nachzudenken.	0,76

Tab. 12: Dimension ANREGUNG aus der Faktorenanalyse der Gesamtbatterie

Die dritte Hauptdimension der Einstellungsbatterie setzt sich zusammen aus einem Item der ursprünglichen Wichtigkeitsskala und einem der ursprünglichen Interessenskala.

Beide Items sprechen in fast identischer Form einen Wirkungsaspekt von Erdkunde an, der weder etwas mit Interesse am Fach, noch mit der Wichtigkeit des Faches zu tun hat. Diese Dimension verkörpert im Bewußtsein der Schüler offensichtlich die ganz persönlichen Konsequenzen aus der Beschäftigung mit geographischen Problemen.

Dieser Aspekt der Einstellung zum Fach Erdkunde wurde bei Erstellung der Itembatterie nicht als relevant erkannt. Dadurch ist diese Dimension nur durch zwei Items abgedeckt, was für eine nachträgliche Subskalenbildung eigentlich nicht mehr ausreicht.

Dimension 4: NÜTZLICHKEIT

VAR		
010 *W1	Das was ich im Erdkundeunterricht erfahre, kann ich auch außerhalb der Schule gut gebrauchen.	0,55
027 *W6	Einer, der in Erdkunde viel weiß, kommt später im Leben besser voran.	0,81
029 *W8	Erdkundliches Wissen braucht man bei vielen Gelegenheiten im Leben.	0,73

Tab. 13: Dimension NÜTZLICHKEIT aus der Faktorenanalyse der Gesamtbatterie

Die Dimension 4 ist praktisch der Rumpf der ursprünglichen Wichtigkeitsskala.

Die inhaltliche Inspektion der Items läßt es allerdings naheliegend erscheinen, die Dimension als Nützlichkeitsdimension umzubenennen. Es geht bei den Items weniger darum, für wie wichtig Erdkunde allgemein gehalten wird, sondern darum, was man konkret mit erdkundlichem Wissen außerhalb (!) der Schule anfangen kann.

Dimension 5: BEDEUTUNG (im Vergleich zu anderen Schulfächern)

VAR		
018 W4	Die meisten Schulfächer sind wichtiger als EK.	0,80
022 W5	EK ist zwar recht interessant, aber im Vergleich zu anderen Fächern ziemlich unwichtig.	0,76

Tab. 14: Dimension BEDEUTUNG aus der Faktorenanalyse der Gesamtbatterie

Diese Dimension wird durch zwei Items der ursprünglichen Wichtigkeitsskala aufgespannt. Allerdings geht es auch hier nicht um Wichtigkeit von Erdkunde schlechthin, sondern um die Bedeutung des Faches im Rahmen des *schulischen* Alltags. Die Dimensionen 4 und 5 repräsentieren also den Innen- bzw. Außenaspekt bei der Relevanzeinschätzung des Faches durch die Schüler. Entscheidend ist, daß diese beiden Aspekte im Bewußtsein der Schüler nicht (!) systematisch zusammenhängen, wie ihre Aufdifferenzierung in zwei verschiedene Meßdimensionen zeigt.

Dieses Ergebnis der Dimensionalanalyse bedeutet zweierlei:

Erstens hat sich bestätigt, daß die zur Wichtigkeitsskala zusammengestellten Items ein Meßinstrument bilden, das mindestens zwei verschiedene Sachverhalte umfaßt. Man kann also aus diesem Items nicht einfach einen „Wichtigkeits"-Index oder dergleichen bilden. Von einer Skala im engeren Sinn kann sowieso nicht die Rede sein.

Zweitens hat sich gezeigt, daß man bei einer empirisch fundierten Itemauswahl die einzelnen Aspekte in den Schülereinstellungen besser hätte abdecken können. Praktisch verwertbar ist nämlich nur die „Interesse"- und die „Schwierigkeits"-Skala. Die restlichen 11 Items haben eine Meßqualität, die man mit einem Schrotschuß vergleichen könnte: Es wird zwar etwas „getroffen", nur ist die Streuung so groß, daß man nie genau weiß, was eigentlich getroffen (d. h. gemessen) wurde.

Könnte man jetzt noch etwas an der Einstellungsbatterie zum Erdkundeunterricht verändern, würde man versuchen, die drei Dimensionen ANREGUNG, NÜTZLICHKEIT und BEDEUTUNG durch *zusätzliche* Items abzudecken, um so das Globalmerkmal „WICHTIGKEIT" treffsicherer zu erfassen.

6. Die Meßqualität des Polaritätsprofils zum Fach Erdkunde allgemein

Neben der oben analysierten Einstellungsbatterie wurde vom RCFP noch ein weiteres Verfahren eingesetzt, um die Schülermeinungen zum Fach Erdkunde zu erfassen: das Polaritätsprofil.

Es handelt sich dabei um ein Profil, das in Anlehnung an *Anwander* und *Havers* (*Anwander* 1974, *Havers* 1972) zusammengestellt wurde. Sowohl das Fach Erdkunde allgemein als auch die jeweils erprobte RCFP-Unterrichtseinheit wurden von den Schülern anhand dieses Profils eingestuft. (Bei der Einheit „Tabi Egbe" verzichtete das RCFP auf die Profile – vermutlich da die Einheit bei sehr jungen Schülern erprobt wurde. Die Einheit INDIOS mußte von uns ausgeschlossen werden, da hier zwar das Polaritätsprofil nicht aber die Einstellungsbatterie den Schülern vorgelegt wurde.)

Das Polaritätsprofil besteht aus 20 (Rating-)Skalen, die durch polare Eigenschaftswörter aufgespannt werden. Die folgende Tab. 15 zeigt die Form, in der das Profil sowohl zur Einstufung des Faches als auch zur Einschätzung der Unterrichtseinheiten den Schülern vorgelegt wurde:

Was trifft für die gesamte Unterrichtseinheit
GASTARBEITERKINDER IN EINER DEUTSCHEN GROSSTADT
eher zu?

		1	2	3	4	5	
71	logisch	o	o	o	o	o	unlogisch
72	modern	o	o	o	o	o	altmodisch
73	stumpfsinnig	o	o	o	o	o	anregend
74	leicht	o	o	o	o	o	schwer
75	unwichtig	o	o	o	o	o	wichtig
76	bedrückend	o	o	o	o	o	erfreuend
77	übersichtlich	o	o	o	o	o	verwirrend
78	langweilig	o	o	o	o	o	interessant
79	sinnvoll	o	o	o	o	o	unsinnig
80	trocken	o	o	o	o	o	lustig
81	beweisbar	o	o	o	o	o	unbeweisbar
82	unbefriedigend	o	o	o	o	o	befriedigend
83	notwendig	o	o	o	o	o	überflüssig
84	kindlich	o	o	o	o	o	erwachsen
85	beengend	o	o	o	o	o	befreiend
86	realitätsbezogen	o	o	o	o	o	weltfremd
87	unmenschlich	o	o	o	o	o	menschlich
88	klar	o	o	o	o	o	unklar
89	unpolitisch	o	o	o	o	o	politisch
90	fortschrittlich	o	o	o	o	o	konservativ

Tab. 15: Das Polaritätsprofil des RCFP

6.1 Methodische Mängel des Profils

Das vom RCFP-Forschungsstab herangezogene Polaritätsprofil enthält leider eine Vielzahl methodischer Schwächen. Da die Erhebung nun einmal damit durchgeführt wurde, hat es wenig Sinn, nachträglich alle Schwachstellen herauszuarbeiten. Deshalb hier nur eine grobe Zusammenfassung der beiden wichtigsten Probleme.

(1) Fehlende Vergleichbarkeit mit anderen Untersuchungen

Das RCFP-Profil entspricht in etwa dem Profil von *Anwander* (*Anwander* 1974). Allerdings wurden von den 20 Polaritäten drei anders formuliert:

aus:	wissenschaftlich	– unwissenschaftlich	wurde
	leicht	– schwer,	
aus:	lebendig	– tot	wurde
	lustig	– trocken	und
aus:	fortschrittlich	– reaktionär	wurde
	fortschrittlich	– konservativ.	

Nur 17 Polaritäten sind also prinzipiell vergleichbar.

Leider verwenden aber sowohl *Anwander* als auch *Havers* 7-stufige Ratingskalen für die Polaritäten, während sich das RCFP-Forschungsteam für 5-stufige Skalen entschied. Damit wird die Vergleichbarkeit zwischen den Profilen stark eingeschränkt.

Die an sich sinnvolle Übernahme eines bereits empirisch erprobten Polaritätsprofils wurde durch diese Modifikation des RCFP-Forschungsstabes also wieder teilweise zunichte gemacht. Dabei wäre es höchst interessant gewesen, unmittelbar vergleichen zu können, wie – bei verschiedenen empirischen Untersuchungen – unterschiedliche Schulfächer eingestuft werden. Der Vergleich von Polaritätsprofilen zu Erdkunde, Religion, Geschichte, Sozialkunde, Deutsch und Physik wäre wesentlich aussagekräftiger als die bloße Einstufung von Erdkunde (bzw. der Erprobungseinheit). Durch die unterschiedliche Abstufung der Skalen lassen sich nur noch die sog. „Profilgestalten" der verschiedenen Fächer mit Erdkunde vergleichen, nicht jedoch die konkreten Profilwerte. Dabei geht erheblich an Information verloren. (Zur Problematik des Profilvergleichs s. *Bergler* 1975, S. 169 ff.

(2) Fehlen einer Qualitätsüberprüfung (Itemanalyse)

Da der RCFP-Forschungsstab im Endeffekt also praktisch ein fast neues Polaritätsprofil konstruiert hat, wären allgemein anerkannte Prinzipien der Profilkonstruktion anzuwenden gewesen. Dies wurde nicht getan.

Einer der wichtigsten Aspekte dabei ist die *Relevanz* der Polaritäten (vgl. *Bergler* 1975). Soweit uns bekannt ist wurde versäumt, durch Vortest

empirisch zu untersuchen, ob die Schüler bei der Einstufung von Unterricht tatsächlich die ausgewählten Eigenschaftspole als relevant erachten, bzw. ob nicht ganz andere, zusätzliche Merkmalspole bedeutungsvoll wären. Z. B. erscheint durchaus fraglich, ob die Polarität „menschlich-unmenschlich" für das Bedeutungsumfeld eines Schulfaches relevant ist (vgl. *Thompson/Stapleton* 1979/80).

Ein weiterer Aspekt ist die *Polaritätseigenschaft* der Skalen. Nirgendwo ist überprüft, ob die ausgesuchten Eigenschaftswörter tatsächlich jeweils *Gegen*pole einer eindimensionalen Skala sind, und zwar Gegenpole im Bewußtsein der Beurteilerpopulation, also der Schüler. Nehmen wir als Beispiel die Polarität „bedrückend-erfreuend". Ist der Gegenpol zu „bedrückend" nicht eher „befreiend"? Dies wäre jedenfalls in einem Pretest empirisch zu klären gewesen, wie er in einschlägigen Handbüchern auch dringend empfohlen wird.

Ein dritter Aspekt betrifft die *„dimensionale Differenzierung"* (*Bergler* 1975, S. 23) der Polaritäten. Es wurde nicht explizit untersucht, ob durch die Polaritäten genügend Bedeutungsdimensionen abgedeckt wurden.

Schließlich blieb ein letzter Aspekt unbeachtet, nämlich die sog. *„Objektzentrierung"* (*Bergler* 1975, S. 84). Dies bedeutet, daß die Einstufung der Schüler gewissermaßen im „luftleeren Raum" bewertet werden muß. Man kann lediglich die jeweilige Einstufung der RCFP-Erprobungseinheit mit der des Faches Erdkunde allgemein (je Schüler) vergleichen. Es fehlt aber beispielsweise eine Vergleichsmöglichkeit mit anderen Schulfächern. *Fürstenberg* und *Jungfer* schreiben, daß das Fach Erdkunde „überwiegend in positiven Kategorien beschrieben" wurde (*Fürstenberg/Jungfer* 1980, S. 58). Dies mag ja zutreffen, nur es stellt sich sofort die Frage, ob andere Schulfächer nicht noch positiver eingestuft worden wären. Da die Polaritätsprofile nicht mit den Untersuchungen von *Anwander* und *Havers* vergleichbar sind, hätte bei den RCFP-Untersuchungen auf jeden Fall mindestens noch ein weiteres Schulfach eingestuft werden müssen, damit die relative Position des Fachs Erdkunde bestimmbar wäre.

Aufgrund dieser methodischen Schwächen ist eine Itemanalyse des Polaritätsprofils besonders wichtig. Vor allem wird die Frage zu klären sein, ob die Qualität des Profils durch Ausschluß schwacher Polaritäten gesteigert werden kann. Optimal für ein solches Vorgehen wäre allerdings ein weit größerer „Itempool" als uns zur Verfügung steht. *Schäfer* und *Fuchs* sprechen von 50 bis 70 Polaritäten, die durch eine Itemanalyse auf ca. 4 Polaritäten je Dimension zu reduzieren wären (vgl. *Bergler* 1975, S. 135). Bei einem Profil mit fünf Meßdimensionen ergäben sich dann am Ende 20 Polaritäten – also die gleiche Zahl an Polaritäten, die wir als Ausgangsmaterial für unsere Analyse vorfinden.

Um die Meßqualität des Polaritätsprofils zu überprüfen – und wenn möglich zu erhöhen – wählen wir diesmal eine etwas andere Vorgehensweise als bei der Einstellungsbatterie. Nach den Erfahrungen mit der Einstellungsbatterie erscheint es sinnvoll, als erstes die Dimensionalität des Pro-

fils zu erfassen und erst danach eine klassische Kennwertanalyse für die einzelnen Dimensionen durchzuführen.

6.2 Faktorenanalyse (Dimensionsüberprüfung) des Gesamt-Polaritätsprofils zum Fach Erdkunde allgemein

Der erste Schritt zur Überprüfung der Meßqualität des Profils ist eine Faktorenanalyse. Die methodischen Voraussetzungen für die Verwendung dieses statistischen Verfahrens dürften weitgehend erfüllt sein. Mehrfach wurde nachgewiesen, daß eine einfache Hauptkomponentenanalyse auf der Basis von Produkt-Moment-Korrelationen bei Polaritätsprofilen durchaus angemessen ist (*Revenstorf* 1973).

(1) Die quadrierte multiple Korrelation der Polaritäten

Zur Vorbereitung für die Faktorenanalyse bestimmen wir zunächst die quadrierten multiplen Korrelationen. Sie geben für jede Polarität an, wie groß deren gemeinsame Varianz mit dem Gesamtprofil ist. Ein niedriger multipler Korrelationskoeffizient kennzeichnet die jeweilige Polarität als sehr spezifisch. Es ergibt sich Tab. 16:

VAR				
030	logisch	–	unlogisch	0,15
031	modern	–	altmodisch	0,22
032	stumpfsinnig	–	anregend	0,45
033	leicht	–	schwer	0,17
034	unwichtig	–	wichtig	0,37
035	bedrückend	–	erfreuend	0,34
036	übersichtlich	–	verwirrend	0,30
037	langweilig	–	interessant	0 52
038	sinnvoll	–	unsinnig	0,42
039	trocken	–	lustig	0,31
040	beweisbar	–	unbeweisbar	0,15
041	unbefriedigend	–	befriedigend	0,37
042	notwendig	–	überflüssig	0,40
043	kindisch	–	erwachsen	0,17
044	beengend	–	befreiend	0,27
045	realitätsbezogen	–	weltfremd	0,17
046	unmenschlich	–	menschlich	0,24
047	klar	–	unklar	0,35
048	unpolitisch	–	politisch	0,03
049	fortschrittlich	–	konservativ	0,24

Tab. 16: Die quadrierte multiple Korrelation (Bestimmtheitsmaß) des Polaritätsprofils zum Fach Erdkunde allgemein

Ein Mindestmaß gemeinsamer Varianz ist Voraussetzung dafür, daß eine Polarität bei der Dimensionsanalyse des Profils überhaupt einer Dimension eindeutig zugeordnet werden kann. Mehrere Polaritäten sind in dieser Hinsicht problematisch. Mit Sicherheit kann man die Polarität „unpolitisch-politisch" als sehr spezifisch bezeichnen. D. h. sie hat mit dem Polaritätsprofil insgesamt so gut wie gar nichts zu tun. Nur 3 % ihrer Varianz kann durch die anderen Polaritäten erklärt werden. Sie ist zweifellos für die weiteren Analysen auszuschließen.

(2) Die Faktorenanalyse des Polaritätsprofils (1. Runde)

Nach Ausschluß der Polarität „unpolitisch-politisch" ergaben sich bei einer Hauptkomponentenanalyse folgende vier signifikante Dimensionen (Eigenwert < 1), die ca. 50 % der Varianz erklären;

VAR			FAKTOR1	FAKTOR2	FAKTOR3	FAKTOR4
030 logisch	–	unlogisch	0,07	0,22	0,46	0,15
031 modern	–	altmodisch	0,12	0,23	0,13	0,52
032 stumpfsinnig	–	anregend	0,55	0,45	0,11	0,19
033 leicht	–	schwer	0,41	0,16	0,56	0,21
034 unwichtig	–	wichtig	0,20	0,77	0,04	0,09
035 bedrückend	–	erfreuend	0,68	0,13	0,14	0,12
036 übersichtlich	–	verwirrend	0,41	0,08	0,59	0,03
037 langweilig	–	interessant	0,61	0,48	0,14	0,13
038 sinnvoll	–	unsinnig	0,19	0,66	0,27	0,21
039 trocken	–	lustig	0,69	0,14	0,04	0,08
040 beweisbar	–	unbeweisbar	0,09	0,10	0,60	0,27
041 unbefriedigend	–	befriedigend	0,57	0,26	0,13	0,30
042 notwendig	–	überflüssig	0,14	0,74	0,18	0,13
043 kindisch	–	erwachsen	0,16	0,04	0,01	0,67
044 beengend	–	befreiend	0,58	0,02	0,08	0,36
045 realitätsbezogen	–	weltfremd	0,13	0,27	0,45	0,31
046 unmenschlich	–	menschlich	0,29	0,02	0,19	0,57
047 klar	–	unklar	0,35	0,21	0,60	0,09
049 fortschrittlich	–	konservativ	0,09	0,22	0,16	0,58

Tab. 17: Ergebnisse der Faktorenanalyse des Polaritätsprofils zum Fach Erdkunde allgemein (1. Runde)

Wie die Tabelle zeigt, erbrachte die Faktorisierung des Profils einigermaßen trennscharf besetzte Faktoren. Lediglich die Polaritäten VAR030 („logisch-unlogisch") und VAR045 („realitätsbezogen-weltfremd") lassen sich nicht eindeutig einer Dimension zuordnen. Sie haben auf keinem Faktor Ladungen über 0,5. Beide Polaritäten haben auch relativ geringe ge-

meinsame Varianz mit dem Gesamtprofil, wie ihre geringe quadrierte multiple Korrelation anzeigt (vgl. Tab. 16). Sie werden deshalb für die endgültige Festlegung der Dimensionen (2. Runde der Faktorisierung) ausgeschlossen.

(3) Die zweite Runde der Faktorenanalyse (endgültige Dimensionen)

Bei der 2. Runde der Faktorisierung erzielten wir durch diesen Ausschluß eine insgesamt etwas bessere Varianzaufklärung von 53%. Es ergaben sich folgende vier Faktoren:

Faktor 1: ANREGUNG – MOTIVATION

	Was trifft eher für EK zu?			
VAR				Ladungen
032	stumpfsinnig	–	anregend	0,59
035	bedrückend	–	erfreuend	0,69
037	langweilig	–	interessant	0,63
039	trocken	–	lustig	0,70
041	unbefriedigend	–	befriedigend	0,55
044	beengend	–	befreiend	0,59

Tab. 18: Dimension ANREGUNG–MOTIVATION aus der Faktorenanalyse des Polaritätsprofils zum Fach Erdkunde allgemein

Die Ladung der Items auf diesem Faktor sind gut bis sehr gut. Berechnet man pro Item die Varianzaufklärung, so kommt man auf eine durchschnittliche Varianzaufklärung von etwa 40%. D. h. 40% der Varianz jeder Polarität wird im Durchschnitt allein durch diesen Faktor 1 erklärt.

Alle Polaritäten beziehen sich auf Bedeutungskomponenten des Faches Erdkunde, die etwas mit der motivierenden, anregenden Kraft des Faches zu tun haben. Die Dimension erfaßt also, ob die Schüler Erdkunde als interessantes, lustiges und befriedigendes Fach empfinden oder als ein langweiliges, trockenes und unbefriedigendes Schulfach. Damit ist diese Dimension des Polaritätsprofils sehr gut vergleichbar mit der ersten Dimension aus der Einstellungsbatterie, die wir als „Interesse"-Dimension bezeichnet hatten. Allerdings umfaßt unser Faktor ANREGUNG-MOTIVATION zusätzliche Einstellungsaspekte, die beim „Interesse"-Faktor aus der Einstellungsbatterie ausgeklammert waren. Es handelt sich um die mehr subjektiven Aspekte der „Freude" und des „Vergnügens", wie sie in den Polaritäten „trocken-lustig", „bedrückend-erfreuend" und „unbefriedigend-befriedigend" zum Ausdruck kommen. Insofern enthält der Faktor „ANREGUNG-MOTIVATION" eher die ganz persönliche Einschätzung der Schüler in bezug auf ihr Interesse am Fach, während in der „Interesse"-

Dimension der Batterie die mehr objektive „Interessantheit" des Faches erfaßt wird.

Faktor 2: RELEVANZ

Was trifft für EK eher zu?				
VAR				Ladungen
034	unwichtig	–	wichtig	0,77
038	sinnvoll	–	unsinnig	0,69
042	notwendig	–	überflüssig	0,76

Tab. 19: Dimension RELEVANZ aus der Faktorenanalyse des Polaritätsprofils zum Fach Erdkunde allgemein

Die drei Items dieses Faktors haben sehr hohe Ladungen. Er erklärt *allein* im Durchschnitt über 50% ihrer Varianz. Der Faktor erfaßt klar, wie die Schüler die Bedeutung des Faches Erdkunde einschätzen. Er ist vergleichbar mit dem Faktor „Bedeutung" aus der Einstellungsbatterie, der allerdings noch zusätzlich angesprochen hatte, wie die Schüler die *Relevanz* des Faches im Vergleich zu den anderen Schulfächern sehen.

Faktor 3: IMAGE

Was trifft eher für EK zu?				
VAR				Ladungen
031	modern	–	altmodisch	0,53
043	kindisch	–	erwachsen	0,67
046	unmenschlich	–	menschlich	0,60
049	fortschrittlich	–	konservativ	0,61

Tab. 20: Dimension IMAGE aus der Faktorenanalyse des Polaritätsprofils zum Fach Erdkunde allgemein

Auch bei diesem Faktor sind die Ladungen der Items noch relativ hoch. Die Polaritäten dieses Faktors sprechen ganz eindeutig Eigenschaften des Faches Erdkunde an, die man als sein Image bezeichnen würde. Typisch dafür ist, daß die Polaritäten Eigenschaften benennen, die im Zusammenhang mit einem Schulfach eigentlich völlig unpassend sind (z. B. ein „kindisches" Schulfach).

Das Image eines Objektes – in unserem Fall eines Schulfaches – kann sehr stark von seinen „objektiven" Eigenschaften entfernt sein. Es ist deshalb bedauerlich, daß nur relativ wenige Items in das Polaritätsprofil aufgenommen wurden, die zur Erfassung des Image geeignet sind. Die meisten Items des Polaritätsprofils erfassen relativ konkrete, handfeste Eigenschaf-

ten des Faches und weniger die subjektiven, emotionalen Aspekte in den Einstellungen der Schüler.

Faktor 4: SCHWIERIGKEIT

	Was trifft für EK eher zu?			
VAR				Ladungen
033	leicht	–	schwer	0,70
036	übersichtlich	–	verwirrend	0,65
040	beweisbar	–	unbeweisbar	0,51
047	klar	–	unklar	0,64

Tab. 21: Dimension SCHWIERIGKEIT aus der Faktorenanalyse des Polaritätsprofils zum Fach Erdkunde allgemein

Die 4. Dimension in den Polaritäten des Porfils erfaßt, wie die Schüler die Schwierigkeit des Faches einschätzen. Sie dürfte Ähnliches messen wie der Schwierigkeitsfaktor aus der Einstellungsbatterie.

(4) Zusammenfassung

- Die beste Lösung bei der Faktorisierung (in Runde 2) erbrachte eine insgesamte Varianzaufklärung von 53 %. Dies ist ein durchaus akzeptables, aber kein sehr gutes Ergebnis. Wenn man also statt der 17 Einzelpolaritäten (20 ursprüngliche minus 3 ausgeschlossene Polaritäten) nur 4 „factor-scores" pro Schüler heranzieht, dann erfaßt man etwas mehr als die Hälfte der Varianz der ursprünglichen Messung.
- Nicht alle Polaritäten lassen sich eindeutig einer Dimension zuordnen. Die Polarität „langweilig-interessant" wurde mit einer Ladung von 0,63 zurecht dem Faktor 1 zugeordnet. Leider hat sie jedoch auch auf dem Faktor 2 eine beachtliche Ladung von knapp unter 0,5. Diese Ambiguität einiger Polaritäten deuten darauf hin, daß noch nicht für jede Meßdimension die eindeutigsten Polaritäten im Profil vorhanden sind.
- Das Polaritätsprofil mißt die Einstellungen der Schüler zum Fach Erdkunde sehr viel differenzierter als dies vom RCFP-Forschungsteam unterstellt wurde. Dies läßt sich trotz obiger „Schwächen" der Faktorisierung feststellen. Das Profil mißt mit Sicherheit mehr als nur eine „positive (oder negative) Grundhaltung dem Erdkundeunterricht gegenüber" (*Fürstenberg/Jungfer* 1979, S. 95). Es erfaßt – wenn auch nicht mit wünschenswerter Eindeutigkeit – vier Einstellungsdimensionen, die inhaltlich konsistent und theoretisch aussagekräftig sind.
- Leider sind die vorhandenen Dimensionen teilweise mit zu wenigen Polaritäten abgedeckt.

6.3 Reliabilitätsanalyse der einzelnen Dimensionen des Gesamt-Polaritätsprofils zum Fach Erdkunde allgemein

Nachdem geklärt ist, daß wir das Polaritätsprofils als *mehrdimensionales* Meßinstrument behandeln dürfen (und sollten), stellen sich als nächstes folgende Fragen:

Wie gut, d. h. wie trennscharf und reliabel messen die einzelnen Polaritäten je Dimension?

Welche Meßqualität hat jede einzelne Meßdimension als Ganzes? Läßt sich die Meßqualität pro Dimension nachträglich weiter verbessern, indem einzelne Polaritäten zusätzlich ausgeschlossen werden? Hierbei müssen wir allerdings beachten, daß eine zu starke Verringerung der Polaritäten pro Dimension die Meßqualität wieder automatisch verschlechtert. Es geht also darum, ein relatives Optimum an Meßqualität zu finden.

In Tabelle 22 finden sich sämtliche Item-Kennwerte für alle Polaritäten – jeweils berechnet in Bezug auf die entsprechende Meßdimension.

Im großen und ganzen kann man sagen, daß alle Dimensionen des Polaritätsprofiles reliabel und trennscharf messen. Allerdings gibt es bemerkenswerte Unterschiede: Die Dimensionen 1 und 2 (ANREGUNG und RELEVANZ) erreichen sehr gute bzw. gute Gesamtreliabilitäten von α=0,81 bzw. α=0,75. Die Dimensionen 3 und 4 (IMAGE undSCHWIERIGKEIT) dagegen kommen nur auf eine Gesamtreliabilität von α=0,58 bzw. α=0,60.

Der Grund für die mangelnde Reliabilität der beiden letzten Dimensionen ist eindeutig die geringe Trennschärfe ihrer Polaritäten. Selbst wenn man die Trennschärfe anhand der Schwierigkeit normiert ($T_{korr/norm}$), haben die Polaritäten dieser beiden Dimensionen nur eine durchschnittliche Trennschärfe von 0,41 bzw. 0,44. Zum Vergleich: Die normierten Trennschärfen der Polaritäten auf der ersten Dimension betragen im Durchschnitt 0,61; bei Dimension 2 liegt der Durchschnitt der Trennschärfe sogar bei 0,74.

S:	Schwierigkeitsindex	R:	Reliabilitätsindex bei Eliminierung der betreffenden Polarität (Cronbachs α)
T_{korr}:	Corrected-Item-Total-Correlation		
$T_{korr/norm}$:	an der max. Trennschärfe normierter Index	T_{fac}:	Trennschärfe in Bezug auf die „factor-scores"

Oft wird als Grund für mangelnde Reliabilität eine zu geringe Anzahl von Items pro Meßdimension vermutet, weil die Reliabilität einer Skala (bzw. Meßdimension) automatisch geringer wird, wenn die Zahl der Items abnimmt- vorausgesetzt alle anderen Kennwerte bleiben gleich. Diese Vermutung trifft in unserem Fall nicht zu. Die Dimension 2 z. B. hat weniger Polaritäten als Dimension 3, aber sowohl deutlich höhere Reliabilität als

DIMENSION 1: ANREGUNG $\alpha_{(gesamt)} = 0,81$

VAR				S	T_{korr}	$T_{korr/norm}$	α	T_{fac}
032	stumpfsinnig	-	anregend	0,73	0,61	0,69	0,77	0,57
035	bedrückend	-	erfreuend	0,42	0,55	0,56	0,78	0,69
037	langweilig	-	interessant	0,71	0,67	0,74	0,75	0,63
039	trocken	-	lustig	0,32	0,54	0,58	0,79	0,70
041	unbefriedigend	-	befriedigend	0,62	0,57	0,59	0,78	0,56
044	beengend	-	befreiend	0,32	0,48	0,51	0,80	0,59
			Mittelwerte	0,52	0,57	0,61		

DIMENSION 2: RELEVANZ $\alpha_{(gesamt)} = 0,75$

VAR				S	T_{korr}	$T_{korr/norm}$	α	T_{fac}
034	unwichtig	-	wichtig	0,76	0,56	0,66	0,68	0,77
038	sinnvoll	-	unsinnig	0,86	0,55	0,79	0,69	0,70
042	notwendig	-	überflüssig	0,80	0,61	0,76	0,62	0,76
			Mittelwerte	0,81	0,57	0,74		

DIMENSION 3: IMAGE $\alpha_{(gesamt)} = 0,58$

VAR				S	T_{korr}	$T_{korr/norm}$	α	T_{fac}
031	modern	-	altmodisch	0,72	0,35	0,39	0,52	0,52
043	kindisch	-	erwachsen	0,59	0,34	0,35	0,53	0,67
046	unmenschlich	-	menschlich	0,70	0,36	0,39	0,52	0,60
049	fortschrittlich	-	konservativ	0,79	0,41	0,50	0,48	0,61
			Mittelwerte	0,70	0,37	0,41		

DIMENSION 4: SCHWIERIGKEIT $\alpha_{(gesamt)} = 0,60$

VAR				S	T_{korr}	$T_{korr/norm}$	α	T_{fac}
033	leicht	-	schwer	0,53	0,34	0,34	0,56	0,34
036	übersichtlich	-	verwirrend	0,68	0,46	0,49	0,46	0,49
040	beweisbar	-	unbeweisbar	0,80	0,25	0,31	0,62	0,31
047	klar	-	unklar	0,82	0,48	0,62	0,45	0,62
			Mittelwerte	0,71	0,38	0,44		

Tab. 22: Item-Kennwerte der einzelnen Dimensionen aus der Faktorenanalyse des Polaritätsprofils zum Fach Erdkunde allgemein

auch wesentlich bessere Trennschärfe (wobei beide Dimensionen relativ „schwierige" Polaritäten haben). Der Grund für die relativ schlechte Meßqualität der Dimension 3 (IMAGE) und 4 (SCHWIERIGKEIT) liegt also in der mangelden Trennschärfe ihrer Polaritäten.

Woher kommt diese geringere Trennschärfe bei den Dimensionen 3 und 4?

Bei der Dimension 3 (IMAGE) liegt dies an den hohen Anforderungen, die an das Abstraktionsvermögen der Schüler gestellt werden: Polaritäten wie „kindlich-erwachsen" oder „menschlich-unmenschlich" haben wenig konkreten Bezug zu einem Schulfach. Ebenso verhält es sich mit der Polarität „fortschrittlich-konservativ". Ein Lehrer mag für die Schüler menschlich, modern oder fortschrittlich erscheinen – oder auch nicht. Um das selbe über ein Schulfach sagen zu können, bedarf es einer erheblichen Fähigkeit zur Abstraktion. Offensichtlich wußten viele Schüler – vor allem vermutlich die jüngeren – nicht so recht, was sie eigentlich antworten sollten. Jedenfalls kam kein so einheitliches, in sich konsistentes Antwortmuster bei dieser Einstellungsdimension zustande, wie bei den anderen Dimensionen. Trotzdem ist die Dimension IMAGE aus theoretischen Gründen von großer Bedeutung, auch wenn ihre Meßqualität mittelmäßig ist.

Bei der Dimension 4 (SCHWIERIGKEIT) resultiert die mangelde Trennschärfe vermutlich aus der sprachlich unglücklichen Auswahl der Polaritäten. Vor allem die Polarität „beweisbar-unbeweisbar", die die schlechtesten Kennwerte hat, ist inhaltlich wenig überzeugend.

Was soll man sich auch unter einem beweisbaren Schulfach vorstellen? Einzelne Inhalte eines Schulfaches, z. B. eine Gleichung in Mathematik, mögen beweisbar (und deshalb gerade „schwierig") sein – aber kaum ein ganzes Schulfach als solches.

Zusammenfassend kann man feststellen:

- Alle Dimensionen des Polaritätsprofils werden von ausreichend trennscharfen und reliablen Polaritäten abgedeckt, wobei allerdings die Dimensionen 3 und 4 (IMAGE und SCHWIERIGKEIT) die Meßstandards nur knapp erreichen.
- Bei einer weiteren Analyse sollten im Prinzip zwei Polaritäten ausgeschlossen werden: In Dimension 4 (SCHWIERIGKEIT): „beweisbar-unbeweisbar" und in Dimension 1 (ANREGUNG): „beengend – befreiend".

  Da jedoch die Dimension 4 dann nur noch aus drei Polaritäten bestehen würde, lassen wir die erste Polarität weiter im Variablensatz, obwohl deren Ausschluß die Reliabilität erhöhen würde. Diese technisch mögliche, geringfügige Reliabilitätsverbesserung erscheint unerheblich in Anbetracht der dann sehr „mager" besetzten Meßdimension. Auch die zweite Polarität wird nicht eliminiert, und zwar deshalb, weil deren Trennschärfe und Reliabilität im Vergleich mit den Polaritäten aus den anderen Dimensionen immer noch recht gut ist, auch wenn sie bei ihrer eigenen Dimension den schlechtesten Wert aufweist.
- Die Kennwertanalyse bestätigt damit im wesentlichen die Ergebnisse der Faktorenanalyse – vor allem die relativ gute Meßqualität der beiden

ersten Dimensionen, und das besonders schlechte Abschneiden der Dimension 3 (IMAGE).

Das oben untersuchte Polaritätsprofil wurde – wie schon erwähnt – zweimal in dem RCFP-Fragebogen verwendet: Einmal zur Einstufung des *Faches Erdkunde allgemein* und einmal zur Einstufung der jeweiligen *RCFP-Erprobungseinheit*. Nachdem wir nun untersucht haben, wie gut sich das Profil zur Einstufung des Faches eignet, müßten wir nun unsere Analysen wiederholen, um seine Meßqualität bei der Einstufung der jeweiligen Unterrichtseinheiten zu überprüfen. Das Profil könnte für die Einstufung des Unterrichtsfaches geeignet sein, aber nicht für die Einstufung der Unterrichtseinheiten.

Wir haben dies ebenfalls überprüft – hier die wichtigsten Ergebnisse:

- Die Faktorenanalyse ergab bei dem Profil, mit dem die Erprobungseinheiten eigestuft wurden, im wesentlichen die gleichen Dimensionen, wie beim Fach Erdkunde.
- Die Analyse der Kennwerte bestätigt auch hier im wesentlichen die Ergebnisse aus der Faktorenanalyse. Alle Dimensionen (mit Ausnahme der Dimension „Schwierigkeit") werden von (noch) ausreichend reliablen und trennscharfen Polaritäten abgedeckt. Die ersten drei Dimensionen „Anregung", „Relevanz" und „Image" haben sogar teilweise sehr gute Kennwerte.
- Es zeigte sich, daß das Profil sogar besser zur Einstufung der RCFP-Einheiten geeignet war als zur Einstufung des Faches allgemein. So ergeben sich bei der Einstufung der RCFP-Einheiten generell bessere Kennwerte für die Trennschärfe und die Reliabilität und auch bessere Faktorenlösungen. D. h. das Polaritätsprofil hatte bei der Erfassung der Schülereinstellungen zu den RCFP-Einheiten bessere Meßqualitäten als bei der Erfassung der Schülereinstellungen zum Fach Erdkunde.
- Durch Eliminierung zweier Polaritäten (nämlich: „politisch-unpolitisch" und „logisch-unlogisch") ist es uns wieder gelungen, eine bessere Aufgliederung des Gesamtprofils auf die einzelnen Meßdimensionen zu erreichen.

Nach dieser Überprüfung der Meßqualität wurden für die Einstellungsbatterie und für die beiden Polaritätsprofile sog. Faktorwerte berechnet, d. h. Meßwerte auf den gefundenen Meßdimensionen. Die Einzelheiten der Berechnung können hier aber nicht dargestellt werden. Damit wollen wir unsere Diskussion zu Problemen der Meßqualität abschließen. An zwei Beispielen aus der RCFP-Untersuchung – einer Einstellungsbatterie und einem Polaritätsprofil – konnte gezeigt werden, wie sich die Meßqualität eines Erhebungsinstruments überprüfen und – sogar nachträglich noch – verbessern läßt. Faktorenanalyse und klassische Testtheorie sind zwei bewährte Verfahren, die sich dafür eignen.

Im nächsten Kapitel soll nun eine ganz besonders problematische Komponente im empirischen Forschungsprozeß behandelt werden: die bivariate Analyse von Variablenzusammenhängen.

7. Bivariate Analysen von Zusammenhängen aus den RCFP-Erhebungen

Der Hauptzweck jeder empirischen Untersuchung ist die Analyse von Zusammenhängen zwischen den gemessenen Variablen. Zwar gibt es empirische Erhebungen, die lediglich beschreiben, wie bestimmte Merkmale (z. B. die „Note im Fach Erdkunde") in der untersuchten Population verteilt sind. Solche rein deskriptiven Arbeiten befinden sich jedoch erst im Vorstadium einer wissenschaftlichen Herangehensweise. Eine wissenschaftliche Untersuchung dagegen fragt nicht nur danach, „wie" bestimmte (soziale) Merkmale verteilt sind, sondern auch „warum" sie so und nicht anders zusammenhängen. Diese „warum"-Fragen werden in Form der Hypothesen (über kausale Zusammenhänge) zu beantworten versucht.

Die grundlegende wissenschaftliche Methode zur Hypothesenprüfung ist das Experiment zur Aufdeckung eines kausalen Variablenzusammenhanges. Seit vielen Jahrzehnten existiert in den Sozialwissenschaften aber auch eine (alternative) Methodologie, um Kausalanalysen auch bei nichtexperimentellem Design (z. B. bei Erhebungen wie der des RCFP) durchführen zu können.[1]

Die prinzipiell einfachste Methode zur Analyse von Variablenzusammenhängen ist der bivariate Ansatz. Im folgenden werden die Möglichkeiten und Grenzen dieser bivariaten Methode demonstriert. Zunächst jedoch verschaffen wir uns einen Überblick über den Datensatz „SCHÜLER":

	Anzahl Schüler	in %
Die Projekte		
FLUG	786	13,9
RHEIN	852	15,0
GELT	802	14,1
BODEN	1042	18,4
BRAND	974	17,2
GAST	543	9,6
MOBI	669	11,8
Das Geschlecht		
männlich	2772	48,9
weiblich	2853	50,3
keine Antwort	43	0,7
Die Erdkundenote		
1	231	4,1
2	1584	27,9
3	2391	42,2
4	1254	22,1
5	133	2,3
6	1	0
keine Antwort	74	1,3
Summe	5668	100,0

	Anzahl Schüler	in %
Die Schultypen		
Hauptschule	490	8,6
Gesamtschule	246	4,3
Realschule	946	16,7
Berufsschule	155	2,7
Gymnasium	3831	67,6
Die Klassenstufen		
Klasse 5	120	2,1
Klasse 6	517	9,1
Klasse 7	1015	17,9
Klasse 8	1346	23,7
Klasse 9	1166	20,6
Klasse 10	849	15,7
Klasse 11	585	10,3
Klasse 12	70	1,2
Summe	5668	100,0

Tab. 23: Deskriptiver Überblick über den Datensatz „SCHÜLER"

7.1. Die Erdkundenote

Die Erdkundenoten sind natürlich *kein* Indikator für den wirklichen Leistungsstand eines Schülers in diesem Fach – jedenfalls nicht ein allgemeingültiger Indikator im streng wissenschaftlichen Sinn: zu groß sind die subjektiven Elemente in der Notengebung des Lehrers, zu stark unterscheidet sich die Bedeutung des Faches (und damit die Strenge in den Zensuren) je nach Schulart, und zu unterschiedlich sind die Lehrpläne je nach Bundesland. Die Note 2 in einem Erdkunde-Leistungskurs an einem bayerischen Gymnasium repräsentiert einen anderen relativen Leistungsstand als die gleiche Note an einer hessischen Hauptschule – um ein beliebiges Beispiel herauszugreifen.

Dies ist in der wissenschaftlichen Diskussion unbestritten (vgl. *Knoche* 1969, S. 101 und S. 11). Ein Vergleich der Erdkundenoten nach Schultyp oder Klassenstufe sagt also weniger etwas über das jeweilige Leistungsniveau der Schüler aus, als vielmehr über Besonderheiten der Notengebung durch die Lehrer. In diesem Sinn sind auch die folgenden Ergebnisse unserer bivariaten Analyse zu sehen: Sie zeigen z. B., daß sich die Praxis der Notenvergabe im Fach Erdkunde sehr stark von Schultyp zu Schultyp unterscheidet.

7.1.1 Erdkundenote und Schultyp

Schon die Durchschnittsnoten im Fach Erdkunde unterscheiden sich deutlich von Schultyp zu Schultyp:

Schultyp	Durchschnittsnote in Erdkunde
Hauptschule	3,16
Gesamtschule	2,65
Realschule	3,12
Berufsschule	2,47
Gymnasium	2,86
Durchschnittsnote insgesamt	2,91

Tab. 24: Durchschnittsnote im Fach Erdkunde je nach Schultyp

Diese Unterschiede in den durchschnittlichen Erdkundenoten je Schultyp sind hochsignifikant, wie eine Varianzanalyse ergab ($F = 42{,}54$; $p = 0{,}000$).

Trotzdem sehen die Unterschiede in der Tabelle 24 relativ unbedeutend aus. Warum ist das so, trotz der festgestellten hohen Signifikanz der Unterschiede?

Der Grund liegt darin, daß sich beim Vergleich der Durchschnittswerte in den Erdkundenoten nicht die Unterschiede in den *Verteilungen* der *einzelnen* Noten je Schultyp widerspiegelt.

Das sieht man sofort an folgender Abbildung:

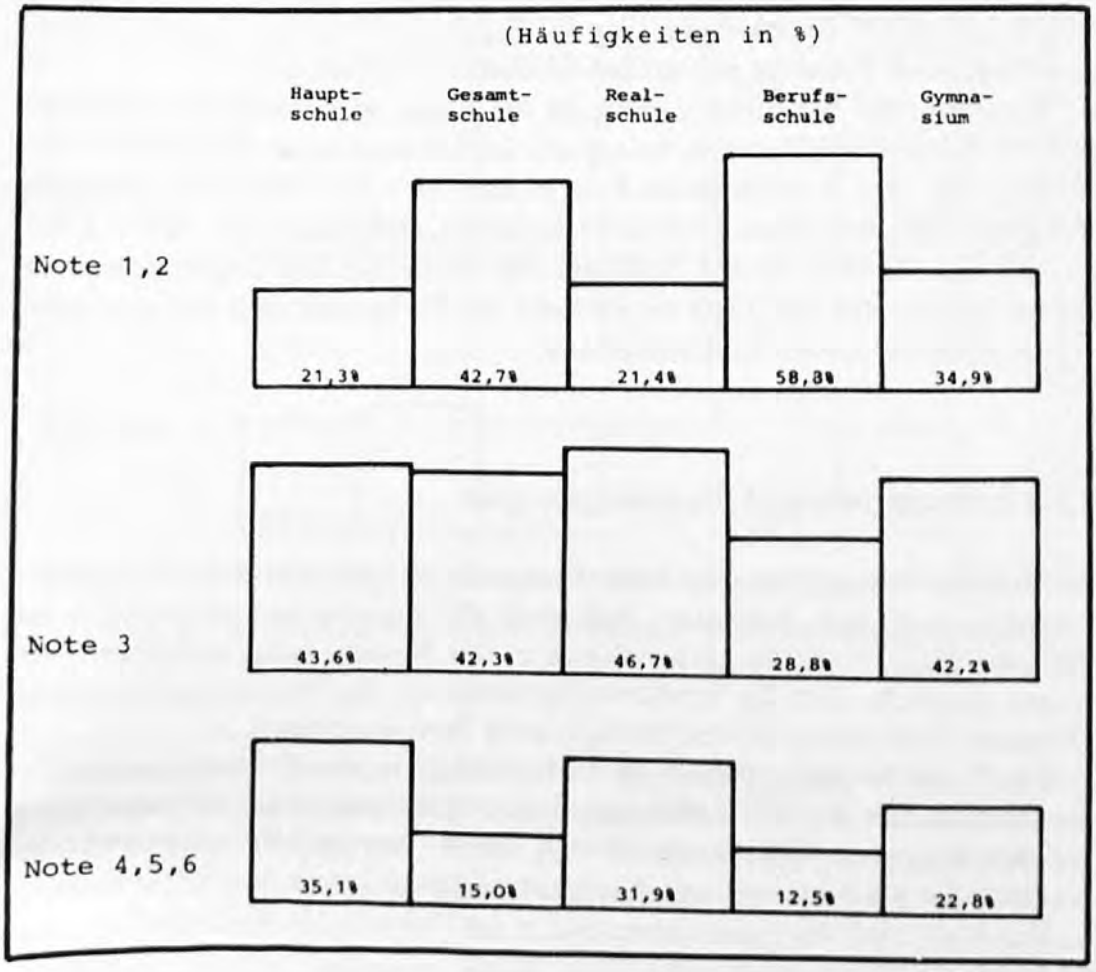


Abb. 7: Die Noten im Fach Erdkunde je nach Schultyp

Betrachet man statt der Durchschnittswerte die einzelnen Notenstufen (bzw. die zu Gruppen zusammengefaßten Notenstufen), so zeigt sich, daß an der Berufsschule wesentlich mehr „Einser" und „Zweier" im Fach Erdkunde vergeben werden als in allen anderen Schularten. Beispielsweise bekommen Berufsschüler fast dreimal häufiger gute oder sehr gute Noten als Haupt- oder Realschüler.

Auch die Gesamtschüler kommen häufig zu Spitzennoten: 43% von ihnen erreichen die Note 1 und 2 aber nur 15% müssen eine 4, 5 oder 6 hinnehmen.

Am schlechtesten ergeht es den Hauptschülern: 35% von ihnen schaffen nur eine 4 oder 5.

Am Gymnasium sind die Noten ziemlich gleichmäßig verteilt: Etwa jeder dritte Gymnasiast schafft eine „Eins" oder „Zwei"; 40% bekommen die Note 3 und jeder Fünfte bekommt Note 4 oder 5.

Wir haben zusätzlich die Schultypen nach *einzelnen* Noten verglichen (und nicht nach zusammengefaßten Gruppen wie oben): Die Unterschiede sind hier noch drastischer. Ein Chi-Quadrat-Test ergab mit $\chi^2 = 201{,}83$, df=20, p=0,0000 ein hochsignifikantes Ergebnis. Es besagt, daß die Notengebung je nach Schultyp mit großer Sicherheit ungleich ist.

Was uns jetzt noch interessiert, ist die Frage, wie *stark* die Zugehörigkeit zu einem bestimmten Schultyp die Erdkundenote bestimmt. Dazu berechnen wir den Koeffizienten Eta-Quadrat (η^2) (für Note als abhängige Variable): Der Wert von $\eta^2 = 0{,}0295$ bedeutet, daß (0,0295 x 100 = 2,95) ca. 3% der Varianz in der Notengebung allein aus der Zugehörigkeit zu einem bestimmten Schultyp zu erklären ist. Es handelt sich also um einen relativ *unbedeutenden* Zusammenhang.

7.1.2 Erdkundenote und Erprobungsprojekt

Da sich die Notengebung im Fach Erdkunde je nach Schultyp also unterscheidet, darf man vermuten, daß auch die Erprobungspopulationen zu den einzelnen Projekten Unterschiede in der Notengebung aufweisen. Wir wissen nämlich, daß die Schülerstichproben für die Projekterprobung die einzelnen Schultypen unterschiedlich stark berücksichtigen.

Die Frage ist nur: können die Unterschiede in der Notengebung je Erprobungseinheit des RCFP *allein* aus der unterschiedlichen Berücksichtigung der Schultypen (je Erprobungseinheit) erklärt werden? Betrachten wir dazu zunächst die Notengebung je Erprobungspopulation:

Man sieht, daß die Erprobungsschüler des Projekts MOBI besonders viele schlechte Erdkundenoten bekommen haben – es sind mehr als ein Drittel, die eine „4" oder eine „5" haben. Die Gastarbeiter-Erprobung dagegen erfolgte bei auffallend „guten" Schülern (jedenfalls was das Fach Erdkunde betrifft): Mehr als 38% von ihnen hatten eine „Eins" oder eine „Zwei" in Erdkunde.

Diese Unterschiede bei der Notengebung sind statistisch höchstsignifikant ($\chi^2 = 118{,}2$, df=30, p=0,000). Allerdings ist auch hier die Stärke des Zusammenhangs nur schwach ($\eta^2 = 0{,}11$), d. h. die Erdkundenoten der Schüler lassen sich aus der Beteiligung an einem bestimmten Projekt nur zu ca. 1% ihrer Varianz erklären ($\eta^2 = 0{,}013$).

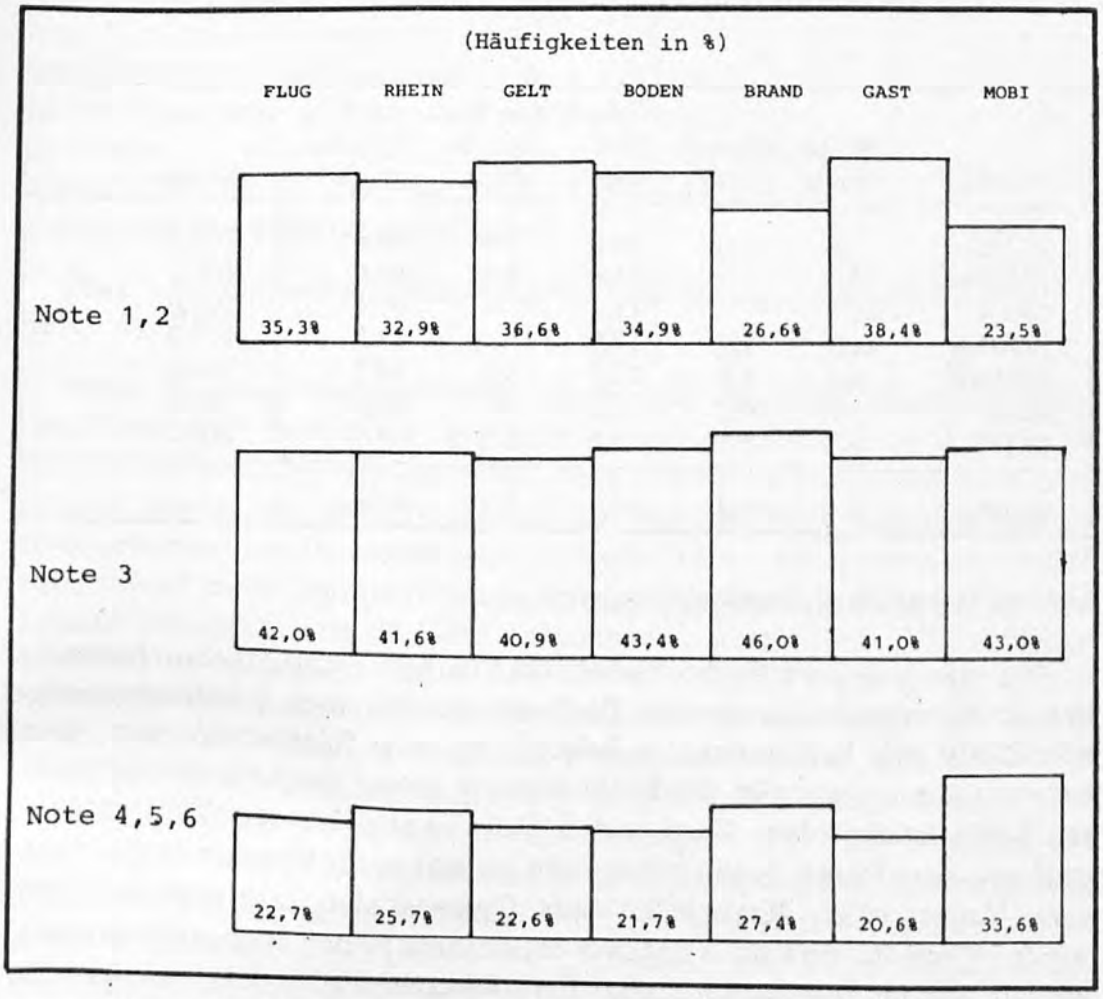


Abb. 8: Die Noten im Fach Erdkunde ja nach RCFP-Erprobungsprojekt

Kommen wir nun zurück auf unsere eingangs gestellte Frage: Ergeben sich diese Unterschiede je Erprobungspopulation einfach aus dem indirekten Effekt der Schultypen? Wäre dies der Fall, müßte beispielsweise die Einheit MOBIlität überwiegend an Hauptschulen, Realschulen oder Gymnasien erprobt worden sein, wo die Zensuren im allgemeinen schlechter ausfallen. Umgekehrt müßte die GASTarbeiter-Erprobung primär an Berufs- oder Gesamtschulen erfolgt sein, was die überdurchschnittlich guten Erdkundenoten erklären würde. Sehen wir uns also die Aufgliederung der Erprobung je Schultyp an.

7.2 Erprobung und Schultyp

Projekt	Schultyp in % Haupt-schule	Gesamt-schule	Real-schule	Berufs-schule	Gymna-schule	Summe
FLUG	7,6	7,0	16,0	8,9	60,4	100,0
RHEIN	3,3	–	23,8	3,4	69,4	100,0
GELT	21,8	–	11,1	–	67,1	100,0
BODEN	11,8	2,6	17,9	–	67,8	100,0
BRAND	6,1	6,8	22,8	3,5	60,9	100,0
GAST	–	14,2	5,7	–	80,1	100,0
MOBI	6,7	3,1	13,3	3,3	73,5	100,0
Summe	8,6	4,3	16,7	2,7	67,6	

Tab. 25: Die RCFP-Erprobungen je nach Schultyp

Wie man sieht, wurde die Einheit GASTarbeiter weit überdurchschnittlich an Gesamtschulen erprobt. Da Gesamtschüler auch überdurchschnittlich häufig gute Erdkundenoten bekommen, ist es folgerichtig, wenn diese Erprobungspopulation in der Erdkundenote besser abschnitt als die anderen Schülerstichproben. Ganz anders sieht es aber bei der MOBIlitäts-Erprobung aus: Deren Schülerstichprobe enthält *nicht* überdurchschnittlich viele Hauptschüler, Realschüler oder Gymnasiasten, wie man erwarten würde. Vielmehr sind die einzelnen Schultypen in der Stichprobe etwa so verteilt wie im Durchschnitt aller Erprobungen. Woher kommen also die schlechteren Zensuren dieser Schüler?

Betrachten wir die Schülerstichprobe aus der Erprobung der GELTinger Bucht. Fast 22% ihrer Schüler sind Hauptschüler und 11% sind Realschüler. D. h. fast ein Drittel dieser Erprobungsschüler kommt aus Schularten, in denen im allgemeinen *schlechte* Erdkundezensuren gegeben werden. Zu erwarten wäre also, daß diese Erprobungsschüler vergleichsweise schlechte Erdkundenoten haben. Genau das Gegenteil ist der Fall: Die Schüler aus der Erprobung der GELTinger Bucht hatten bei den Erdkundenoten das zweitbeste Ergebnis aller Erprobungsprojekte.

Die Konsequenz aus diesen Ergebnissen ist eindeutig: Die Notengebung im Fach Erdkunde war für die Erprobungsschüler der sieben Projekte unterschiedlich. Diese Unterschiede sind *nicht* (jedenfalls nicht allein) eine Folge aus der unterschiedlichen Zusammensetzung der Stichproben hinsichtlich der Schultypen. Einige Projekte wurden vielmehr anscheinend bei besonders strengen Lehrern (bzw. eher schlechten Schülern) erprobt, andere bei überdurchschnittlich guten Schülern (bzw. besonders mild zensierenden Lehrern). Dies ist ein weiterer Beleg dafür, daß von einer Repräsentativität der Schülerstichproben für die Projekte nicht gesprochen werden kann.

Betrachten wir nochmals die Aufgliederung der Erprobungen je Schultyp: Hier zeigt sich nämlich noch eine weitere, viel offensichtlichere Besonderheit der Erprobungsstichproben. Fast 68% der Erprobungen erfolgten an Gymnasien und nur knapp 9% an Hauptschulen – was mit Sicherheit nicht die wahre Bedeutung der Schularten widerspiegelt. Schlimmer noch: Einige Projekte wurden überhaupt nicht mit Berufsschülern erprobt, andere nicht an einer Gesamtschule.

Dies ist ein schwerer methodischer Fehler in der Anlage der Erhebung durch das RCFP.

Wenn es schon nicht gelingen konnte, die einzelnen Schultypen entsprechend ihrer Bedeutung (gemessen an der Schülerzahl) in den Erprobungsstichproben zu repräsentieren, dann hätte man zumindest bei jedem Projekt jeweils die *gleichen Anteile* an Hauptschülern, Gesamtschülern, Gymnasiasten usw. berücksichtigen müssen. Dann wären zwar die Schultypen nicht mehr repräsentativ in den Erprobungsstichproben, aber man könnte wenigstens die Projekte untereinander vergleichen. Völlig unverständlich ist es jedoch, daß das RCFP bei einigen Projekten bestimmte Schultypen überhaupt nicht in der Evaluation berücksichtigte. Es bedarf wohl keiner Diskussion, daß sich die Unterrichtssituation, das Anspruchsniveau, die Leistungsbereitschaft usw. je nach Schultyp unterscheiden. Oben haben wir nachgewiesen, daß sich die Notengebung je nach Schultyp unterscheidet. Durch die unterschiedliche Berücksichtigung der einzelnen Schularten bei den sieben Projekten tritt ein sog. „Interaktionseffekt" zwischen den Variablen „Projekt" und „Schultyp" auf. Dieser Interaktionseffekt führt bei einer *bivariaten* Analyse zu dem Dilemma, daß man nicht entscheiden kann, welcher Anteil bei den Unterschieden in der Notengebung auf das „Konto" der „Schulart" geht, und welcher Anteil aus der unterschiedlichen Berücksichtigung der Schultypen je „Projekt" resultiert. Die unter 7.1.1 und 7.1.2 vorgestellten *bivariaten* Zusammenhänge sind also nicht die „reinen" Zusammenhänge zwischen der „Erdkundenote und Schultyp" bzw. der „Erdkundenote und dem Projekt". Es handelt sich vielmehr jeweils um ein *Gemisch aus direkten und indirekten Zusammenhängen.*

Dies ist das Grundproblem *jeder* bivariaten Analyse bei nichtexperimenteller Anlage der Erhebung, und wir werden noch ausführlich darauf zu sprechen kommen. Hier bleibt jedoch festzuhalten, daß die Schultypen für jedes Projekt in den gleichen Proportionen hätten berücksichtigt werden müssen. Dies bezeichnet man als „Konstanthalten" der Variablen „Schultyp" in Bezug auf die Variable „Projekt".

Der mögliche Einwand: einzelne Unterrichtseinheiten ließen sich wegen ihres Inhalts nicht an allen Schularten erproben, trifft übrigens nicht zu. Die folgende Tabelle zeigt, daß bis auf die Einheit GAST alle Unterrichtseinheiten von den Autoren für alle Schultypen bestimmt gewesen waren[2].

Projekt	Klassen	Schultyp
FLUG	8-10	alle
RHEIN	8-10 (9-10)	Hau/Real/Gym/Ges
GELT	6-8	alle
BODEN	7-8	alle
BRAND	7-8	alle
GAST	Sekundarstufe II (11-13)	Gym/Ges
MOBI	8-10 bzw. 9-11	Hau/Real/Gym/Ges

Tab. 26: Die Schultypen, für welche die RCFP-Einheiten entwickelt wurden

7.3 Einstellungen zum Fach Erdkunde (Die Einstellungsbatterie)

Die Schülereinstellungen zum Fach Erdkunde sind ein weiteres Bündel von grundlegenden Untersuchungsvariablen. Sie wurden durch ein Polaritätsprofil und eine Einstellungsbatterie erhoben. Wie in den Abschnitten 4, 5 und 6 dargestellt, haben wir diese Meßinstrumente genauestens auf innere Konsistenz und Dimensionalität überprüft und in Form von Faktorenwerten (bzw. „factor-scores") verdichtet. Wir können uns deshalb hier auf die Betrachtung der Faktorenwerte zu den Einstellungsdimensionen beschränken. Da wir die Profilfaktoren später berücksichtigen werden, sollen hier nur die Einstellungsfaktoren: INTERESSE, SCHWIERIGKEIT, ANREGUNG, NÜTZLICHKEIT, BEDEUTUNG aus der Batterie behandelt werden.

Die Faktorenwerte zu diesen Einstellungsdimensionen liegen in Form zweier Werte vor: Einmal als *Mittelwert* über alle Faktorenwerte hinweg und zum anderen als *dichotomisierter Meßwert* pro Schüler. Wenn wir wissen wollen, welche Einstellungen *insgesamt gesehen* die Schüler zum Fach Erdkunde haben, ist es sinnvoller, den Mittelwert zu betrachten. Wollen wir dagegen aufzeigen, wie die Einstellungen für verschiedene Untergruppierungen aussehen, ist es angemessener, mit den dichotomisierten Faktorwerten zu arbeiten.

Über alle Schüler gemittelt ergibt sich, daß Erdkunde für
- mittelmäßig interessant,
- mittelmäßig schwierig,
- etwas anregend,
- ziemlich nützlich für außerhalb der Schule und
- für etwas unbedeutend im Vergleich zu anderen Schulfächern

gehalten wird.

Das zeigt sich, wenn man die jeweiligen *Mittelwerte* der Faktorwerte-„Skala" betrachtet (s. Abb. 9).

Das Ergebnis entspricht im wesentlichen dem bisher bekannten „Image" des Faches Erdkunde (vergl. *Fürstenbert/Jungfer* 1980). Leider ist dieses Ergebnis wenig aussagekräftig: Es fehlt der Bezug zu anderen Schulfächern. Ist Erdkunde interessanter als Geschichte, oder nützlicher als Physik, oder schwieriger als Deutsch? Wir können es nicht beantworten. Wie wir dies schon bei den Polaritätsprofilen beklagen mußten, so fehlt auch hier jegliche „Verortung" des Einstellungsraumes.

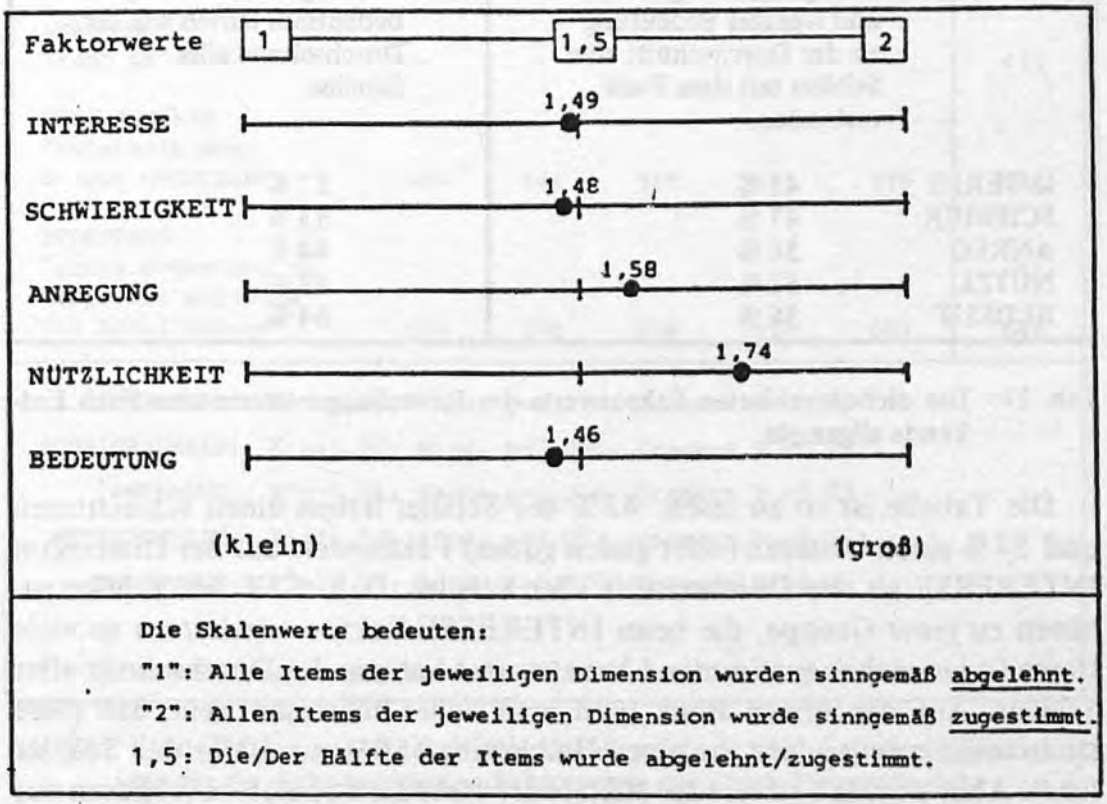


Abb. 9: Die Mittelwerte der Faktorenwerte-„Skala" aus der Einstellungsbatterie zum Fach Erdkunde

Wir können also keine absoluten Einstellungsunterschiede zu verschiedenen Schulfächern untersuchen – wohl aber *relative* Einstellungsunterschiede zum Fach Erdkunde für verschiedene Schülergruppen. Dabei werden wir jedoch nicht die Mittelwerte der jeweiligen Einstellungsfaktoren vergleichen, sondern die *dichotomisierten* Faktorwerte. Der Grund dafür ist, daß die Ursprungsitems aus der Einstellungsbatterie auch nur dichotomes Datenniveau hatten. D. h. wir schöpfen nur einen Teil der Information aus den Faktorwerten aus, sind also „vorsichtiger" und „konservativer" in Bezug auf das Niveau der Daten.

Die Dichotomisierung erfolgte so, daß jeder Schüler den Wert „2" bekam, der pro Dimension so viele oder mehr Items „zustimmend" beant-

wortete wie der Durchschnitt aller Schüler. Hatte der Schüler weniger Items zustimmend beantwortet als der Durchschnitt der Schüler, bekam er eine „1".

Das Ergebnis zeigt die folgende Tabelle 27:

	1 Anteil der Schüler, die *weniger* Interesse, mehr Schwierigkeit, weniger Anregung, weniger Nutzen und weniger Bedeutung als der Durchschnitt aller Schüler mit dem Fach verbinden.	2 Anteil der Schüler, die das Fach *mindestens* für so interessant, so leicht, so anregend, nützlich und bedeutsam halten wie der Durchschnitt aller Schüler.
INTERES	43 %	57 %
SCHWIER	47 %	53 %
ANREG	56 %	44 %
NÜTZL	52 %	48 %
BEDEUT	36 %	64 %

Tab. 27: Die dichotomisierten Faktorwerte der Einstellungsbatterie zum Fach Erdkunde allgemein

Die Tabelle ist so zu lesen: 43 % der Schüler haben einen schlechteren und 57 % einen besseren (oder gleich guten) Faktorwert auf der Dimension INTERESSE als der Durchschnitt aller Schüler. D. h. 57 % der Schüler gehören zu jener Gruppe, die beim INTERESSE-Faktor *mindestens* so viele Items (oder mehr) zustimmend beantwortet hat wie der Durchschnitt aller Schüler. Auf den ersten Blick wird man sicher überrascht sein, daß diese Dichotomisierung *nicht* zu einer Halbierung (50 % zu 50 %) der Schüler führt. Aber erstens halbiert der Mittelwert *nicht* die Population (sondern nur der Median), und zweitens wirkt sich die teilweise geringe Zahl der Items verzerrend aus.

Wir können dieses Problem im folgenden unbeachtet lassen, da uns nicht die „absolute" Zustimmung (oder Ablehnung) zu einer Einstellungsdimension interessiert, sondern nur die *relativen* Unterschiede bei verschiedenen Gruppierungen. Dafür aber spielt es keine wesentliche Rolle, wo wir die Trennungslinie für die Dichotomisierung hinlegen – solange sie inhaltlich vertretbar ist. Alle folgenden Analysen gehen von den dichotomen Einstellungsfaktoren aus.

7.3.1 Einstellungen und Schultyp

Schüler verschiedener Schularten finden Erdkunde gleich interessant. Das Fach regt sie auch gleich oft an, über andere Probleme nachzudenken. Die

Schwierigkeit des Faches wird dagegen unterschiedlich beurteilt. Ebenso seine Nützlichkeit und Bedeutung im Vergleich zu anderen Schulfächern:

Faktor	Haupt-schule	Gesamt-schule	Real-schule	Berufs-schule	Gymna-sium	Durch-schnitt
INTERESSE "ist interessant"	-*	-	-	-	-	57%
SCHWIERIGKEIT "ist schwierig"	65%**	63%	52%	44%	52%	53%
ANREGUNG "regt an"	-	-	-	-	-	44%
NÜTZLICHKEIT "außerhalb der Schule nützlich"	54%	44%	51%	56%	47%	48%
BEDEUTUNG "große Bedeutung im Vergleich mit anderen Schulfächern"	61%	69%	60%	62%	66%	64%

INTERESSE: $\chi^2 = 6{,}32$; df=4; p=0,176; Cramers V =0,03

SCHWIERIGKEIT: $\chi^2 = 42{,}88$; df=4; p=0,000; Cramers V =0,08

ANREGUNG: $\chi^2 = 5{,}81$; df=4; p=0,214; Cramers V =0,03

NÜTZLICHKEIT: $\chi^2 = 16{,}14$; df=4; p=0,003; Cramers V =0,06

BEDEUTUNG: $\chi^2 = 11{,}06$; df=4; p=0,026; Cramers V =0,05

* - bedeutet: kein signifikanter Unterschied! ** gerundet!

Tab. 28: Einstellungen zum Fach Erdkunde allgemein je nach Schultyp (Häufigkeiten für die dichotomisierten Faktorwerte der Einstellungsbatterie)

Im Mittel gehören (nach der von uns gewählten Dichotomisierung) 53% der Schüler zu jener Gruppe, die Erdkunde eher für schwierig halten. Bei den Hauptschülern sind es dagegen 65%, bei den Berufsschülern nur 44%. 48% der Schüler halten Erdkunde für „nützlich". 4% weniger sind es bei den Gesamtschülern, 8% mehr bei den Berufsschülern. Besonders häufig für „wichtig" halten die Gesamtschüler das Fach (+ 5%). Kleiner ist diese Gruppe bei den Realschülern (– 4%).

7.3.2 Einstellungen und Erdkundenote

Es ist nicht erstaunlich, *daß* ein Zusammenhang zwischen Erdkundenote und Einstellung zum Fach besteht. Überraschend jedoch ist die Enge des Zusammenhanges im Vergleich zu den bisher dargestellten bivariaten Zusammenhängen.

Faktor	Note 1	Note 2	Note 3	Note 4	Note 5,6	Durchschnitt
INTERESSE "ist interessant"	81,7	70,3	54,1	44,5	38,9	57,3
SCHWIERIGKEIT "ist schwierig"	40,0	44,2	54,5	62,9	69,7	53,2
ANREGUNG "regt an"	62,1	49,3	42,2	36,3	32,7	43,6
NÜTZLICHKEIT "außerhalb der Schule nützlich"	59,5	52,3	46,6	44,9	41,7	48,3
BEDEUTUNG "große Bedeutung im Vergleich mit anderen Schulfächern"	76,0	69,1	64,5	58,0	50,5	64,5

INTERESSE: χ^2=271,06; df=4; p=0,000; Cramers V =0,22
SCHWIERIGKEIT: χ^2=129,13; df=4; p=0,000; Cramers V =0,15
ANREGUNG: χ^2= 73,57; df=4; p=0,000; Cramers V =0,13
NÜTZLICHKEIT: χ^2= 29,05; df=4; p=0,000; Cramers V =0,08
BEDEUTUNG: χ^2= 49,59; df=4; p=0,000; Cramers V =0,11

Tab. 29: Einstellungen zum Fach Erdkunde je nach Erdkundenote (Häufigkeiten für die dichotomisierten Faktorwerte der Einstellungsbatterie)

Zwischen den Schülereinstellungen und den Erdkundenoten besteht sogar ein streng *monotoner* (wenn nicht gar ein linearer) Zusammenhang – und zwar bei *allen* Einstellungsdimensionen: Je besser die Erdkundenoten, desto höher ist grundsätzlich der Anteil der Schüler, die Erdkunde für interessant, leicht, anregend, nützlich und für bedeutend im Vergleich zu anderen Schulfächern halten.

Wir wollen dies für die beiden besonders wichtigen Dimensionen INTERESSE und SCHWIERIGKEIT nochmals graphisch darstellen.

7.3.3 Einstellungen und Geschlecht

Männliche Schüler finden Erdkunde interessanter[3], weibliche Schüler schwerer. Sonst besteht kein (wesentlicher) Unterschied in den Einstellungen. Auf diese Kurzformel kann man das Ergebnis der geschlechtsspezifischen Einstellungsunterschiede bringen (siehe Tab. 30).

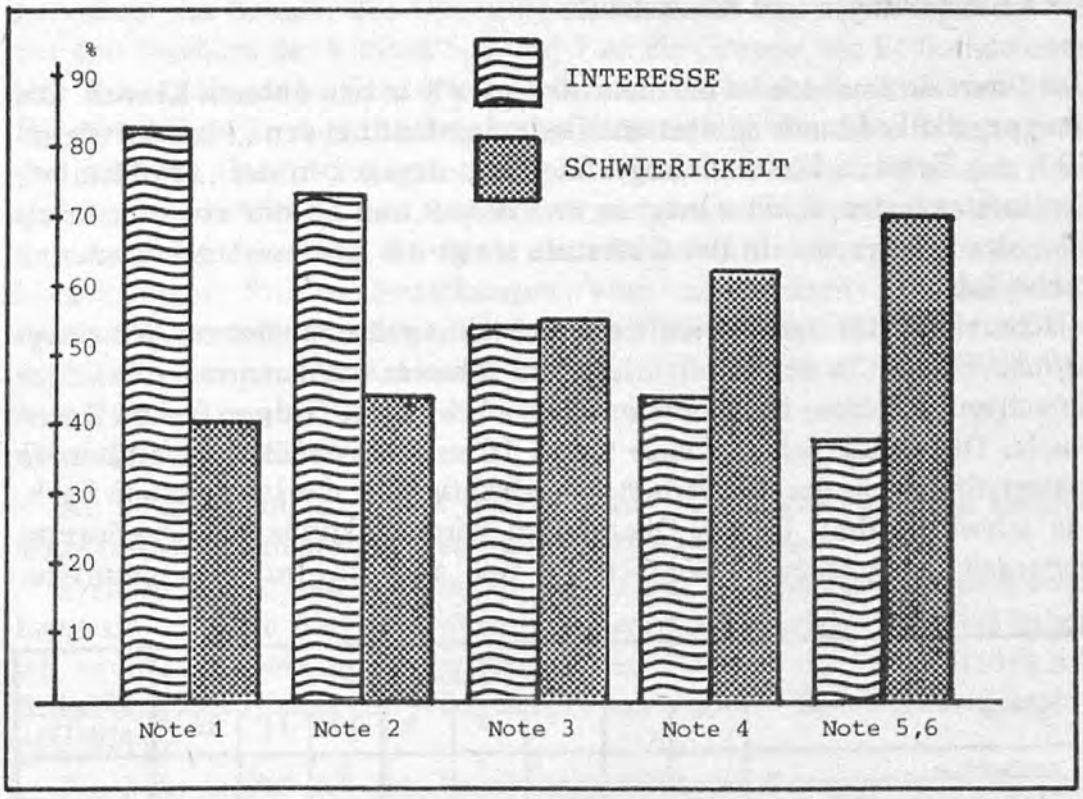


Abb. 10: Die Einstellungsdimensionen INTERESSE und SCHWIERIGKEIT je nach Erdkundenote (Häufigkeiten für die dichotomisierten Faktorenwerte)

Faktor	männl. Schüler	weibl. Schüler	Durch-schnitt
INTERESSE "ist interessant"	63,1	51,7	57,
SCHWIERIGKEIT "ist schwierig"	46,6	59,4	53,
ANREGUNG "regt an"	-*	-	44,
NÜTZLICHKEIT "außerhalb der Schule nützlich"	50,6	45,8	48,
BEDEUTUNG "große Bedeutung im Vergleich mit anderen Schulfächern"	-	-	64,

INTERESSE: χ^2 =71,15; df=1; p=0,000; PHI=0,12
SCHWIERIKGEIT: χ^2 =90,89; df=1; p=0,000; PHI=0,13
ANREGUNG: χ^2 = 0,17; df=1; p=0,680; PHI=0,006
NÜTZLICHKEIT: χ^2 =11,23; df=1; p=0,001; PHI=0,05
BEDEUTUNG: χ^2 = 0,23; df=1; p=0,630; PHI=0,007
* bedeutet: kein signifikanter Unterschied !

Tab. 30: Einstellungen zum Fach Erdkunde je nach Geschlecht (Häufigkeiten für die dichotomisierten Faktorwerte aus der Einstellungsbatterie)

7.3.4 Einstellungen und Klassenstufe

Das *Interesse* am Fach ist nie mehr so groß wie in den unteren Klassen. Die Gruppe, die Erdkunde interessant findet, umfaßt bei den „Fünftklässlern" 80% der Schüler. Ein eindeutiges Tief liegt dagegen in der „Mittelstufe". Die interessierten Schüler machen in Klasse 9 und 10 nur etwas mehr als 50% der Schüler aus. In der Oberstufe steigt das Interesse dann wieder an (siehe Tab. 31).

Komplementär dazu verläuft die Einstellung der Schüler zur *Schwierigkeit* des Faches: In der Unterstufe wird Erdkunde von den meisten Schülern für schwer gehalten, besonders in Klasse 7, wo diese Gruppe fast 60% ausmacht. Dann plötzlich in Klasse 9 und 10 sind die Schüler, die Erdkunde schwer finden, in der Minderheit. Gleichzeitig fällt das Interesse am Fach, wie schon erwähnt. In den Oberklassen wird Erdkunde wieder schwerer (jedenfalls nach Meinung der Schüler) und auch das Interesse nimmt zu.

Faktor	Klassenstufe 5	6	7	8	9	10	11	12	Durchschnitt
INTERESSE "ist interesant"	80,0	70,5	56,0	59,4	52,1	53,6	54,5	65,7	57,3
SCHWIERIGKEIT "ist schwierig"	56,7	52,6	58,8	53,4	47,1	48,5	60,7	65,7	53,2
ANREGUNG "regt an"	37,9	39,1	33,2	45,4	45,4	47,6	48,0	52,2	43,5
NÜTZLICHKEIT "außerhalb der Schule nützlich"	75,7	66,2	54,5	47,3	43,1	40,5	39,8	42,4	48,2
BEDEUTUNG "große Bedeutung im Vergleich mit anderen Schulfächern"	73,4	71,9	61,5	64,0	58,8	65,2	69,2	83,9	64,4

INTERESSE: χ^2 = 86,47; df=7; p=0,000; Cramers V =0,13
SCHWIERIGKEIT: χ^2 = 54,67; df=7; p=0,000; Cramers V =0,10
ANREGUNG: χ^2 = 51,23; df=7; p=0,000; Cramers V =0,11
NÜTZLICHKEIT: χ^2 =150,45; df=7; p=0,000; Cramers V =0,17
BEDEUTUNG: χ^2 = 45,30; df=7; p=0,000; Cramers V =0,10

Tab. 31: Einstellungen zum Fach Erdkunde je nach Klassenstufe (Häufigkeiten für die dichotomisierten Faktorwerte der Einstellungsbatterie)

Daß Erdkunde *anregt* (Faktor: ANREG), über einige Probleme weiter nachzudenken, sehen besonders die Schüler der Ober- und Mittelstufen, weniger dagegen die Schüler aus den Klassen 5, 6 und 7. Gerade umgekehrt verhält es sich mit der Einstellung der Schüler zur *Nützlichkeit* des Faches

außerhalb der Schule: Die Oberstufenschüler sind hier eher skeptisch. Nur bei den Schülern der Klassen 5, 6 und 7 ist die Gruppe, die Erdkunde nützlich findet, überdurchschnittlich groß, nämlich zwischen 76% und 55%. Die obige Tabelle zeigt die Ergebnisse im einzelnen. Da diese Resultate für den weiteren Gang der Argumentation im Rahmen unserer bivariaten Analyse besonders wichtig sind, wollen wir sie gleich noch einmal in anschaulicherer Form darstellen. Es wurde dazu für jede Einstellungsdimension die klassenweisen Prozentabweichungen vom allgemeinen (prozentualen) Durchschnittswert berechnet. Also z. B. so: Im Durchschnitt aller Schüler umfaßt die Gruppe, die Erdkunde interessant findet, 57,3% der Schüler. In Klasse 5 sind es aber 80,0% der Schüler, also 22,7% mehr. Es ergibt sich dabei die folgende Abbildung.

An dieser Abbildung wird deutlich, daß in den klassenweisen Einstellungsunterschieden gewisse Regelmäßigkeiten auftreten.

Erstens besteht eine auffallende *Parallelität* zwischen dem Wandel beim Interesse am Fach und der wahrgenommenen Schwierigkeit. Darauf haben wir bereits hingewiesen. Hinzu kommt, daß sich auch die Einschätzung der Bedeutung des Faches (im Vergleich zu den anderen Schulfächern) analog dazu ändert.

Zweitens ergibt sich eine (noch augenfälligere) *Komplementarität* zwischen den Einstellungsdimensionen: „Nützlichkeit" des Faches und „Anregung" durch das Fach. Die Schüler der unteren Klassen finden besonders häufig das Fach (auch für außerhalb der Schule) nützlich, gleichzeitig aber seltener anregend. Je älter die Schüler werden, desto mehr dreht sich dieser Zusammenhang um: In der Oberstufe wird das Fach häufiger für anregend gehalten, aber seltener für nützlich.

Wir wollen diese beiden Erscheinungen nun noch etwas genauer analysieren.

(1) Parallele Einstellungsveränderungen während der Schulzeit

Die auffällige Parallelentwicklung zwischen dem Interesse am Fach, der perzipierten Schwierigkeit des Faches und seiner wahrgenommenen Bedeutung im Vergleich zu anderen Schulfächern gibt einige Rätsel auf.

So könnte man vermuten, daß ein allzu leicht empfundenes Schulfach nicht besonders interessant sein kann. Dies wird bestätigt durch die Tatsache, daß in der 9. Klassenstufe das Fach Erdkunde am seltensten für schwierig und gleichzeitig auch am seltensten für interessant gehalten wird. Unerklärlich bleibt dann aber, warum auf Klassenstufe 7 zwar ein Interesse am Fach nur *unter*durchschnittlich vorhanden ist, während der Anteil jener Schüler, die das Fach für schwierig halten, deutlich *über* dem Durchschnitt liegt.

Eine Erklärung dafür könnte sein, daß das Interesse am Fach weniger mit dem Anspruchsniveau (= perzipierte Schwierigkeit) zusammenhängt als

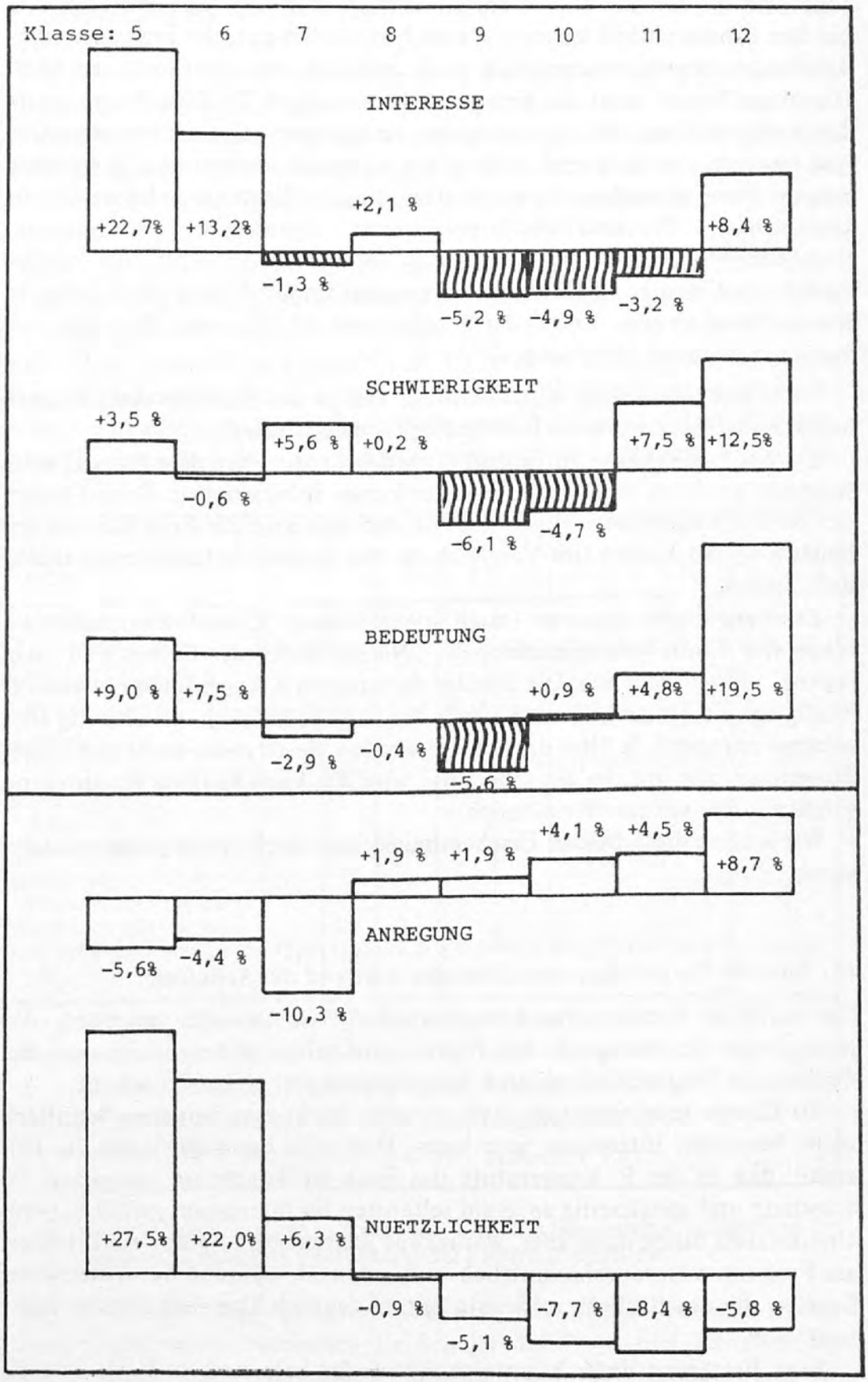


Abb. 11: Parallele und komplementäre Einstellungen zum Fach Erdkunde je nach Klassenstufe

vielmehr mit der Bedeutung, die das Fach im Rahmen der anderen Schulfächer hat. Die klassenweisen Veränderungen im Interesse der Schüler verlaufen auch tatsächlich parallel zu der vermuteten Bedeutung des Faches in bezug auf die anderen Schulfächer (Faktor BEDEUTUNG).

Dies zeigt nochmals besonders deutlich die Abbildung unten: Sie entspricht genau den Mittelwertsabweichungen aus der Abbildung auf der vorigen Seite. Nur wurden hier die Prozentsätze je Klasse durch eine Linie verbunden, was die optische Beurteilung der „Verlaufsgestalten" der Einstellungen über die verschiedenen Klassenstufen erleichtert.

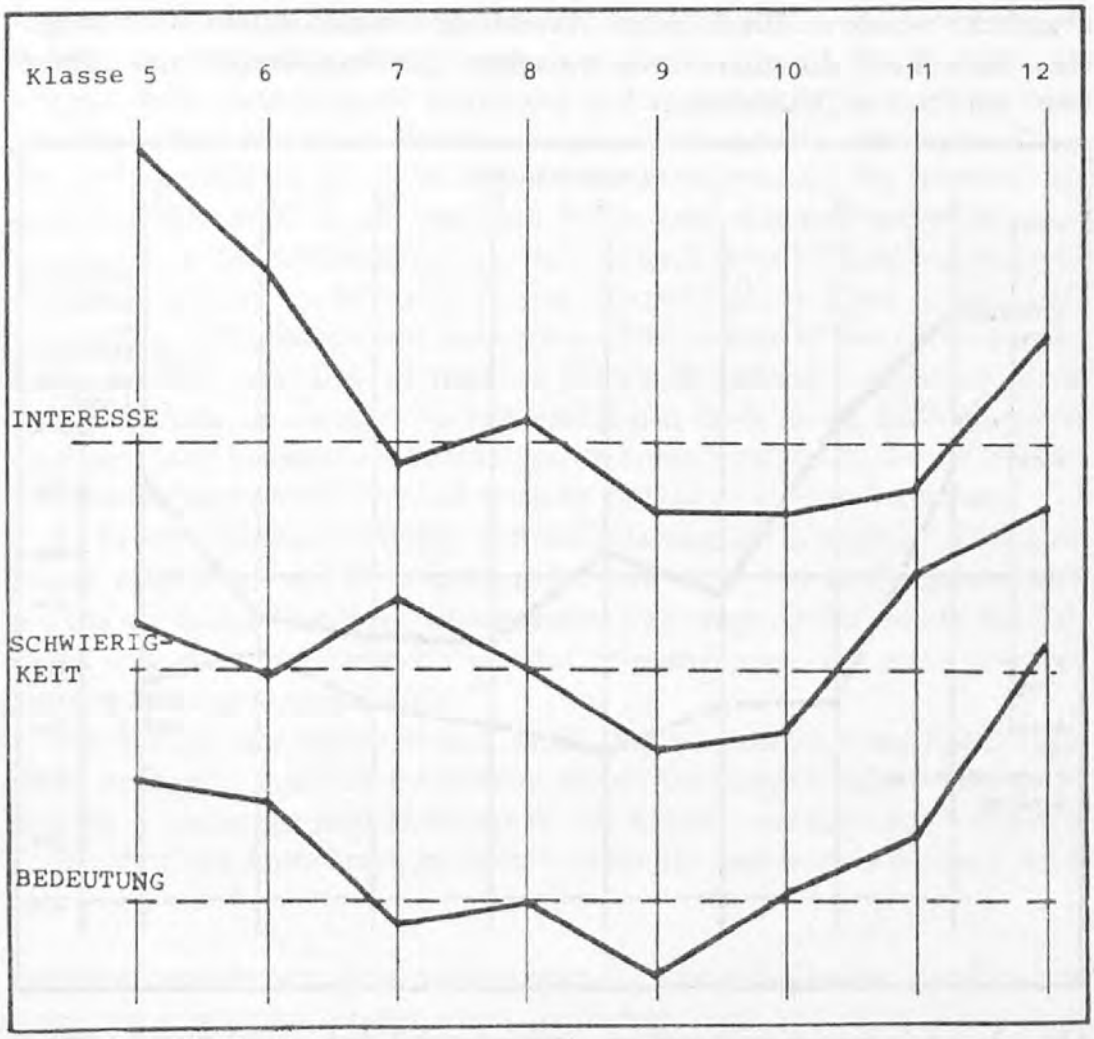


Abb. 12: Veränderungen der Einstellungsdimensionen INTERESSE, SCHWIERIGKEIT und BEDEUTUNG je nach Klassenstufe

Halten wir also fest: Es besteht eine augenfällige Parallelität zwischen dem Interesse am Fach Erdkunde und der vermuteten Bedeutung dieses Faches in Bezug auf die anderen Schulfächer: In jenen Klassenstufen, in denen Erdkunde nach Meinung der Schüler als Fach eine große Bedeutung hat, wird es auch häufiger als interessant empfunden.

Können wir uns mit dieser „Erklärung" schon zufrieden geben? Zweifellos wäre dies ein Trugschluß: Denn erstens korreliert[4] das Interesse am Fach Erdkunde z. B. gleichzeitig auch hoch mit der Erdkundenote – wie wir bereits wissen. Und die Note wiederum variiert mit der Klassenstufe. Die klassenweisen Unterschiede im Interesse *und* in der vermuteten Bedeutung des Faches könnten also beide eine Folge der unterschiedlichen Praxis der Notenvergabe in den einzelnen Klassenstufen sein. Zweitens dürften noch eine Reihe weiterer Variablen einen indirekten Einfluß auf das Interesse je nach Klassenstufe haben, wie z. B. das Geschlecht und die Stichprobenpopulation je RCFP-Einheit (wo nicht alle Klassenstufen gleichmäßig abgedeckt wurden). Die folgende Abbildung veranschaulicht einen möglichen indirekten Zusammenhang zwischen „Erdkundenote" und „Interesse" am Fach je „Klassenstufe":

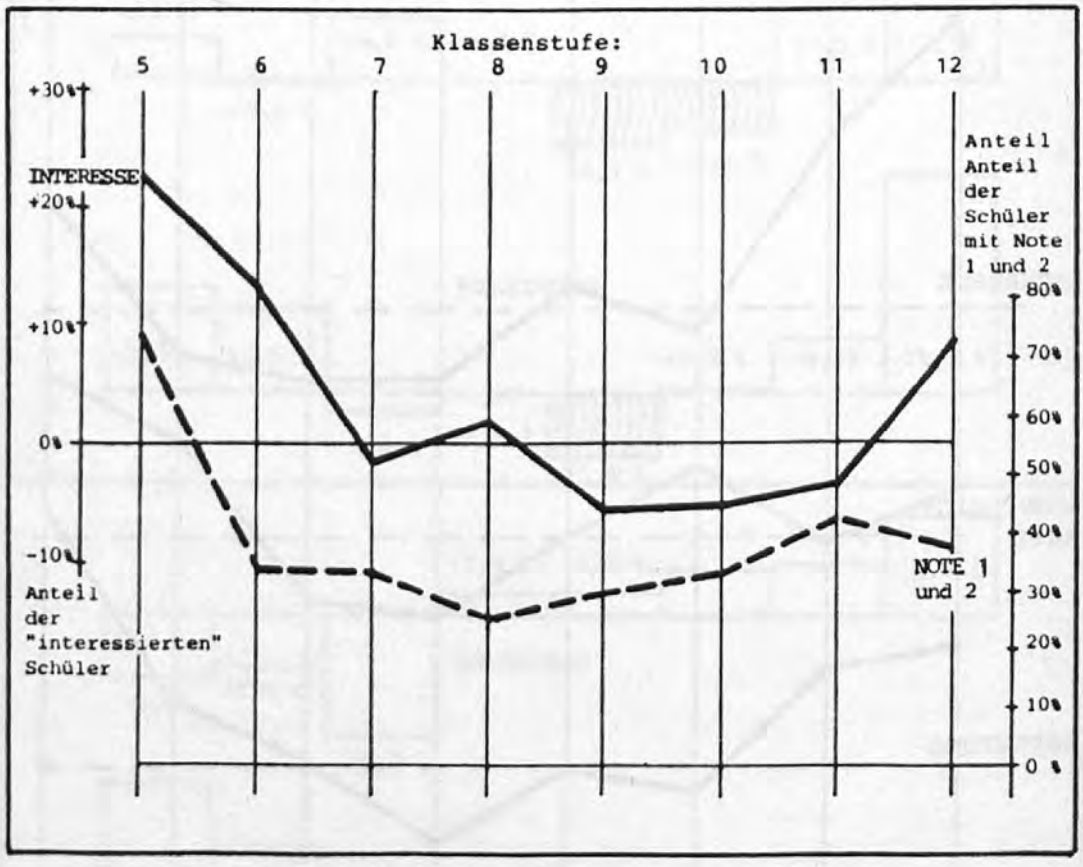


Abb. 13: Veränderunge im Anteil an „interessierten" und „(sehr) guten" Schülern je Klassenstufe

Wie man sieht, erhalten wir hier einen Zusammenhang wie aus einem Lehrbuch: Der Anteil der am Fach interessierten Schüler steigt und fällt in den einzelnen Klassenstufen fast in gleicher Weise wie der Anteil jener Schüler, die auf dieser Klassenstufe gute oder sehr gute Erdkundenoten haben. Die klassenweisen Schwankungen im Interesse der Schüler am Fach Erdkunde können also mehrere (!) Ursachen haben:

- Reifungsprozesse, wie sie in bestimmten Klassenstufen auftreten,
- Veränderungen in der Schwierigkeit des Faches,
- Veränderungen in der Bedeutung des Faches, verglichen mit den anderen Schulfächern,
- unterschiedliche Praxis der Notenvergabe in bestimmten Klassenstufen, oder
- alles zusammen (in verschiedenen Kombinationen).

Kann man eine dieser Variablen als hauptverantwortlich bezeichnen? Reifungsprozesse werden von „Praktikern" im Schuldienst gern als die wichtigste Variable angesehen, um die Schwankungen im Interesse der Schüler (je nach Klassenstufe) zu „erklären". Für das Fach Erdkunde stimmt diese monokausale Erklärung sicher nicht. Sehen wir uns dazu nochmals die Abb. 13 an: Schüler aus der Klasse 5 hatten im letzten Zeugnis im Durchschnitt zu *72 % eine 1 oder 2* bekommen. Ihr Interesse am Fach war sehr groß. In der nächsten Klasse (also Klasse 6) schwächte sich das Interesse der Schüler *etwas* ab. Befragt nach ihrer Erdkundenote gaben nur noch *33 %* an, sie hatten im *letzten* Zeugnis (also in Klasse 5) eine gute oder sehr gute Erdkundenote bekommen. Wieder eine Klasse weiter kommt der große Einbruch: Das Interesse am Fach Erdkunde sinkt drastisch. Diese Schüler hatten im Zeugnis vorher wieder nur noch zu ca. 30 % eine gute oder sehr gute Erdkundenote erhalten, zu einem Zeitpunkt, wo die Schüler (üblicherweise) ja noch ziemlich stark an Erdkunde interessiert waren.

Es ist also durchaus möglich, daß das Interesse der Schüler am Fach Erdkunde eine *Folge* der Notengebung ist, und nicht von Reifungsprozessen (erfaßt durch die Variable „Klassenstufe") abhängt. Dafür spricht die Tatsache, daß die „Interessen-Kurve" der „Noten-Kurve" mit einer gewissen *Zeitverzögerung* hinterherhinkt.

Die Schüler der Klasse 6 sind noch relativ *interessiert* am Fach – *obwohl* sie bereits deutlich *schlechtere* Noten bekommen haben. Erst nachdem sie – sozusagen zum zweitenmal – in Klasse 7 wieder relativ schlechte Erdkundenoten hinnehmen mußten – sinkt ihr Interesse in Klasse 7 deutlich.

Fassen wir zusammen: Interpretiert man die uns vorliegenden Querschnittsdaten wie Längsschnittdaten, dann „verändern" sich bestimmte Einstellungen von Schülern in augenfällig paralleler Weise im Lauf der Schulzeit. Es sind das *Interesse* am Fach Erdkunde, die Einschätzung der *Schwierigkeit* des Faches und die Beurteilung seiner *Bedeutung* im Vergleich mit den anderen Schulfächern.

Die (fast parallelen) „Verlaufskurven" dieser Einstellungen über die einzelnen Klassenstufen haben einen U-förmigen Verlauf, d. h. sie kennzeichnen einen *nicht*-linearen Zusammenhang zwischen den Variablen „Klassenstufe" und „Einstellung". In mittleren Klassenstufen erreicht das Interesse am Fach, die Bedeutung des Faches und sein Anspruchsni-

veau ein Minimum, während in unteren und oberen Klassenstufen wesentlich positivere „Werte" erreicht werden.

Der parallele Wandel dieser Einstellungen kann nicht *allein* auf Reifungsprozesse zurückgeführt werden, wie sie in der Variablen „Klassenstufe" operationalisiert sind. Erstens könnten Zusammenhänge zwischen den Einstellungsdimensionen dafür verantwortlich sein. Zweitens könnte die Notengebung im Fach Erdkunde alle drei Dimensionen in gleicher Weise beeinflußen. Drittens könnten andere strukturelle Variablen (wie Geschlecht und Schultyp) sowohl die Notengebung als auch die Stichprobenzusammensetzung der einzelnen Klassenstufen und letztlich dadurch auch die klassenweise parallelen Einstellungsveränderungen bewirken. Die folgende Skizze verdeutlicht diese komplexen Zusammenhänge:

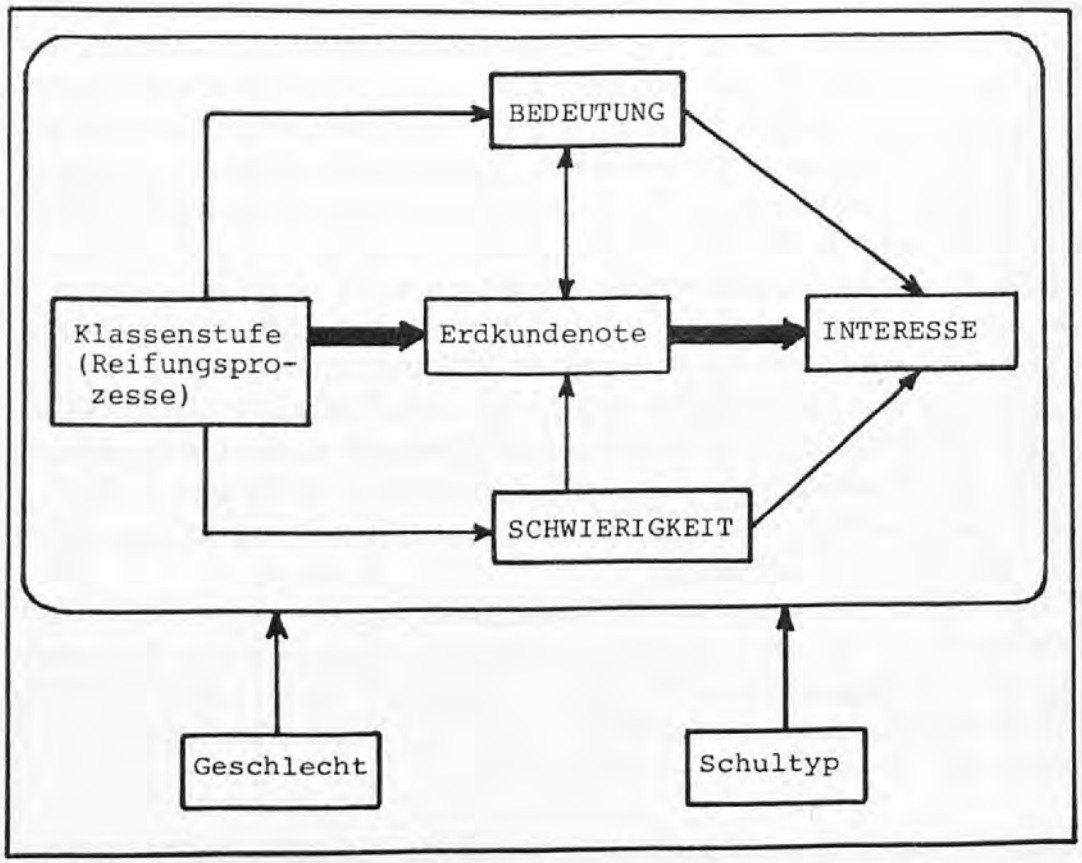


Abb. 14: Mögliche Zusammenhänge zwischen verschiedenen Einstellungsvariablen und der Klassenstufe, bzw. der Erdkundenote

(2) Komplementäre Einstellungsveränderungen während der Schulzeit

Wir wollten uns noch ein weiteres Phänomen in den klassenweisen Einstellungsveränderungen genauer ansehen: die gegenläufige Veränderung in den Einstellungen der Schüler zur Nützlichkeit des Faches Erdkunde und zur Anregung, die aus diesem Schulfach kommt. Sehen wir uns dazu zunächst wieder die Verlaufskurven an, die aus Abb. 11 entnommen sind (s. Abb. 15).

Besonders in den oberen Klassen regt der Erdkundeunterricht die Schüler an, „über einige seiner Themen weiter nachzudenken". Liegt dies daran, daß jene Schüler einfach reifer sind, um sich eigene Gedanken zu machen? Oder liegt es vielleicht daran, daß in der Oberstufe der Unterricht in Erdkunde immer esoterischer, immer weniger konkret verwertbar wird? Die Schüler der Oberstufe jedenfalls glauben immer weniger daran, daß in Erdkunde Dinge gelernt werden, die „man später im Leben brauchen kann". Die Frage ist hier nicht zu entscheiden. Fest steht aber eines: Würde ein fiktiver „Durchschnittsschüler" die Klassen 5 bis 12 der vorliegenden Erhebung durchlaufen, würde sich sein Bild vom Fach Erdkunde ziemlich bedenklich verändern. Er müßte quasi zu folgender Auffassung kommen: Je interessanter und anregender die Inhalte in diesem Schulfach sind, desto weniger kann man mit ihnen außerhalb der Schule anfangen. Dies ist beileibe kein schmeichelhaftes Ergebnis für den „Realitätsgehalt" eines Schulfaches. Offenbar wird Erdkunde in den Oberklassen immer mehr zu einem „weichen" Schulfach, das der Konkurrenz der „harten" Fächer (wie Mathematik oder Physik) nicht gewachsen ist.

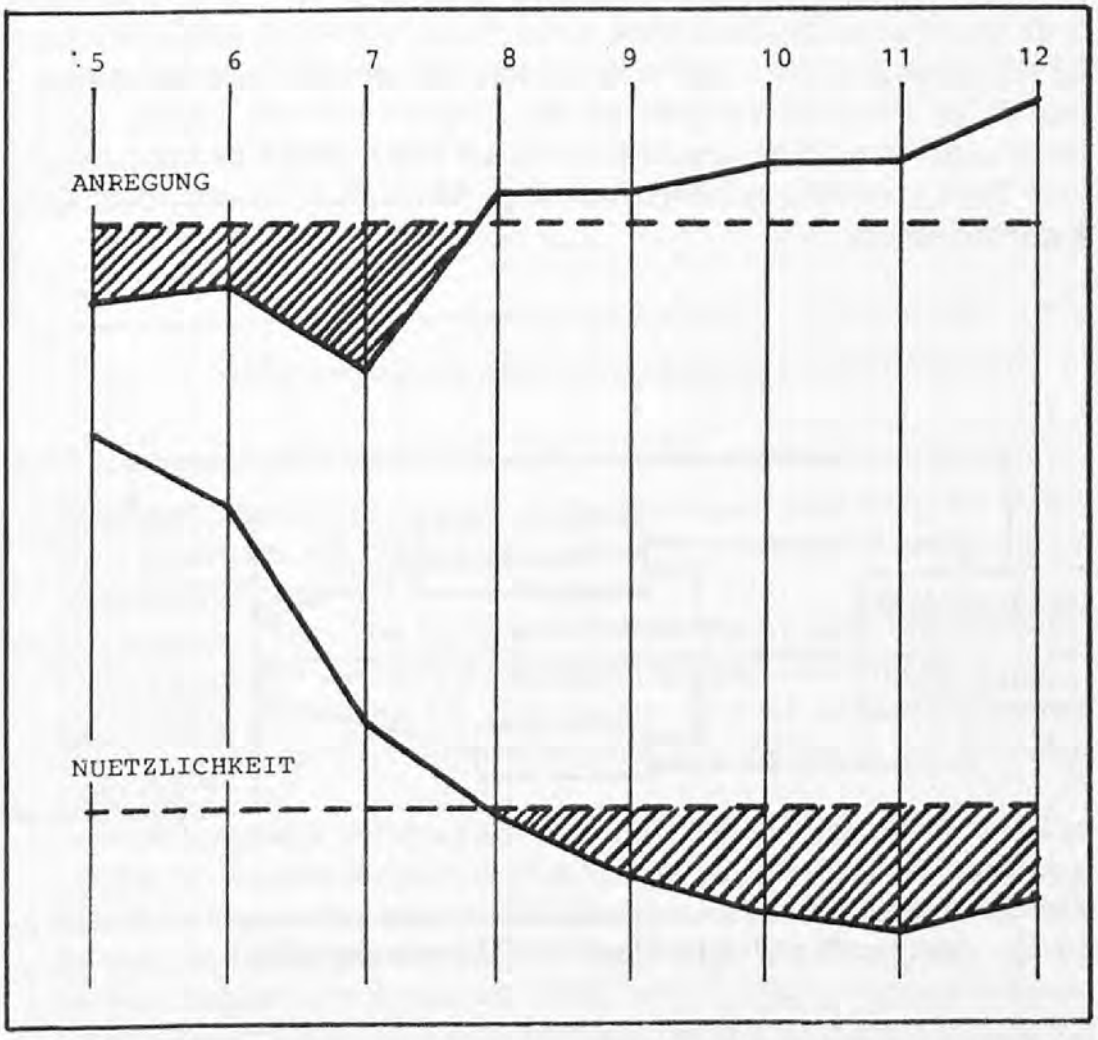


Abb. 15: Gegenläufige Verlaufskurven der Einstellungsdimensionen ANREGUNG und NÜTZLICHKEIT je nach Klassenstufe

Die Notengebung hat auf diesen „Verfall" der Nützlichkeit des Faches *keinen* Einfluß. Auch das Geschlecht dürfte keinen Effekt haben; zwar finden männliche Schüler Erdkunde durchweg nützlicher als Mädchen, aber die Geschlechterproportionen auf den einzelnen Klassenstufen ist ungefähr gleich (bis auf einzelne Ausreißer-Klassen).

Nur die Variable Schultyp ist zu beachten: Berufs- und Hauptschüler halten Erdkunde für nützlicher als Gymnasiasten. Diese sind aber in den Oberklassen ausschließlich vertreten. Also könnte die anscheinend gering eingestufte Nützlichkeit des Faches in der Oberstufe bloß ein Effekt der Schulart sein, und weder durch Reifungsprozesse noch durch Veränderungen im „Charakter" des Faches bedingt sein.

7.3.5 Einstellungen und Projekt

Wie wir bereits wissen, unterscheiden sich die Erprobungspopulationen der einzelnen Projekte in ihrer Erkundenote. Außerdem sind die verschiedenen Schultypen unterschiedlich stark vertreten. Da die Einstellungen der Schüler zum Fach Erdkunde aber auch mit der Erdkundenote zusammenhängen, und da sie ebenfalls teilweise mit der Schulart variieren, müßten die Erprobungsschüler der verschiedenen Projekte unterschiedliche Einstellungen dem Fach gegenüber gehabt haben. Eine kleine Skizze verdeutlicht diese Zusammenhänge:

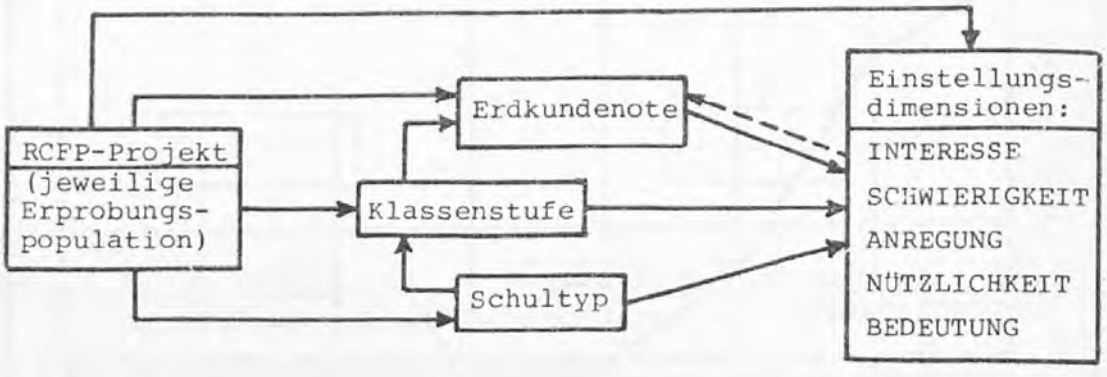


Abb. 16: Mögliche direkte und indirekte Zusammenhänge zwischen den Schülereinstellungen und dem jeweiligen RCFP-Erprobungsprojekt.

Sehen wir uns nun das empirische Ergebnis an:

Faktor	FLUG	RHEIN	GELT	BODEN	BRAND	GAST	MOBI	Durchschnitt
INTERESSE "ist interessant"	61%*	54%	[64%]	60%	56%	57%	[47%]	57%
SCHWIERIGKEIT "ist schwierig"	48%	[46%]	55%	55%	[60%]	50%	57%	53%
ANREGUNG "regt an"	[51%]	47%	35%	40%	[29%]	48%	45%	44%
NÜTZLICHKEIT "außerhalb der Schule nützlich"	48%	[40%]	[59%]	48%	56%	41%	41%	48%
BEDEUTUNG "große Bedeutung im Vergleich mit anderen Schulfächern"	63%	66%	68%	[61%]	65%	[70%]	[61%]	64%

INTERESSE: χ^2 =53,49; df=6; p=0,000; Cramers V =0,099

SCHWIERIGKEIT: χ^2 =51,88; df=6; p=0,000; Cramers V =0,096

ANREGUNG: χ^2 =49,51; df=6; p=0,000; Cramers V =0,103

NÜTZLICHKEIT: χ^2 =97,68; df=6; p=0,000; Cramers V =0,141

BEDEUTUNG: χ^2 =15,45; df=6; p=0,000; Cramers V =0,058

* gerundet!

Tab. 32: Schülereinstellungen zum Fach Erdkunde je nach RCFP-Projekt

Sehr vereinfacht gesprochen kann man folgendes sagen:

- Besonders interessiert am Fach Erdkunde waren die Schüler der Erprobung der Einheit GELTinger Bucht. Weniger interessiert waren die Schüler der Einheit Innenstädtische MOBIlität.
- Für besonders *schwierig* hielten die Schüler der BRAND-Erprobung das Fach, für weniger schwierig die Schüler der RHEIN-Erprobung.
- Die Schüler der Einheit BRAND fanden das Fach auch am seltensten anregend. Die der FLUG-Erprobung dagegen hielten es für besonders anregend.
- Besonders *nützlich* war das Fach für die Schüler der Erprobung GELTinger Bucht, weniger nützlich für die Schüler aus der RHEIN-Erprobung.
- Im Vergleich mit den anderen Schulfächern hielten die Erprobungsschüler der GAST-Einheit das Fach Erdkunde häufiger für *wichtig*, die Erprobungsschüler der Einheiten MOBI und BODEN dagegen seltener. Dieser Zusammenhang ist jedoch vergleichsweise schwach (Cramers V = 0,058).

Wichtiger als die obigen (deskriptiven) Details ist wieder eine andere Frage: Sind diese Unterschiede in den Schülereinstellungen je Erprobungspopulation eine (indirekte) Folge der ungleichgewichtigen „Vertretung" der Schultypen und der unterschiedlichen Praxis der Notengebung?

Sehen wir uns dazu einen konkreten Fall genauer an: Die Erprobungsschüler der Einheit MOBIlität. Sie waren diejenigen, die weitaus die schlechtesten Erdkundenoten bekommen hatten. Schlechte Noten im Fach Erdkunde stehen im Zusammenhang mit einem geringerem Interesse am Fach. Wir folgern daraus, daß die MOBIlitäts-Schüler weniger interessiert waren als andere Schüler. Ein Blick auf die Tab. 32 bestätigt uns dies.

Die Einstellungen der Schüler unterscheiden sich – wie wir wissen – von Schultyp zu Schultyp. Hauptschüler finden Erdkunde überdurchschnittlich *schwierig*. Wir wissen außerdem, daß bei der Erprobung der Einheit GELTinger Bucht weitaus mehr Hauptschüler beteiligt waren als bei den anderen Einheiten. Unsere Folgerung – daß diese Schülerpopulation Erdkunde schwierig finden müßte – wird aber *nur schwach* bestätigt (s. Tab. 32).

Wir können also folgende Schlußfolgerung ziehen: Die Schülerpopulationen der 7 Erprobungsprojekte unterscheiden sich in ihren Einstellungen zum Fach Erdkunde. Dies ist vermutlich *teilweise* auch eine Folge der unterschiedlich vertretenen Schultypen und der unterschiedlichen Notengebung im Fach Erdkunde.

7.4 Einstellungen zur jeweiligen RCFP-Unterrichtseinheit

Die Einstellungen der Schüler zu den Erprobungseinheiten des RCFP wurden durch folgende Variablen erhoben:

- Durch die Variable: „Einschätzung von Spaß und Nutzen" der Einheit (bzw. teilweise ihrer Unterabschnitte).
- Durch die Variable: „Interesse" an der Einheit allgemein.
- Durch ein „Polaritätsprofil" zur Erprobungseinheit.

7.4.1 Spaß und Nutzen je nach RCFP-Projekt

Betrachtet man die SPASS- und NUTZEN-Einschätzungen der Schüler je nach RCFP-Projekt (s. Abb. 17), dann zeigen sich außerordentlich deutliche Unterschiede. Leider gab es bei den beiden Variablen SPASS und NUTZEN Probleme mit zu kleinen Fallzahlen (und das bei 5668 Fällen!). Da die Schüler bei bestimmten Projekten offenbar äußerst unwillig waren nach der (mißglückten?) Erprobung noch einen Fragebogen auszufüllen, traten bei diesen Variablen erhebliche Ausfälle (missing values) auf. Das Projekt MOBI mußte aus diesem Grund sogar völlig ausgeschlossen werden, da nur ein Bruchteil der Schüler geantwortet hatte.

Betrachten wir nun die Ergebnisse im einzelnen: Zunächst zum SPASS an der jeweiligen Unterrichtseinheit:

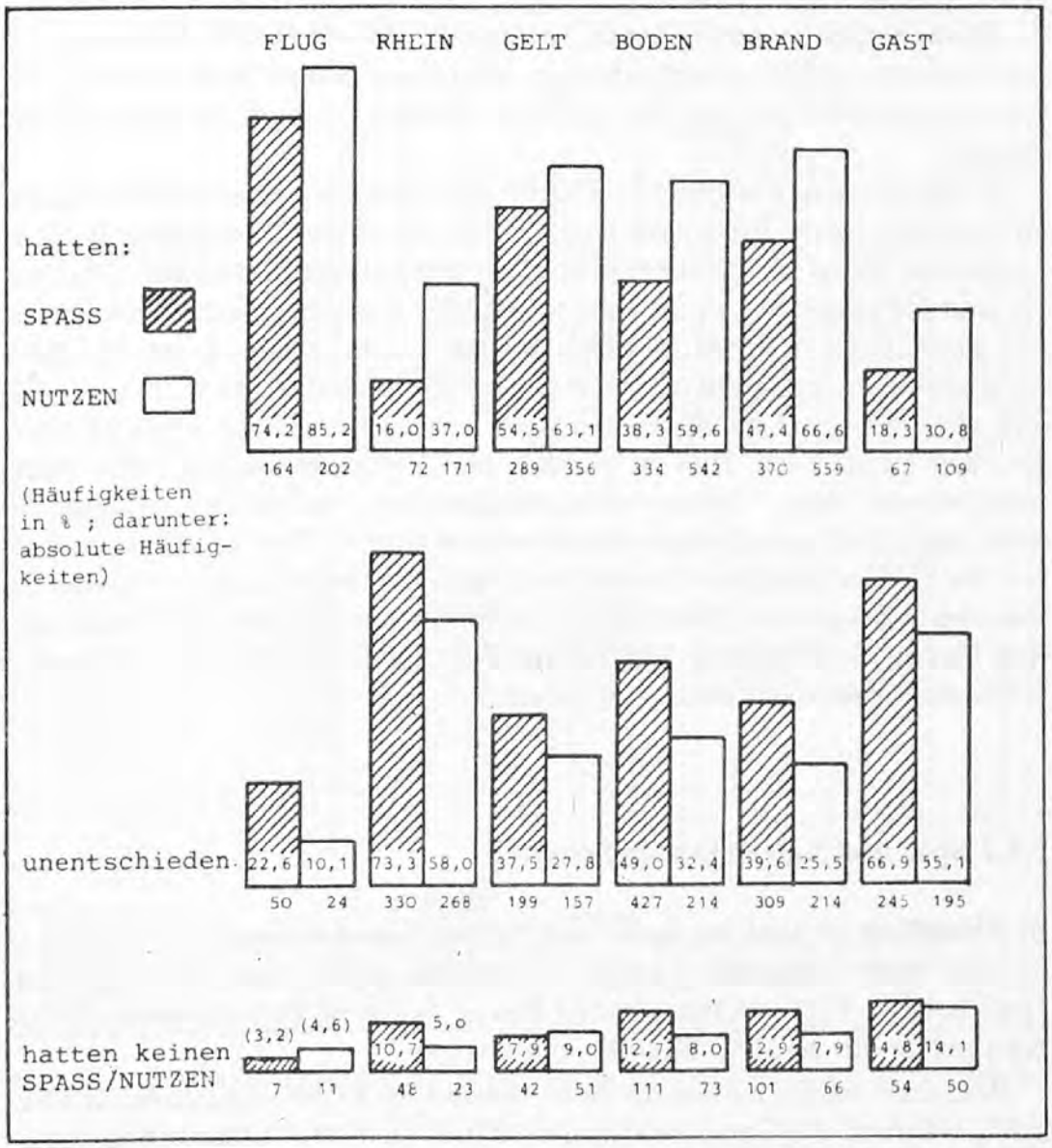


Abb. 17: Spaß- und Nutzen-Einschätzung der Erprobung je nach RCFP-Projekt.

Bei der Erprobung der Einheit FLUG hatten die Schüler am häufigsten Spaß: Fast dreiviertel der Schüler (74 %) geben dies an. Der Rest der Schüler war unentschieden (23 %). Nur 3 % der Schüler stellten ausdrücklich fest, daß sie *keinen* Spaß gehabt hätten. Ohne Zweifel war es die Einheit FLUG, die bei den Schülern die Freude am Lernen am stärksten gefördert hat.

Das Projekt RHEIN konnte den Schülern dagegen keine große Begeisterung entlocken: Nur 16 % meinten, sie hätten Spaß gehabt und 73 % konnten sich nicht entscheiden. Der sehr hohe Prozentsatz unentschlossener Schüler läßt darauf schließen, daß diese Einheit widersprüchliche Empfin-

dungen bei den Schülern auslöst. Wir werden später sehen, daß dies tatsächlich der Fall war: Hohe Erwartungen an die Thematik der Einheit konnten – nach Ansicht der Schüler – im Erprobungsunterricht nicht erfüllt werden.

Ähnlich negativ beurteilten die Schüler das Projekt GAST. Hier gaben sogar 15% der Schüler ausdrücklich an, sie hätten *keinen* Spaß gehabt. 67% waren unentschieden und nur 18% der Schüler meinten, sie hätten Spaß gehabt.

Vergleicht man die Projekte RHEIN und GAST mit dem Projekt FLUG, so wird man sagen dürfen, daß derart deutliche Unterschiede in den Schülermeinungen kaum auf Besonderheiten der Erprobungssituation (z. B. unterschiedlich engagierte Lehrer, unterschiedliche Zusammensetzung der Erprobungspopulation nach Alter, Schultyp oder Leistungsniveau, usw.) zurückzuführen sind – jedenfalls nicht *nur* darauf. Die Einheit FLUG hatte den Schülern aber nicht nur am häufigsten Spaß gemacht: 85% der Schüler hielten sie auch für nützlich. Dies ist wirklich ein hervorragendes Ergebnis, wenn man bedenkt, daß – lernpsychologisch gesehen – subjektive Erfolgserlebnisse kaum hoch genug eingeschätzt werden können. Bei dieser Einheit hatten die Schüler nicht nur Freude am Lernen, sie fanden auch, sie hätten etwas nützliches gelernt. Dies ist per se schon positiv zu bewerten. Am seltensten hielten die Schüler die Einheit GAST für nützlich (nur 30%), 14% meinten sogar, sie sei nicht nützlich gewesen.

7.4.2 Spaß und Nutzen je nach Schultyp

In Abbildung 18 sind die Spaß- und Nutzen-Einschätzungen der Schüler je nach Schulart dargestellt. Es zeigt sich, daß besonders Hauptschüler großen Spaß hatten: 52% von ihnen gaben dies an, während beispielsweise die Gesamtschüler nur zu 35% dieser Meinung waren.

Besonders häufig hielten die Berufsschüler die RCFP–Einheiten für nützlich[5], während die Gesamtschüler am seltensten einen Nutzen in dem jeweiligen Projekt sahen.

Vergleicht man die Realschüler mit den Gymnasiasten, so fällt auf, daß sie ähnlich oft Spaß an den Projekten hatten, wobei die Realschüler den Nutzen deutlich höher bewerteten als die Gymnasiasten.

7.4.3 Spaß und Nutzen je nach Klassenstufe

Deutliche Unterschiede ergeben sich auch, wenn man die Spaß- und Nutzen-Einschätzungen der Schüler klassenweise vergleicht: (s. Abb. 19). Es zeigen sich dabei zwei auffällige Tendenzen, und zwar sowohl für die Spaß- als auch für die Nutzen-Einschätzung:

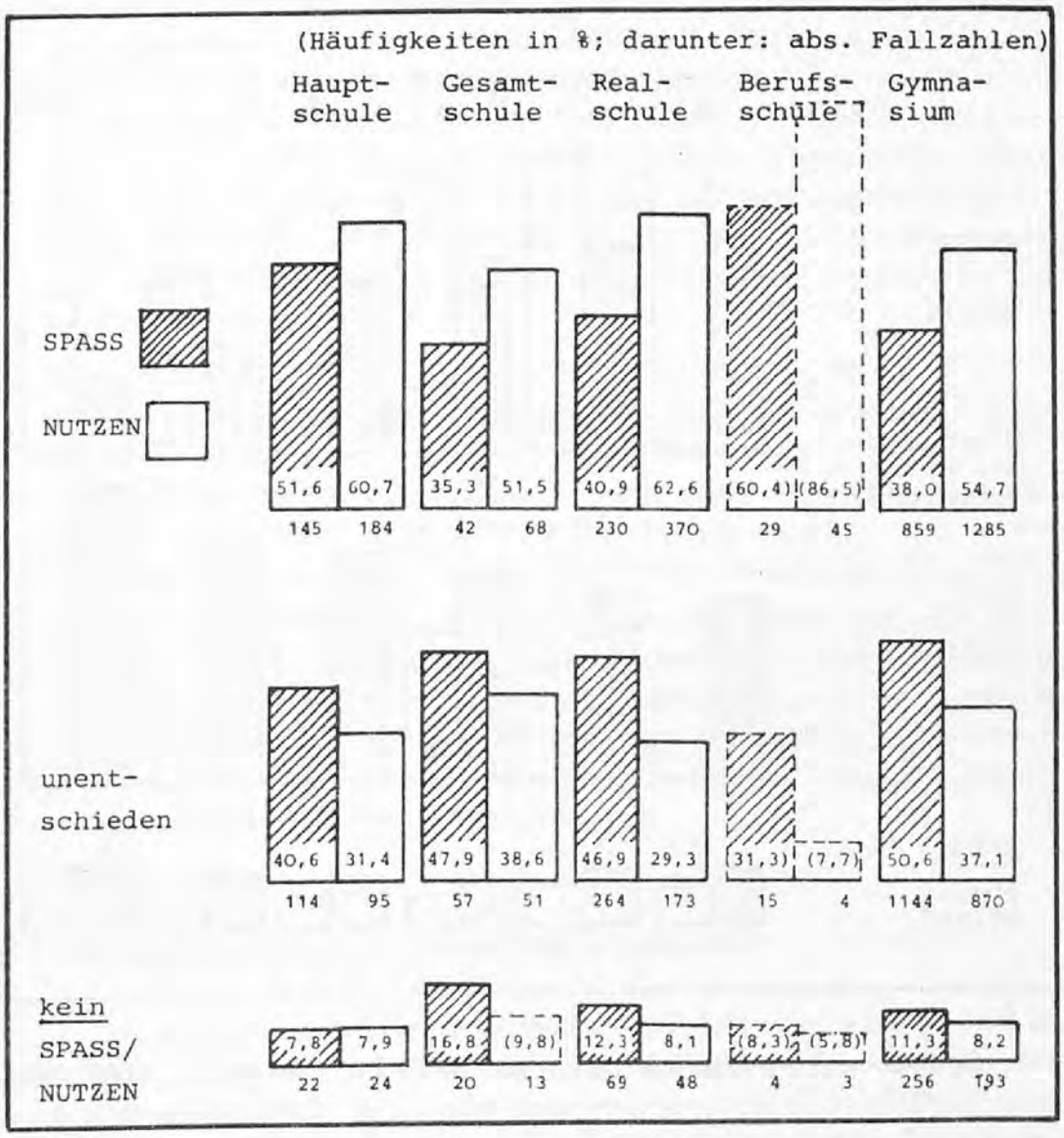


Abb. 18: Spaß- und Nutzen-Einschätzung (der RCFP-Erprobungen) je nach Schultyp

Erstens nimmt der Anteil der Schüler, die Spaß hatten, und die die Einheit nützlich fanden, kontinuierlich ab. In Klasse 6 hatten fast 50% Spaß an der Erprobung, und fast 64% fanden sie nützlich. In Klasse 11 dagegen hatten nur noch 18% der Schüler Spaß und nur noch 36% der Schüler fanden die Einheit nützlich.

Zweitens wird der Anteil unentschlossener Schüler mit höherer Klassenstufe immer größer. In den Klassen 6 konnten sich 40% nicht entscheiden, ob ihnen die Erprobung eher Spaß gemacht hat oder nicht, und 28% waren sich darüber unschlüssig, ob die Einheit eher nützlich war oder nicht. In der Klassenstufe 11 dagegen konnten oder wollten sich fast 67% der Schüler nicht entscheiden, ob die Erprobung Spaß gemacht hat, und über 50% der Schüler waren sich in ihrer Nutzen-Einschätzung unschlüssig.

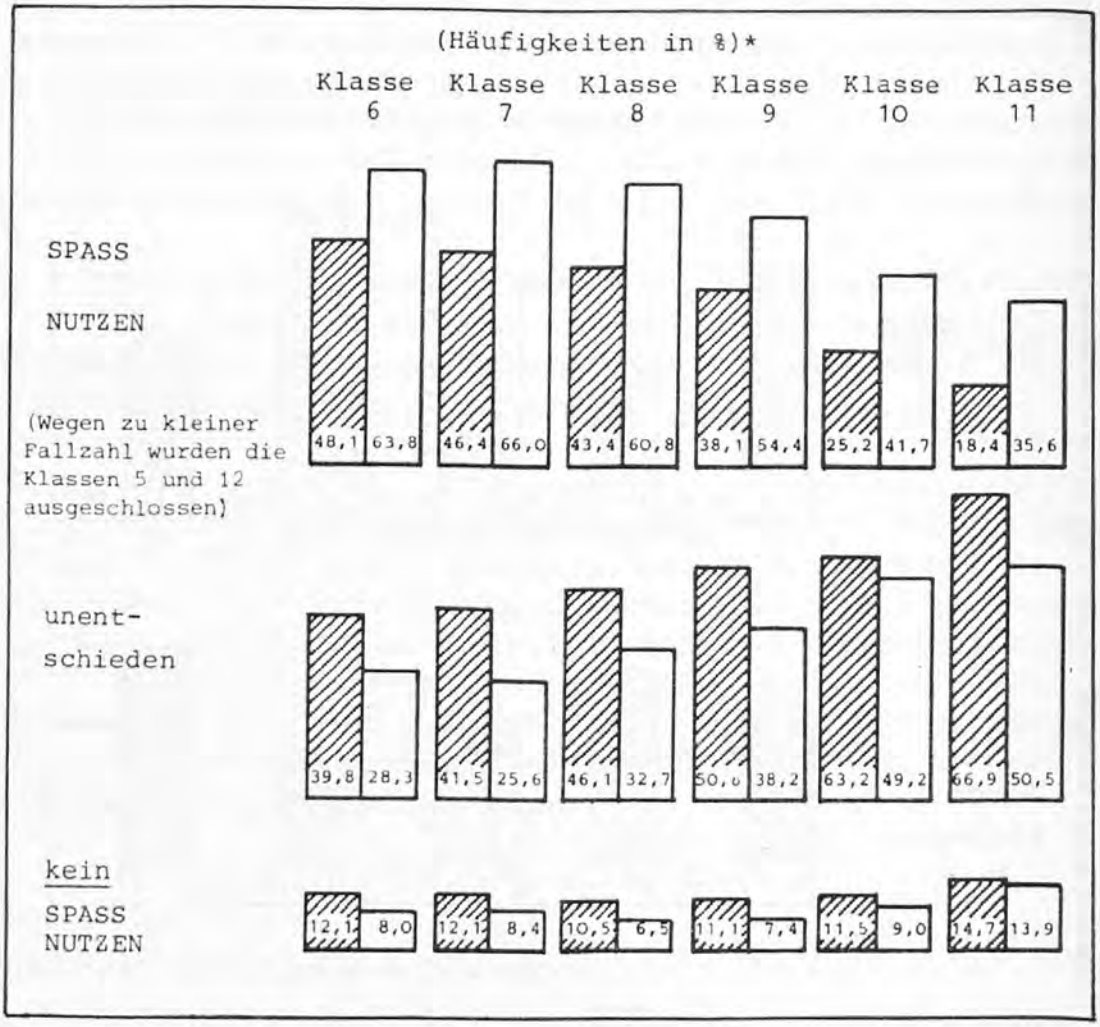


Abb. 19: Spaß- und Nutzen-Einschätzung (der RCFP-Erprobungen) je nach Klassenstufe

Dieses Phänomen ist schwer zu interpretieren. Man könnte vermuten, daß in den oberen Klassenstufen, bei denen nur Gymnasiasten und Gesamtschüler vertreten sind, „kritischere" und „vorsichtigere" Schüler geurteilt haben. Dies ist jedoch ein Trugschluß: Wie Abb. 18 zeigte, liegt bei den Gymnasiasten und Gesamtschülern der Anteil Unentschlossener nur gering über dem Anteil, den er bei Schülern der anderen Schultypen auch hat. Überraschend ist auch die Regelmäßigkeit der beiden Trends, die kontinuierlich mit der Klassenstufe fallen (bzw. wachsen).

7.4.4 Zusammenfassung

- Im allgemeinen bewerten die Schüler den Spaß an der Erprobung höher als den Nutzen der Einheit. Dies gilt für alle Einheiten.

- Es bestehen extreme Unterschiede hinsichtlich der einzelnen Projekte. Die Ergebnisse für die Einheiten RHEIN und GAST kann man als niederschmetternd bezeichnen. Nur 16 % der Schüler gaben an, sie hätten bei der Erprobung der Einheit RHEIN Spaß gehabt (bei GAST: 18 %). Bei der Einheit GAST gaben sogar 15 % der Schüler an, sie hätten keinen Spaß gehabt. Ganz im Gegensatz dazu verlief die Erprobung der Einheiten FLUG und GELT: 74 % der Schüler hatten bei der FLUG-Erprobung Spaß, und 85 % fanden sie nützlich (54 % bzw. 63 % waren es bei der RHEIN-Erprobung).
- Bei der Einheit MOBI konnten die Antworten wegen zu geringer Fallzahlen überhaupt nicht statistisch ausgewertet werden, obwohl die Einheit an 669 Schülern erprobt wurde. Die vorhandenen Antworten lassen allerdings den Schluß zu, daß die Erprobungen völlig „danebengingen", so daß die Schüler nicht mehr dazu bewegt werden konnten, nach der Erprobung auch noch einen längeren Fragebogen vollständig auszufüllen.
- Die Erprobungsbedingung der Projekte waren allerdings auch sehr unterschiedlich: FLUG, GELT und BODEN wurden bei Schülern erprobt, die eher *interessiert am Fach Erdkunde* waren, während die Einheiten RHEIN und MOBI bei unterdurchschnittlich interessierten Schülern erprobt wurden. Dies relativiert die oben dargestellten Ergebnisse, erklärt jedoch nicht die extremen Unterschiede.

7.5 Gesamtbewertung der bivariaten Analysen

- Die im obigen Kapitel demonstrierten Verfahren der Analyse *bivariater* Zusammenhänge bei nichtmetrischen Variablen erlauben einen deskriptiven Überblick über einige grundlegende Strukturzusammenhänge im Datensatz. Sie sind jedoch nicht für eine tiefergehende Zusammenhangsanalyse geeignet.
- Dies folgt aus der Tatsache, daß bei einem nichtexperimentellen Design der Untersuchung immer direkte und indirekte Variablenzusammenhänge vermischt auftreten. Man kann also niemals sicher sein, ob man wirklich den interessierenden Zusammenhang zwischen zwei untersuchten Variablen erfaßt oder den einer dritten (und vierten) Variablen, die indirekt auf die beiden betrachteten Variablen einwirkt.
- Dies wirkt sich bei einem nichtexperimentellen Design, wie dem der RCFP-Untersuchung, deshalb besonders gravierend aus, weil hier die wichtigen *Strukturvariablen* wie Geschlecht, Klassenstufe, Schultyp usw. nicht „konstant" gehalten werden können. Bei jedem betrachteten Zusammenhang (beispielsweise zwischen zwei Einstellungsvariablen) können indirekte Effekte dieser Strukturvariablen wirksam sein und dadurch den Zusammenhang verzerren oder ihn auch verdecken.

- Die obigen Mängel einer bivariaten Betrachtungsweise bedeuten jedoch nicht, daß solche Analysen völlig überflüssig sind. Jeder Datensatz muß in mehreren Schritten durchanalysiert werden, und eine bivariate Analyse stellt immer den ersten Schritt einer solchen Auswertungsprozedur dar. Sie kann Hinweise darauf geben, welche Variablen in eine spätere multiple oder multivariate Analyse einbezogen werden sollten. Bei der Analyse paralleler und gegenläufiger Einstellungsänderungen je nach Klassenstufe haben wir dargestellt, wie sich aus der bivariaten Betrachtung Hinweise auf multiple Variablenzusammenhänge ergeben.

8. Mehrvariablenanalysen: Bildung und Auswertung mehrdimensioner Kontingenztabellen

Im vorigen Kapitel haben wir uns mit der bivariaten Analyse von Variablenzusammenhängen befaßt. An mehreren Analysebeispielen konnten wir zeigen, daß die bivariate Analyse von Schülereinstellungen sehr schnell an ihre Grenzen stößt: Die unabhängigen, also erklärenden, Variablen sind keine isolierten Einflußgrößen, sondern hängen untereinander zusammen. Dadurch ergeben sich eine Vielzahl direkter und indirekter Effekte auf die abhängigen Variablen. Versucht man mit den Mitteln der bivariaten Analyse diesen Zusammenhängen nachzugehen, so verliert man sehr schnell jede Übersicht.

Einen Ausweg aus dieser Situation bieten nur Mehrvariablenanalysen. Bei metrischem Datenniveau werden seit Jahrzehnten solche Mehrvariablenanalysen routinemäßig durchgeführt, wie beispielsweise die multiple oder multivariate Regressionsanalyse. Es ist wenig bekannt, daß auch für nichtmetrische Daten analoge Verfahren der Mehrvariablenanalyse existieren – und zwar sowohl für den multiplen als auch für den multivariaten Fall. Wir werden im folgenden einige dieser Verfahren darstellen. Dies soll wieder anhand konkreter Analysebeispiele geschehen.

Zuvor jedoch wollen wir nochmals im Detail nachweisen, warum der *gleichzeitigen* Untersuchung mehrerer Variablen eine so große Bedeutung zukommt. Dabei werden wir gleich die grundlegenden Schritte einer mehrdimensionalen Kreuztabellenanalyse darstellen.

8.1 Zur Notwendigkeit von Mehrvariablenanalysen

8.1.1 Das bivariate Erklärungsschema

Bei der bivariaten statistischen Erklärung von Zusammenhängen geht es immer um folgendes Grundproblem: Man betrachtet Paare von Variablen auf dem Hintergrund einer „Theorie". Aus der Theorie ergibt sich, welche Variable als unabhängig, und welche als abhängig zu betrachten ist. Die unabhängige Variable wird dabei entweder als *kausale Ursache* der abhängigen Variablen betrachtet, oder – vorsichtiger – als eine *Bedingung* für das Auftreten bestimmter Ausprägungen der abhängigen Variablen.

Kein statistisches Verfahren der Zusammenhangsanalyse kann aus sich heraus die Richtung und die theoretische Bedeutung des jeweiligen Zusammenhangs festlegen.

Wir wollen dies an einem Beispiel demonstrieren.

Betrachten wir zwei Variablen aus dem uns vorliegenden Datensatz „SCHÜLER“:

T : „Thema der RCFP-Erprobung für später wichtig“[1]
K : „Klassenstufe“ (trichotomisiert in: Unterstufe, Mittelstufe, Oberstufe)

Zunächst müssen wir festlegen, welche Variable als abhängige, und welche als unabhängige zu betrachten ist[2], und welche Art von Zusammenhang postuliert werden soll. Da wir uns für Schülereinstellungen interessieren, setzen wir Variable T als abhängige Variable fest. Außerdem postulieren wir einen „Bedingungszusammenhang“.

Damit erhalten wir folgendes (Grund-) Schema.

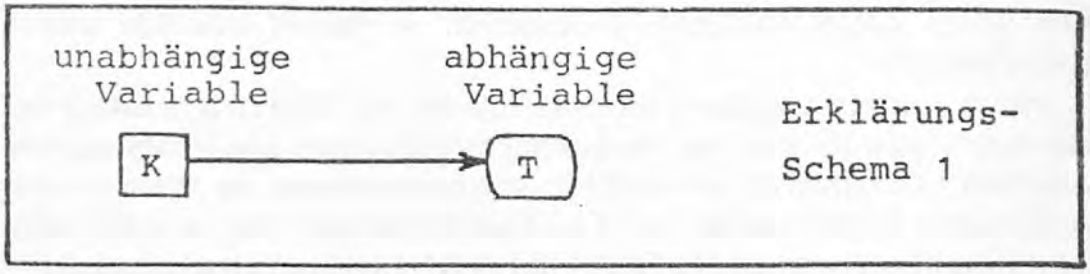


Abb. 20: Die Logik einer bivariaten Hypothese (Grundschema der Hypothesenprüfung)

Wir überprüfen die Behauptung (oder Hypothese), die in diesem Schema repräsentiert wird, durch ein angemessenes statistisches Verfahren, in unserem Fall durch eine einfache Kreuztabelle mit χ^2-Test. Es ergibt sich folgendes:

	"Thema für später wichtig?"		
	Unterstufe	Mittelstufe	Oberstufe
Thema wichtig	60,2%	67,6%	77,7%
Thema nicht wichtig	39,8%	32,4%	22,3%

χ^2=111,57; df=2; p=0,000; $Gamma_{KT}$=-0,25

Tab. 33: Schülereinschätzung der Bedeutung des Themas (der RCFP-Einheit) vor Erprobung je nach Klassenstufe

Damit haben wir bestätigt, daß mit höherer Klassenstufe die Gruppe jener Schüler signifikant zunahm, die die Themen der RCFP-Erprobungen für wichtig hielten. Um ein Maß für die „Enge" dieses Zusammenhanges zu erhalten, wurde eine der gebräuchlichen Zusammenhangsmaßzahlen („Gamma") berechnet.[3]

Bei den von uns durchgesehenen empirischen Arbeiten zur Didaktik der Geographie werden „Strukturvariablen" wie Geschlecht, Klassenstufe, Schulart usw. *generell* mit dieser bivariaten Betrachtungsweise analysiert. Auch unsere Ausführungen in Kapitel 7 beruhen auf dieser Logik.

8.1.2 Die Mehrvariablenanalyse – Ein Verfahren zur Aufdeckung von scheinzusammenhängen

Betrachten wir wieder das oben eingeführte Beispiel: Wir waren zu dem (vorläufigen) Ergebnis gekommen, daß in den höheren Klassenstufen bereits vor den Erprobungen die Themen der RCFP-Einheiten häufiger für wichtig gehalten wurden als in den unteren Klassenstufen. Je reifer die Schüler sind – so könnten wir schlußfolgern – desto größer ist die Aufgeschlossenheit für die (neuen) Themen des RCFP.

(1) Einführung einer dritten Variablen:

Da wir von dieser „Erklärung" nicht ganz überzeugt sind, stellen wir uns die Frage, ob es nicht auch einen anderen Grund für Unterschiede in der Aufgeschlossenheit gegenüber den RCFP-Themen geben könnte. Nach Durchsicht der vorhandenen Variablen entschließen wir uns, die Variable „Vortragserfahrung" in unsere Betrachtung mit einzubeziehen. Wir vermuten nämlich, daß nicht Alter oder „Reife" der Schüler schlechthin (gemessen durch die Klassenstufe) für die Aufgeschlossenheit verantwortlich sind, sondern die *schulische* Erfahrung: Schüler, die z. B. häufiger selbst kleine Vorträge halten, werden den neuen Themen des RCFP aufgeschlossener gegenüber sein als Schüler, die noch wenig Erfahrung in dieser Hinsicht haben. Da nun aber Schüler in höheren Klassenstufen vermutlich auch häufiger Vortragserfahrung haben, könnte die entscheidende Variable eben die Vortragserfahrung sein und nicht die Klassenstufe. Der bivariat festgestellte Zusammenhang mit der Klassenstufe wäre dann nur ein *Scheinzusammenhang*, hinter dem sich der Effekt der Vortragserfahrung verbirgt.

(2) Die Logik einer „echten" Dreivariablenanalyse:

Der Verdacht eines solchen Scheinzusammenhanges kann nur durch eine Mehrvariablenanalyse geklärt werden. Die einfachste Form einer Mehr-

variablenanalyse besteht darin, den Zusammenhang *zweier* Variablen unter gleichzeitiger *Berücksichtigung* einer Drittvariablen zu betrachten. Das Schema einer solchen Drei-Variablen-Analyse sieht so aus:

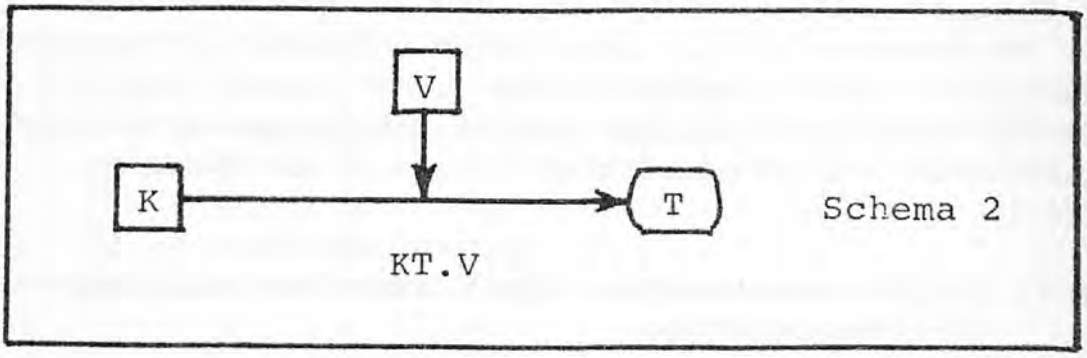


Abb. 21: Die Logik einer Dreivariablen-Analyse

Es verdeutlicht, daß wir den Effekt der Klassenstufe K auf die Aufgeschlossenheit der Schüler gegenüber den RCFP-Themen T betachten wollen, wobei der Effekt der „Vortragserfahrung" V statistisch eliminiert werden soll. Diesen reinen Effekt der Klassenstufe bezeichnet man dann als *Partialzusammenhang:* KT.V. Die statistische Methode nennt man „Auspartialisierung" oder „Konstanthaltung". Bei Kategorialdaten ist diese Auspartialisierung recht einfach: man bildet Partial- (Kreuz-) Tabellen und berechnet dafür die jeweiligen statistischen Kennwerte.

(3) Berechnung des Partialzusammenhanges: Einführung der Kontrollvariablen „Vortragserfahrung"

Bei unserem Beispiel ergibt sich dabei folgende Tabelle:

"Thema für später wichtig?"							
Schüler mit Vortragserfahrung				Schüler ohne Vortragserfahrung			
	Unterst.	Mittelst.	Oberst.		Unterst.	Mittelst.	Oberst.
Thema wichtig	68,8%	72,0%	78,1%	Thema wichtig	56,1%	61,1%	76,0%
Thema nicht wichtig	31,2%	28,0%	21,9%	Thema nicht wichtig	43,9%	38,9%	24.0%
χ^2=20,72; df=2; p=0,000; Konditionales $Gamma_{KT.V1}$=-0,153				χ^2=37,67; df=2; p=0,000; Konditionales $Gamma_{KT.V2}$=-0,198			

Tab. 34: Schülereinschätzung der Bedeutung des Themas der RCFP-Einheit (vor Erprobung) je nach Klassenstufe *und* Vortragserfahrung

Wir betrachten in Tabelle 34 den Zusammenhang zwischen der Klassenstufe und der Aufgeschlossenheit der Schüler – getrennt bei Schülern *mit* oder *ohne* Vortragserfahrung. Bei beiden Schülergruppen bleibt dieser Zusammenhang bestehen. Beide Partialtabellen[4] haben einen signifikanten χ^2-Wert. Aber bei *beiden* Schülergruppen wird der Zusammenhang *schwächer:* Gamma sinkt von −0,25 auf −0,15 bei Schülern *mit* Vortragserfahrung bzw. auf −0,20 bei Schülern *ohne* Vortragserfahrung. Dieses „Schwächer-Werden" des Zusammenhanges läßt sich noch klarer ausdrücken, wenn man aus den beiden konditionalen Gamma-Koeffizienten *einen* Wert berechnet, das sog. „partielle" Gamma[5]. Es beträgt in unserem Fall: $Gamma_{KT.V} = -0{,}17$. Das bedeutet folgendes: Betrachtet man den Zusammenhang zwischen K (der Variablen Klassenstufe) und T (der Variablen „Thema wichtig") unter Konstanthaltung von V (Variable Vortragserfahrung), so wird der Effekt von K deutlich schwächer! Die Enge des Zusammenhangs fällt von $Gamma_{KT} = -0{,}25$ auf $Gamma_{KT.V} = -0{,}17$.

Damit haben wir also in der Tat aufgedeckt, daß der Zusammenhang zwischen „Klassenstufe" und „Anregung durch RCFP-Themen" zum Teil ein *Scheinzusammenhang* ist. Man sieht dies, wenn man das Schema 1 und 2 gegenüberstellt:[6]

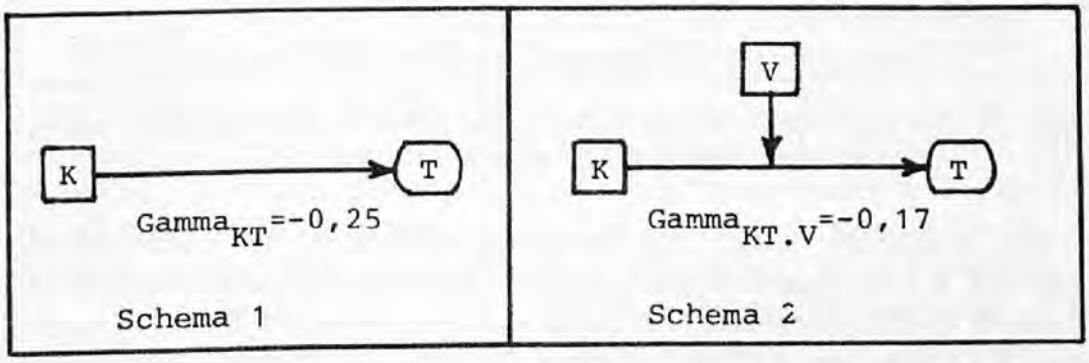


Abb. 22: Gegenüberstellung des bivariaten (Schein-)Zusammenhanges und des „auspartialisierten" Zusammenhanges aus obigem Beispiel.

Der scheinbare Effekt der Klassenstufe in Erklärungs-Schema 1 ist in Wirklichkeit also ein kombinierter Effekt aus der Variablen Klassenstufe und der Variablen Vortragserfahrung: Eliminiert man den Effekt der Variablen Vortragserfahrung in diesem kombinierten Klassenstufen-Effekt, dann sinkt der Effekt in seiner „Stärke" um ca. ein Drittel (Schema 2). Dieses Zusammenwirken von K und V bezeichnet man auch als „Interaktionseffekt".

Das Verfahren der Auspartialisierung (oder Konstanthaltung) ist also eine statistische Methode[7] zur Aufdeckung von Scheinzusammenhängen. Durch die Einführung einer zusätzlichen Variablen wird kontrolliert, ob ein Zusammenhang bestehen bleibt oder nicht. Man nennt diese konstant gehaltenen Variablen deshalb auch „Kontrollvariablen".

(4) Generalisierung des Verfahrens zur Aufdeckung von Scheinzusammenhängen:

Es ist unmittelbar einsichtig, daß diese Methode der Auspartialisierung durch Einführung *zusätzlicher* Kontrollvariablen ausgeweitet werden kann. Wir haben dies auch getan und in unserem Beispiel zusätzlich die Variable G „Gruppenarbeit"[8] eingeführt: Damit ergab sich ein Schema aus 4 Variablen. Die Erstellung der entsprechenden Partialtabellen sowie die Berechnung der χ^2-Tests, der „konditionalen" und des „partiellen" Gamma-Koeffizienten erfolgt völlig analog zum oben beschriebenen Vorgehen.
Wir erhielten folgendes Ergebnis:

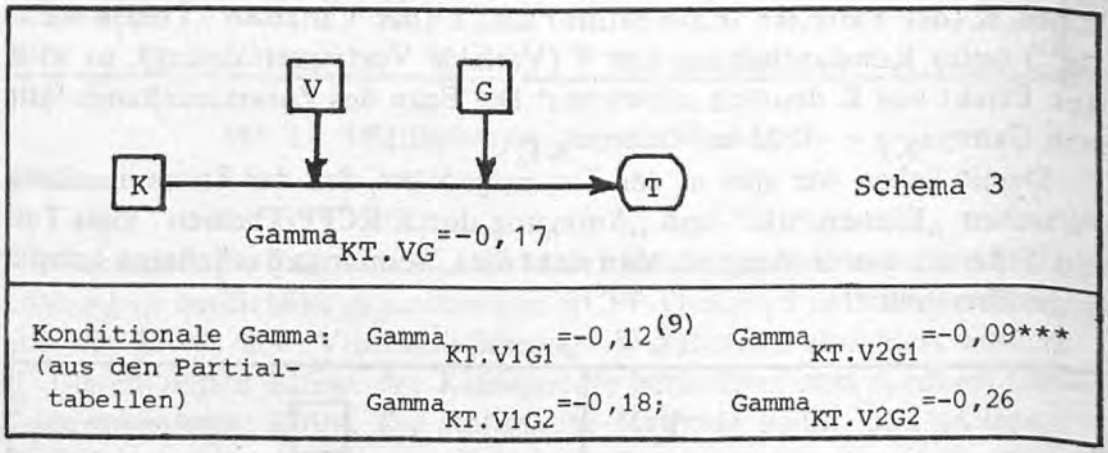


Abb. 23: Die Logik einer Viervariablen-Analyse (zwei Kontrollvariablen) mit den entsprechenden Ergebnissen für das obige Beispiel.

Die Abbildung besagt, daß bei Auspartialisierung des Einflusses von Variable V (Vortragserfahrung) *und gleichzeitiger* Auspartialisierung von Variable G (Gruppenarbeit) der Zusammenhang zwischen K (Klassenstufe) und T (Thema der RCFP-Erprobung wichtig) so erhalten *bleibt,* wie er nach der Auspartialisierung von V allein bestand. D. h. ob die Schüler „häufig" oder „selten" Gruppenarbeit gemacht haben, ist *ohne* Bedeutung für den Zusammenhang zwischen „Klassenstufe" und „Vortragserfahrung" einerseits und „Aufgeschlossenheit für das RCFP-Thema" andererseits. Unser 3-Variablen-Schema beschreibt also in Bezug auf die Variable Gruppenarbeit *keinen* Scheinzusammenhang. Die zusätzliche Einführung einer *zweiten* Kontrollvariablen verändert diesen Zusammenhang nicht mehr weiter. (s. Abb. 22/23: $\text{Gamma}_{KT.V} = \text{Gamma}_{KT.VG}$)

(5) Zusammenfassung:

- Eine generelle Methode zur Aufdeckung von Scheinzusammenhängen besteht in der (sukzessiven) Einführung von (Kontroll-) Variablen.
- Auch bei Kategorialdaten ist eine solche Mehrvariablenanalyse zur Aufdeckung von Scheinzusammenhängen möglich, wie oben demonstriert wurde.

- Je mehr (sinnvolle) Kontrollvariablen im Modell berücksichtigt werden, desto größer ist die Chance, einen Scheinzusammenhang aufzudecken.
- Bei *bivariaten* Analyseverfahren ist es *prinzipiell nicht* möglich (mit statistischen Mitteln), einen Scheinzusammenhang aufzudecken.

8.1.3 Die Mehrvariablenanalyse – Ein Verfahren zur Aufdeckung „verschütteter" Zusammenhänge

Wir hatten die Notwendigkeit einer Mehrvariablenanalyse damit begründet, daß nur dadurch „Scheinzusammenhänge" entlarvt werden können. Eine noch interessantere Eigenschaft solcher Verfahren ist jedoch, daß durch sie Zusammenhänge *aufgedeckt* werden können, die bei einer bivariaten Betrachtung verborgen bleiben. Auch dies soll wieder an einem konkreten Beipsiel aus dem uns vorliegenden RCFP-Datensatz demonstriert werden.

(1) Die „bivariate" Analyse:

Wir betrachten dazu folgende zwei Variablen:

T : „Thema" der RCFP-Erprobung wichtig? (*vor* Erprobung erhoben)
A : Anregung aus der RCFP-Erprobung (*nach* Erprobung erhoben)[10]

Wir vermuten, daß jene Schüler, die bereits vor der Erprobung das jeweilige Thema *wichtig* fanden, auch anschließend durch die RCFP-Unterrichtserprobung eher angeregt waren. Und umgekehrt: Jene Schüler, die das Thema für *nicht* wichtig hielten, werden vermutlich auch seltener die Erprobung für anregend gehalten haben. Diese Hypothese läßt sich theoretisch durch die größere Motivation der ersten Schülergruppe untermauern. Um diese bivariate Hypothese zu überprüfen, bilden wir die entsprechende Kreuztabelle und berechnen die nötigen Statistiken:

RCFP-Einheit war	Thema der RCFP-Einheit ist wichtig	Thema der RCFP-Einheit ist nicht wichtig
anregend	50,0%	51,9%
nicht anregend	50,0%	48,1%
χ^2=1,77; df=1; p=0,18	n.s.	
$Gamma_{TA}$=0,03		

Tab. 35: Anregung der Schüler durch die RCFP-Erprobung je nachdem, ob die Schüler die Thematik schon vorher anregend fanden oder nicht.

Das Ergebnis ist überraschend: Es besteht *kein* Zusammenhang (χ^2-Wert ist nicht signifikant). Wir würden das Ergebnis also so interpretieren, daß die „Motivation" der Schüler in Bezug auf das Thema der RCFP-Einheit ohne Bedeutung dafür ist, wie anregend die anschließende Erprobung der Einheit empfunden wird. Bei einer bivariaten Betrachtungsweise wären wir damit am Ende der Analyse.

(2) Einführung der Kontrollvariablen „Geschlecht":

Es ist nun allerdings ziemlich unwahrscheinlich, daß die „Erwartungshaltung" der Schüler keinerlei Einfluß darauf hätte, wie der (Erprobungs-) Unterricht empfunden wurde. Wir überlegen uns deshalb, durch welche Variable dieser, an sich bestehende, Zusammenhang verdeckt sein könnte. Nach Durchsicht mehrerer „Kandidaten" stoßen wir auf die Variable „Geschlecht". Bei Kontrolle dieser Variablen ergibt sich folgendes:

"RCFP-Einheit war anregend?"

Thema der RCFP-Einheit	männliche Schüler: wichtig	männliche Schüler: nicht wichtig	Thema der RCFP-Einheit	weibliche Schüler: wichtig	weibliche Schüler: nicht wichtig
Einheit war anregend	54,1%	51,2%	Einheit war anregend	45,7%	52,5%
nicht anregend	45,9%	48,8%	nicht anregend	54,3%	47,5%
χ^2=108,21 df = 1 p = 0,16			χ^2=11,3; df=1; p=0,00;		
$Gamma_{TA.G1}$=0,06			$Gamma_{TA.G2}$=0,14		

Tab. 36: Anregung durch die RCFP-Erprobung (nach Erprobung) je nachdem, ob die Schüler die Thematik vorher wichtig fanden oder nicht – getrennt nach Geschlecht

Bei den weiblichen Schülern besteht also sehr wohl ein (höchstsignifikanter) Zusammenhang zwischen der Erwartungshaltung gegenüber dem Thema der RCFP-Erprobung und der nachträglich beurteilten Anregung durch die Erprobung: Von den Schülerinnen, die das Thema vorher für wichtig hielten, fanden nur 46 % die darauf folgende Einheit anregend, aber 54 % *nicht* anregend! Von den Schülerinnen, die das Thema vorher für *nicht* wichtig hielten, fanden nachher ca. 53 % die Einheit anregend und nur 48 % *nicht* anregend. Bei den männlichen Schülern besteht dagegen nach wie vor kein signifikanter (!) Zusammenhang; allerdings deutet sich auch hier eine Verschiebung an: Bei männlichen Schülern, die vorher das Thema der RCFP-Einheit für wichtig hielten, fanden 54 % *auch* die Erprobung anregend und nur 46 % sie *nicht* anregend.

Anhand der Partialtabellen ist also zu erkennen, *warum* ohne Berücksichtigung des Geschlechts der an sich bestehende Zusammenhang zwischen den Variablen T und A verborgen blieb: Weibliche Schüler reagieren genau *gegenläufig* zu den männlichen Schülern. Und da wir bei der bivariaten Betrachtungsweise die Geschlechter nicht unterscheiden, *heben sich die Effekte gerade auf:* Der Zusammenhang bleibt verborgen.

Das Ergebnis dieser Mehrvariablenanalyse scheint auch theoretisch recht interessant zu sein: Man könnte es so interpretieren, daß bei weiblichen Schülern eine Art „negative Verstärkung" bei einer Erwartungshaltung wirksam wird, während dies bei männlichen Schülern nicht der Fall ist; evtl. wird bei ihnen sogar eine „positive Verstärkung" wirksam. Gerade Schülerinnen, die *hohe Erwartungen* mit dem Thema der RCFP-Erprobung verbanden, waren *besonders kritisch* gegenüber der tatsächlichen Erprobung.

(3) Zusammenfassung:

- Durch eine Mehrvariablenanalyse ist es möglich, Variablenzusammenhänge aufzudecken, die bei einer bivariaten Betrachtung verborgen bleiben.
- Dies ist z. B. dann der Fall, wenn ein nichtsignifikanter (bivariater) Variablenzusammenhang in „Wirklichkeit" aus zwei (oder mehreren) *gegenläufigen* Effekten besteht, die sich in ihrer Wirkung gerade aufheben.
- Mehrvariablenanalysen haben damit eine größere analytische Kraft als bivariate Verfahren. Wo diese vor „insignifikanten" Zusammenhängen kapitulieren müssen, können Mehrvariablenanalysen unter Umständen eine „Tiefenstruktur" aufdecken.

8.1.4 Der gemeinsame Effekt mehrerer unabhängiger Variablen

Nachdem wir in den beiden letzten Abschnitten gezeigt haben, wie durch Einführung von Kontrollvariablen und Erstellung entsprechender Partialtabellen der ursprüngliche bivariate Zusammenhang näher überprüft werden kann, bleibt uns jetzt noch folgendes Problem: Wie ergibt sich der *„gemeinsame"* Effekt zweier unabhängiger Variablen in Bezug auf die abhängige? Wir greifen dazu wieder das in 8.1.1 eingeführte Beispiel auf. Zur Erinnerung die ursprüngliche Kontingenztabelle:

abhängige Variable T : "Thema der RCFP-Einheit wichtig?"			
	Unterstufe	Mittelstufe	Oberstufe
Thema wichtig	60,2%	67,6%	77,7%
Thema nicht wichtig	39,8%	32,4%	22,3%
χ^2=108,21; df=2; p=0,00; Gamma$_{KT}$=–0,25			

Tab. 37: Schülereinschätzung der Bedeutung des Themas der RCFP-Einheit (vor Erprobung) je nach Klassenstufe

Wir wollen nun *gleichzeitig* den Effekt der Variablen *Klassenstufe, Vortragserfahrung* und *Gruppenarbeit* auf die Variable T („Thema der RCFP-Erprobung wichtig") erfassen. Dies ist mehr ein technisches als ein methodisches Problem, denn der Gesamteffekt der drei unabhängigen Variablen ergibt sich unmittelbar aus der Zusammensetzung der jeweiligen Partialtabellen zu einer neuen Gesamttabelle.

Neben gewissen technischen Schwierigkeiten bei der Erstellung der mehrdimensionalen Kontingenztabellen im Rahmen des Statistikprogramms SPSS führte der *formale* Aufbau der Tabelle öfters zu Verständnisschwierigkeiten, da er von der „zellenweisen" Konstruktion einer üblichen Kreuztabelle[11] abweicht. Betrachten wir dazu unser Beispiel:

Wir haben drei unabhängige Variablen, deren gemeinsamer Effekt dargestellt werden soll. Wir fassen dazu die drei Variablen zu einer „Supervariablen" zusammen. Diese Supervariable enthält als Ausprägungen alle Merkmalskombinationen der drei ursprünglichen Variablen.

Jede Ausprägung der Supervariablen betrachtet man als *Subpopulation*, die hinsichtlich aller einbezogenen unabhängigen Variablen – den sog. Prädiktorvariablen – homogen ist. Für jede Subpopulation werden dann die Häufigkeiten in den Merkmalen der abhängigen Variablen – man bezeichnet sie auch als Zielvariable – ausgezählt. Für unser Beispiel ergibt sich damit folgende mehrdimensionale Kontigenztabelle.

Zusammenfassung:

- Der *gemeinsame* Effekt mehrerer unabhängiger Variablen auf eine abhängige Variable (Zielvariable) ergibt sich aus den Partialtabellen für die neu hinzugefügten Variablen.

	SUPER-VARIABLE: (S)			abhängige Variable (T) (Zielvariable)	
Subpop. Nr.:	Klassen-stufe	Vortrags-erfahrung	Gruppen-arbeit	"Thema" wichtig	nicht wichtig
1	U	m	h	71,5%	28,5%
2	U	m	s	66,1%	33,9%
3	U	o	h	64,6%	354,%
4	U	o	s	50,0%	50,0%
5	M	m	h	75,8%	24,2%
6	M	m	s	67,7%	32,3%
7	M	o	h	66,0%	34,0%
8	M	o	s	56,2%	43,8%
9	O	m	h	79,2%	20,8%
10	O	m	s	77,0%	23,0%
11	O	o	h	74,6%	25,4%
12	O	o	s	77,0%	23,0%

χ^2=218,71; df=11; p=0,000; $Gamma_{ST}$=- 0,17

Erklärung: U = Unterstufe, M = Mittelstufe, O = Oberstufe
m = mit Vortragserfahrung, o = ohne Vortragserfahrung
h = häufig Gruppenarbeit, s = selten Gruppenarbeit

Tab. 38: Mehrdimensionale Kreuztabelle (Kontingenztabelle) für das Beispiel aus 8.1.2 (4)

- Multidimensionale Gesamttabellen werden mit Hilfe einer „Supervariablen" erstellt, die alle Ausprägungen der einbezogenen Variablen in Form von Subpopulationen enthält.

8.2 Die Methodik von Mehrvariablenanalysen bei nichtmetrischen Daten

Kreuztabellen – oder wie man genauer sagt Kontingenztabellen – sind das Rückgrat jeder empirischen Untersuchung. Denn obwohl sich die sozialwissenschaftlichen Disziplinen schon immer um eine exakte, metrische Quantifizierung ihrer Variablen bemüht haben, bleibt letztlich die Tatsache, daß sich viele interessante Variablen dieser metrischen Quantifizierung widersetzen. Bei den Verfahren zur Analyse von nicht-metrischen Daten kann man zwei Stufen unterscheiden.

(1) Grundlegende Verfahren

Da gibt es zunächst die Verfahren der Erstellung und Analyse multidimensionaler Kontingenztabellen (vergl. Reynolds 1977, Kap. 1 bis Kap. 4): Dazu gehören z. B. Methoden der schrittweisen Variablenselektion anhand von Chi-Quadrat-Statistiken (vergl. Chi 1979; *Higgins/Koch* 1977) sowie verschiedene – eher traditionelle – Methoden zur Analyse von Kontingenztabellen, wie die Methode der Standardisierung mit einer Kontroll- oder Testvariablen (vergl. *Rosenberg* 1962), die Methoden zur Berechnung von konditionalen und partialen Zusammenhangskoeffizienten oder die verschiedenen Varianten der Chi-Quadrat-Zerlegung (*Iversen* 1979).

Das Kennzeichen dieser ersten Stufe bei nichtmetrischen Mehrvariablenanalysen ist, daß die Basis der Verfahren die *Primärdaten* (bzw. die empirischen Häufigkeiten) sind.

(2) Höhenentwickelte Verfahren (Modelle)

Bei der zweiten Stufe der Mehrvariablenanalyse geht es dagegen nicht mehr *direkt* um die Primärdaten, sondern um die Analyse von Meßzahlen, die aus den *Zellenbesetzungen der mehrdimensionalen Kontingenztabellen* abgeleitet werden. Diese Verfahren verarbeiten also gewissermaßen die Ergebnisse der ersten Stufe weiter. Es ist allgemein üblich, bei den Verfahren der zweiten Stufe von Modellen[12] zu sprechen, da sie das Ziel haben, die komplizierten Strukturzusammenhänge einer mehrdimensionalen Kontingenztabelle auf wenige grundlegende „Parameter" zu reduzieren.
Je nach Art der abgeleiteten Maßzahlen unterscheidet man:

- „log-lineare" Modelle, bzw. „logit"-Modelle[13]
- und Modelle, die auf regressionsanalytischen Betrachtungen der als „Wahrscheinlichkeiten" definierten Zellenhäufigkeiten beruhen; wobei wieder GLS-Modelle (d. h. Modelle, die auf kleinste-Quadrate-Schätzungen beruhen = Generalized-Least-Squares) und WLS-Modelle (d. h. gewichtete-kleinste-Quadrate = Weighted-Least-Squares) unterschieden werden[14].

Zu der letzten Art von Modellen gehört der sog. GSK-Ansatz von *G*rizzle, *S*tarmer und *K*och, der in jüngster Zeit auch in der Bundesrepublik Deutschland starke Beachtung gefunden hat. Das GSK-Modell wurde übrigens auch schon in der Unterrichtsforschung (in der Bundesrepublik) diskutiert (*Bedall* 1974). Im Kapitel 9 dieser Arbeit werden wir mit einem solchen GSK-Modell eine Analyse ausgewählter Schülereinstellungen durchführen. Im folgenden Abschnitt 8.3 soll zunächst jedoch eine Vorgehensweise zur einfachen multidimensionalen Kontingenztabellenanalyse (also ein Verfahren aus Stufe 1) vorgestellt werden, in dessen Zentrum die sog. „Clark-Higgins-Koch Variable-Selection Procedure" steht (vergl. Chi 1979).

8.3 Die Vorgehensweise bei einer schrittweisen multidimensionalen Kontingenztabellenanalyse (Terminologie und Konzepte)

Die Vorgehensweise zur multidimensionalen Kontingenztabellenanalyse nach der Methode einer schrittweisen Variablenselektion besteht aus drei Schritten:

- Der Definition von Prädiktor- und Zielvariablen,
- einer varianzmaximierenden Variablenselektion und
- der Aufbereitung der Ergebnisse nach den Subpopulationen einer „Super-Variablen".

Diese drei Schritte werden wir nun im einzelnen an einem konkreten Beispiel darstellen:

8.3.1 Festlegung der Zielvariablen und der „Kandidatenliste" für die Prädiktorvariablen

Bei einer multidimensionalen Kontingenztabellenanalyse beginnen die Überlegungen mit der Auswahl einer Kandidaten-Liste für die Prädiktorvariablen und der Festlegung einer (oder mehrerer) Zielvariablen. Die Prädiktorvariablen (oder Prädiktoren, wie wir sie abkürzend nennen wollen) entsprechen den unabhängigen Variablen bei der bivariaten Kreuztabellenanalyse. Die Zielvariable ist dementsprechend die abhängige (oder zu erklärende) Variable[15]. Es können auch mehrere Zielvariablen gleichzeitig untersucht werden – was einer multivariaten Betrachtungsweise im engeren Sinn entspricht. *Alle* Variablen können dichotome oder politome Ausprägungen haben.

Wir wollen dies gleich an einem konkreten Beipsiel präzisieren: Die Variable T („Thema der RCFP-Erprobung für später wichtig") sei unsere Zielvariable.

Als nächstes müssen wir eine Kandidatenliste für die Prädiktorvariablen zusammenstellen. Dazu tragen wir *alle* Variablen zusammen, von denen wir – aufgrund theoretischer Überlegungen – vermuten, daß sie mit unserer Zielvariablen irgendwie direkt oder indirekt zusammenhängen[16]. Aus dem uns vorliegenden Datensatz mit seinen 34 Variablen[17] wurden 8 sinnvolle ausgewählt: (siehe Tab. 35)

Wie man sieht, handelt es sich bei den „Kandidaten" für die Prädiktorvariablen nicht nur um „echte" Kategorial-Daten, sondern auch um ursprünglich ordinale Variablen (wie Note) und sogar um Variablen auf ursprünglich Intervallniveau (wie Klassenstufe). Diese Variablen wurden zu polytomen Kategorialvariablen umgewandelt, bzw. zusammengefaßt. Wir verwenden also für die Analyse *nur* den *nicht*-metrischen „Informations-

Zielvariable: T "Thema der RCFP-Einheit wichtig?" (1) wichtig (2) nicht wichtig	
Kandidaten für die Prädiktorvariablen:	
Variablenname	**Ausprägungen**
Klassenstufe	(1) Unterstufe (2) Mittelstufe (3) Oberstufe
Schultyp	(1) Hauptschule (2) Real- + Berufsschule (3) Gesamtschule - Gymnasium
Geschlecht	(1) männlich (2) weiblich
Note in Erdkunde	(1) Note 1 + 2 (2) Note 3 (3) Note 4 + 5 (+6)
Rollenspieler-fahrung	(1) ja (2) nein
Vortragserfahrung	(1) ja (2) nein
Gruppenarbeit	(1) häufig (2) selten
Projekt	(1) FLUG (2) RHEIN (3) GELT (4) BODEN (5) BRAND (6) GAST (7) MOBI

Tab. 39: Liste der „Kandidatenvariablen" zur Erklärung der abhängigen Variablen T

teil" dieser Variablen. Damit wird klar, daß Variablen auf *jedem* Meßniveau in die Analyse eingehen können.

8.3.2 Varianzmaximierende Variablenselektion – Die „Clark-Higgins-Koch-Procedure"

Der zweite Schritt der Analyse besteht darin, aus der Liste der Kandidaten für die Prädiktorvariablen jene auszuwählen, die zusammen die Zielvariable optimal erklären. Das Problem dabei ist, daß die ausgewählten Variablen *zusammen* eine optimale Aufklärung erbringen sollten: Wie wir aber bereits nachgewiesen haben, addieren sich mehrere unabhängige Variablen nicht einfach in ihren Effekten auf die abhängige Variable (d. h. die Zielvariable). Vielmehr muß bei jeder *neu* eingeführten Variablen der Erklärungsanteil der bereits ausgewählten Variablen berücksichtigt, d. h. auspartialisiert werden. Wollte man durch ausprobieren aus den 8 Kandidaten-Variablen 4 als Prädiktoren auswählen, dann müßte man

$$\frac{8!}{(8-4)! \cdot 4!} = 70$$

verschiedene Prädiktor-Kombinationen[18] durchprobieren und prüfen, welche dieser „Vierer"-Kombinationen die Zielvariable optimal erklärt. Durch Probieren kommt man also kaum ans Ziel.

Einen Lösungsweg bietet die sog. „Clark-Higgins-Koch Variable Selection Procedure" (vergl. Chi 1979). Dabei handelt es sich um eine schrittweise Prozedur zur Variablenauswahl, die (gleichzeitig) auf (drei) verschiedenen χ^2-Statistiken beruht. Um die generelle Funktionsweise der Prozedur näher zu beschreiben, bringen wir zunächst einige Zitate aus dem Originalaufsatz von Higgins und Koch:

„In summary, the selection algorithm described here proceeds in the same spirit as forward stepwise regression." . . . It „relies on appropriately constructed Pearson chi-square statistics divided by their degree of freedom, which are used as a measure of the relative importance of certain combinations of variables in a multivaraite relationship. The first variable selection is one having the larges chi-square per degree of freedom with regard to first-order relationsships . . .". „Additional variables are chosen by similar selection rules using chi-square per degree of freedom computed for the appropriate higher order contingency tables for the eligible combined sets of variables" (*Higgins/Koch* 1977, S. 52).

Diese Vorgehensweise kann am besten an einem konkreten Beispiel erläutert werden. Die folgende Tabelle 40 zeigt den gesamten Selektionsvorgang für unsere Kandidatenliste auf einen Blick.

In der *ersten Runde* der Variablenselektion werden Zusammenhänge 1. Ordnung (bivariate Zusammenhänge) getestet. Die Variable „Vortragserfahrung" hat den höchsten χ^2/df-Wert, d. h. den engsten Zusammenhang mit der abhängigen Variablen T[19].

In der *zweiten Runde* bildet man mit der Variablen „Vortragserfahrung" Tabellen 2. Ordnung, also z. B. Klassenstufe/Vortragserfahrung, Schultyp/Vortragserfahrung usw.. D. h. es werden also mehrdimensionale Zusammenhänge mit T getestet, wobei die neue Gesamtvariable jedesmal die als erstes selegierte Variable „Vortragserfahrung" enthält. Als beste Kombination von Prädiktoren ergibt sich dabei die Kombination „Gruppenarbeit" mit „Vortragserfahrung". Sie hat den höchsten χ^2/df-Wert. Das bedeutet, daß der multiple Effekt dieser beiden Prädiktoren die Zielvariable T am besten (verglichen mit den anderen Prädiktorkombinationen) erklärt.

Ab der zweiten Runde der Variablenselektion muß jedoch eine weitere χ^2-Statistik beachtet werden: die sog. „termination-statistic" T_a[20]. Sie gibt an, wann die Selektion weiterer Prädiktoren abzubrechen ist, weil der jeweilige Hinzugewinn an „Erklärung" nicht mehr signifikant ist. T_a ist nichts anderes als die Summe der χ^2-Werte der *Partial*tabellen. Die Freiheitsgrade für T_a ergeben sich aus der Summe der Freiheitsgrade der Partialtabellen. Da mit den Partialtabellen – wie ausgeführt – der spezifische Effekt der kontrollierenden Variablen gemessen wird, erfaßt auch die T_a-Statistik den jeweiligen *spezifischen Beitragsgewinn* des jeweiligen Prädik-

tors. Der spezifische Effekt der „Gruppenarbeit" beispielsweise (über den Beitrag von „Vortragserfahrung" hinaus) hat einen χ^2-Wert von 40,68, der bei 2 Freiheitsgraden höchstsignifikant ist. Wir können also die Prädiktorkombination „Gruppenarbeit/Vortragserfahrung" auswählen (siehe Tab. 40, 2. Runde).

In der *dritten Runde* wird sinngemäß wie oben verfahren: Den höchsten χ^2 /df-Wert erreicht die Prädiktorkombination „Rollenspiel/Gruppenarbeit/Vortragserfahrung" (23,08). Praktisch identisch jedoch ist der Wert für die Kombination „Projekt/Gruppenarbeit/Vortragserfahrung" (23,06). Bei beiden ist T_a signifikant, so daß man sowohl die Variable „Rollenspiel" als auch die Variable „Projekt" in den Prädiktorsatz aufnehmen könnte. Da uns die letztgenannte Kombination theoretisch einleuchtender erschien, wurde sie ausgewählt.

Die Kombination „Rollenspiel/Gruppenarbeit/Vortragserfahrung" wurde darüberhinaus in einer *vierten Selektionsrunde* weiterverfolgt. Es ergab sich, daß auch hier die Variable „Projekt" den größten Zugewinn brachte (T_a war auch signifikant). Allerdings sind bei dieser Kombination die Subpopulationen zum Teil bereits sehr schwach besetzt, weswegen wir diese Lösung nicht weiterverfolgen wollen.

Die mehrdimensionalen Tabellen, die sich anhand der oben dargestellten Selektionsprozedur ergaben, „erklären" die Erwartungen der Schüler gegenüber dem Thema der RCFP-Erprobung optimal, weil der Selektionsalgorithmus für die Prädiktoren „... seeks to maximize the variation among the proportions to be studied" (*Higgins/Koch* 1977, S. 53).

Bemerkenswert erscheint uns auch, welche Variablen nach dieser Selektionsprozedur *keinen* Effekt auf die Erwartungen der Schüler haben: Es sind vor allem das „Geschlecht" und die „Note in Erdkunde". Der Effekt des „Geschlechts" ist bereits in der bivariaten Betrachtung (erste Runde) insignifikant. Aber auch die „Klassenstufe" und der „Schultyp" sind im Vergleich zu den ausgewählten Prädiktoren in ihrem Erklärungswert unbedeutend. Entscheidend für die Schülererwartungen in bezug auf ein bestimmtes Thema einer RCFP-Einheit sind offensichtlich primär *individuelle Erfahrungen der Schüler im Unterricht* und (erwartungsgemäß) die Attraktivität der jeweiligen Thematik (d. h. der Prädiktor: „Projekt").

Tab. 40: Vier Runden der Variablenselektion zur Bestimmung der „besten" Prädiktoren

Variablenselektion für das Modell zur Erklärung der Variablen: "Thema für später wichtig?"

1. Runde der Variablenselektion (bivariate Zusammenhänge)

Variable	χ^2	df	χ^2/df	T_a	df	Sign.
Klassenstufe	111,57	2	55,79			
Schultyp	13,72	2	6,86			
Geschlecht	0,26	1	0,26			n.s.
Note	21,52	2	10,76			
Rollenspiel	6,59	1	6,59			
Vortragserfahrung	108,54	1	108,54			
Gruppenarbeit	46,51	1	46,51			
Projekt	503,69	6	83,95			

2. Runde der Variablenselektion (zweidimensionale Tabellen)

Variable	χ^2	df	χ^2/df	T_a	df	Sign.
Klassenstufe Vortragserfahrung	168,47	5	33,70	58,39	4	
Schultyp Vortragserfahrung	134,91	5	26,98	25,55	4	
Geschlecht Vortragserfahrung	106,57	3	35,52	0,27	2	n.s.
Note Vortragserfahrung	120,00	5	24,00	17,02	3	
Rollenspiel Vortragserfahrung	110,02	3	36,67	2,84	2	n.s.
Gruppenarbeit Vortragserfahrung	149,45	3	49,82	40,68	2	
Projekt Vortragserfahrung	559,65	13	43,05	452,21	12	

3. Runde der Variablenselektion (zweidimensionale Tabellen)

Klassenstufe Gruppenarbeit Vortragserfahrung	218,72	11	19,88	65,86	8	
Schultyp Gruppenarbeit Vortragserfahrung	199,23	11	18,11	48,21	8	
Geschlecht Gruppenarbeit Vortragserfahrung	150,49	7	21,50	4,25	4	n.s.
Note Gruppenarbeit Vortragserfahrung	162,78	11	14,80	17,17	8	n.s.
Rollenspiel Gruppenarbeit Vortragserfahrung	161,58	7	23,08	17,03	4	
Projekt Gruppenarbeit Vortragserfahrung	622,71	27	23,06	470,56	24	

4. Runde der Variablenselektion (vierdimensionale Tabellen)

Klassenstufe Rollenspiel Gruppenarbeit Vortragserfahrung	250,282	23	10,88	83,89	16	
Schultyp Rollenspiel Gruppenarbeit Vortragserfahrung	212,426	23	9,24	48,92	16	
Geschlecht Rollenspiel Gruppenarbeit Vortragserfahrung	179,486	15	11,97	20,37	8	n.s.
Note Rollenspiel Gruppenarbeit Vortragserfahrung	184,328	23	8,01	26,70	16	n.s.
Projekt Rollenspiel Gruppenarbeit Vortragserfahrung	675,227	55	12,28	509,54	48	

χ^2-Wert nicht signifikant ($p > 0,01$): n.s.; alles andere ist signifikant oder höchstsignifikant.

8.3.3 Erstellung der Gesamttabelle und Aufbereitung der Ergebnisse

Damit ergibt sich folgende Gesamttabelle:

			T		
V	G	P	wichtig	nicht wichtig	
m	h	FLUG	72,7%	27,3%	m=mit Vortragserf.
m	h	RHEIN	85,2%	14,8%	o=ohne Vortragserf.
m	h	GELT	73,9%	26,1%	
m	h	BODEN	55,0%	45,0%	h=häufig Gruppenar.
m	h	BRAND	76,7%	23,3%	s=selten Gruppenar.
m	h	GAST	81,3%	18,7%	
m	h	MOBI	68,4%	13,6%	
m	s	FLUG	70.9%	29,1%	
m	s	RHEIN	86,6%	13,4%	
m	s	GELT	46,4%	35,6%	
m	s	BODEN	44,7%	55,3%	
m	s	BRAND	69,8%	30,2%	
m	s	GAST	73,7%	26,3%	
m	s	MOBI	81,8%	18,2%	
o	h	FLUG	71,5%	28,5%	
o	h	RHEIN	81,5%	18,5%	
o	h	GELT	61,8%	38,2%	χ^2=622,71; df=27;
o	h	BODEN	45,7%	54,3%	p=0,000;
o	h	BRAND	74,2%	25,8%	
o	h	GAST	70,7%	29,3%	
o	h	MOBI	70,5%	29,5%	
o	s	FLUG	71,1%	28,9%	
o	s	RHEIN	79,7%	20,3%	
o	s	GELT	36,9%	63,1%	
o	s	BODEN	38,4%	61,6%	
o	s	BRAND	66,1%	33,9%	
o	s	GAST	69,4%	30,6%	
o	s	MOBI	86,1%	13,9%	

Tab. 41: Mehrdimensionale Kontingenztabelle für die drei besten Prädiktoren Vortragserfahrung (V), Gruppenarbeit (G) und Projekt (P)

Nun muß man allerdings zugeben, daß diese mehrdimensionale Kontingenztabelle alles andere als übersichtlich ist. Es fällt schwer, aus den Häufigkeiten (bzw. Prozentsätzen) der Zielvariablen irgendeine konsistente Interpretation in bezug auf die Subpopulation abzuleiten. Vermutlich ist nicht zuletzt die Komplexität solcher Kontingenztabellen ein Grund für das relativ geringe Interesse an nichtmetrischen Mehrvariablenanalysen. Es gibt jedoch ein einfaches graphisches Verfahren, um diese Tabellen in einer verständlicheren Form aufzuarbeiten: Die Merkmalskombinationen werden dabei als Baumstruktur angelegt. Das Ergebnis sieht so aus: T1 ist der Schüleranteil, der das „Thema wichtig" fand.

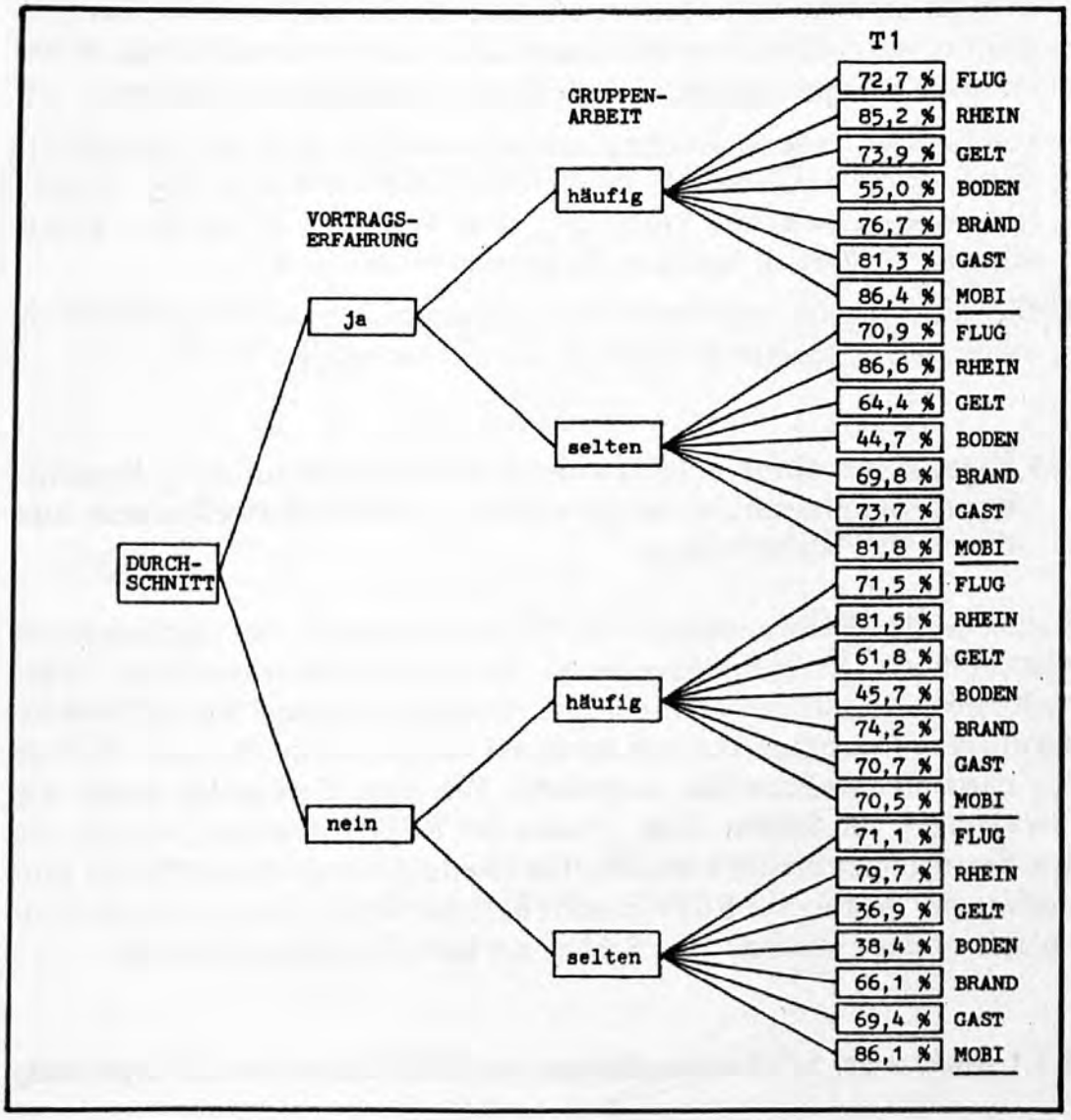


Abb. 24: Graphische Aufbereitung der mehrdimensionalen Kreuztabelle aus Tab. 41 in Form eines „Baumdiagrammes"

8.3.4 Zusammenfassung

- Nichtmetrische Mehrvariablenanalysen haben besondere Bedeutung für jede sozialwissenschaftliche Untersuchung, da ein Großteil der interessierenden Variablen nur nominales oder ordinales Meßniveau aufweisen. Dies gilt im besonderen für empirische Untersuchungen des Unterrichtsgeschehens.
- Nichtmetrische Analyseverfahren beruhen auf Kontingenztabellen. Man unterscheidet zwei Stufen: In einer ersten Stufe geht es um Verfahren zur Bildung mehrdimensionaler Kontingenztabellen bzw. um die dazu nötigen Statistiken, Maßzahlen und Auswertungsmethoden.

In einer zweiten Stufe geht es um Modelle, die an bereits vorhandene Kontingenztabellen angepaßt werden. Das Ziel der Modellbildung ist die optimale Beschreibung der Tabelle durch einige wenige Parameter.

- Für die erste Stufe der Kontingenztabellenanalyse eignet sich eine schrittweise varianzmaximierende Prädiktoren-Selektion wie die sog. „Clark-Higgins-Koch Selection Procedure". Das Verfahren ist mit der (metrischen) schrittweisen multiplen Regression vergleichbar.
- Die sich daraus ergebenden mehrdimensionalen Kontingenztabellen können als übersichtliches Baumdiagramm dargestellt werden.

8.4 Ergebnisse einer schrittweisen mehrdimensionalen Kontingenztabellenanalyse ausgewählter Schülereinstellungen aus den RDFP-Erhebungen

Nachdem wir die Notwendigkeit von Mehrvariablenanalysen begründet und eine Methodik für mehrdimensionale Kontingenztabellenanalysen vorgestellt haben, soll in den nächsten drei Abschnitten gezeigt werden, welche inhaltlichen Ergebnisse sich mit solchen Ansätzen gewinnen lassen. Wir haben dazu drei Zielvariablen ausgesucht: Die erste Zielvariable erfaßt die Erwartungen der Schüler zum „Thema der RCFP-Einheiten", wie sie *vor* den Erprobungen erhoben wurden. Die zweite Zielvariable betrifft die Einstellung der Schüler zur RCFP-Einheit *nach* der Erprobung. Die dritte Zielvariable ist das „Interesse" der Schüler am Fach Erdkunde allgemein.

8.4.1 Analyse der Schülereinstellungen zur RCFP-Einheit vor der Erprobung

Die Variable „Thema der RCFP-Einheit wichtig" erfaßt – so würden wir jedenfalls interpretieren – die *Erwartungen* der Schüler in Bezug auf eine bestimmte Thematik: Wie wir aus einer bivariaten Analyse wissen, war z. B. das Thema „Bodenzerstörung und Bodenerhaltung" wesentlich seltener für wichtig gehalten worden als die Thematik von „Tatort Rhein". Wir sind deshalb ziemlich überrascht, daß *nicht die Art des Themas* (also die Variable „Projekt") primär darüber entscheidet, ob die Schüler dieses Thema für wichtig halten oder nicht. Entscheidend ist vielmehr, ob die Schüler „Vortragserfahrung" besitzen, d. h. an selbständiges Arbeiten im Fach Erdkunde gewöhnt sind. Dies geht aus der Prozedur zur Variablenselektion hervor, die in Abschnitt 8.3.2 vorgeführt wurde. Auf diese Besonderheit sollte nochmals kurz verwiesen werden, um deutlich zu machen, daß die schrittweise Variablensektoren nicht einfach nur auf umständliche Weise das produziert, „was man sowieso schon weiß."

Zur inhaltlichen Interpretation der Ergebnisse ist es sinnvoll, nochmals die Endtabelle als Baumdiagramm darzustellen. Wir nehmen dabei eine

kleine Modifikation gegenüber Abb. 24 vor und tragen die Häufigkeiten für alle Zwischenschritte der Selektionsprozedur mit ein.
Es ergibt sich folgendes Bild:

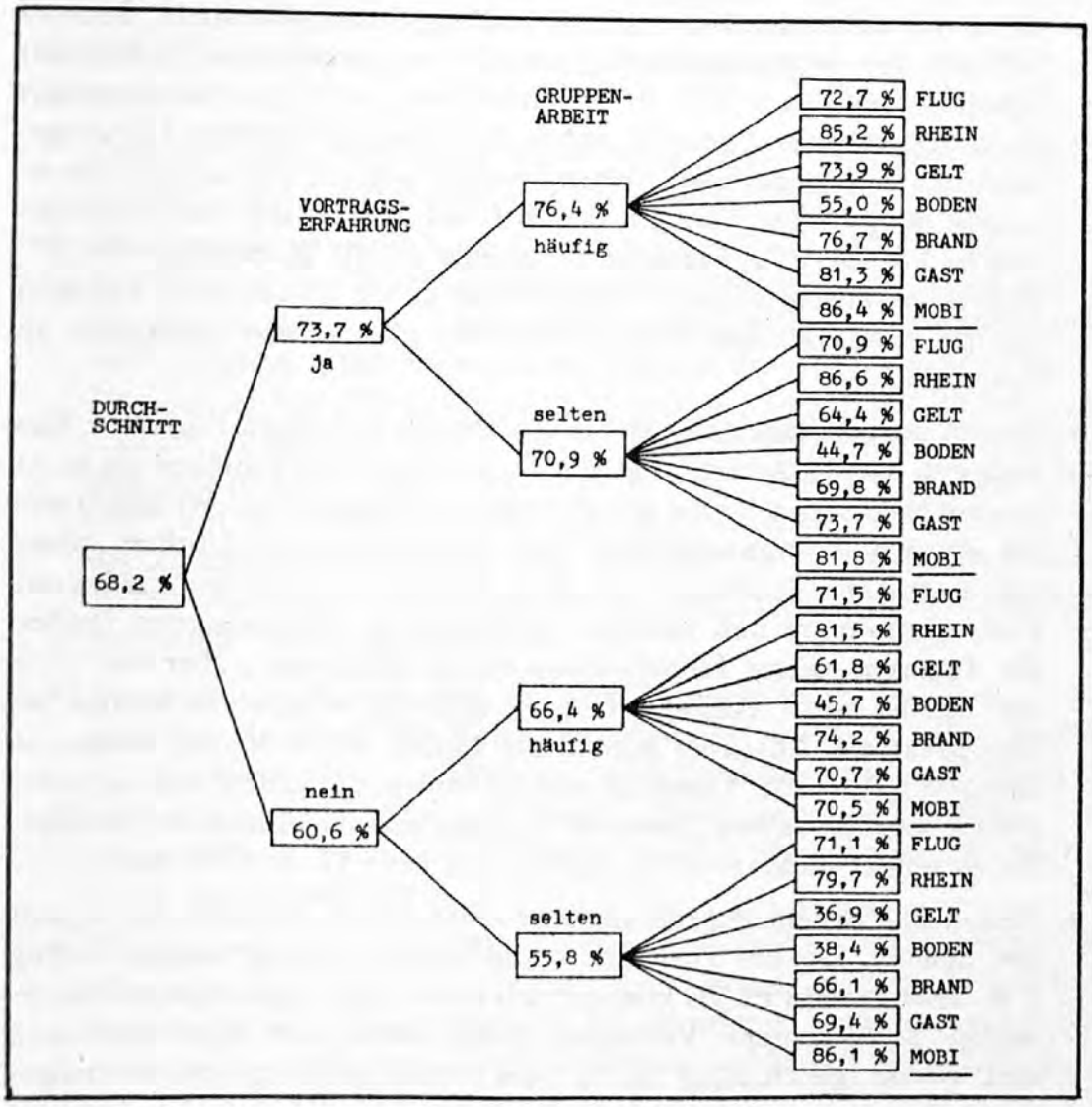


Abb. 25: Baumdiagramm der mehrdimensionalen Kontingenztabelle mit den erklärenden Variablen (Prädiktoren) V, G und P

Abbildung 25 läßt sich so interpretieren:

- 68 % *aller* Schüler fanden die Thematik der RCFP-Einheit schon vor der Erprobung wichtig. Bei den Schülern *mit* Vortragserfahrung stieg ihr Anteil um gute 5 % auf 73,7 %, bei den Schülern *ohne* Vortragserfahrung fiel ihr Anteil um knapp 8 % auf 60,6 %. D. h. je nach schulischer (Vortrags-) Erfahrung ist der Anteil der Schüler, die die Thematik der folgenden RCFP-Einheit wichtig fanden, um 13 % größer oder kleiner.

- Nach der „Vortragserfahrung" ist „Gruppenarbeit" die zweitwichtigste Einflußvariable. Generell erwarten Schüler, die häufig Gruppenarbeit gemacht haben, mehr von der Thematik der RCFP-Einheiten als solche Schüler, die selten Gruppenarbeit kennengelernt haben, und zwar *auch dann, wenn man den Effekt der Vortragserfahrung eliminiert:* Bei den Schülern mit Vortragserfahrung beträgt der zusätzliche Einfluß der Gruppenarbeit ca. ± 3%, bei Schülern ohne Vortragserfahrung sogar ca. ± 5%. Insgesamt gesehen haben diese beiden Variablen (Vortragserfahrung + Gruppenarbeit) einen beachtlichen Einfluß auf die Erwartungen der Schüler: Während ca. 76% der Schüler mit Vortragserfahrung und häufiger Teilnahme an Gruppenarbeit die Thematik der RCFP-Einheit interessant fanden, waren es nur ca. 56% (also 20% weniger) bei den Schülern ohne Vortragserfahrung und seltener Teilnahme an Gruppenarbeit.
- Erst an dritter Stelle rangiert der Einfluß des jeweiligen Projektes. Das bedeutet, daß die Attraktivität des Themas einer RCFP-Einheit gar nicht so ausschlaggebend für die Erwartungen der Schüler war. Wichtiger waren die Schüler*voraussetzungen.* Das sieht man ganz deutlich, wenn man einmal nur das Projekt GELT betrachtet: ca. 74% der Schüler mit Vortragserfahrung und häufiger Teilnahme an Gruppenarbeit fanden die Thematik dieser Einheit schon vorher interessant, aber nur 37% der Schüler ohne Vortragserfahrung und mit seltener Teilnahme an Gruppenarbeit. Ähnliches gilt für die Einheit BODEN: Hier waren es einmal 55%, die die Thematik wichtig fanden, das andere Mal ca. 38%. *Bei ein und der selben Thematik* variieren also die Anteile der Schüler, die sie schon vorher wichtig fanden, um (74–37 =) 37% bzw. 17%.
- Eine solche Regelmäßigkeit gilt aber *nicht* für alle Projekte. Der Anteil der Schüler, die die Thematik schon vorher wichtig fanden, betrug z. B. beim Projekt FLUG konstant um die 71,5% – ganz egal, ob die jeweilige Schülergruppe Vortragserfahrung hatte oder nicht und ganz egal, ob sie (gleichzeitig) häufig oder selten bei Gruppenarbeit teilgenommen hatte.
- Dieses scheinbar widersprüchliche Ergebnis bedeutet, daß der Effekt der Prädiktoren „Vortragserfahrung" und/oder „Gruppenarbeit" abhängt von der Ausprägung des Prädiktors „Projekt". D. h.: Zwischen diesen Prädiktoren bestehen *Interaktionseffekte.* Solche Interaktionseffekte werden wir später noch genauer untersuchen.

8.4.2 Analyse der Schülereinstellungen zur RCFP-Einheit nach Erprobung

Die Variable ANREG („Anregung durch die RCFP-Einheit") ist der „factor-score" aus dem 1. Faktor des Polaritätsprofils zur RCFP-Einheit (Näheres dazu wurde ausgeführt unter 6.3).

(1) Prädiktoren-Selektion:

Um die Anregung der Schüler durch die RCFP-Einheit zu erklären, verwenden wir eine ähnlich Kandidaten-Liste wie oben: Zusätzlich aufgenommen wurde lediglich die vorhin als Zielvariable fungierende Variable: „Thema der RCFP-Einheit für später wichtig?", die *vor* der Erprobung erhoben wurde. Das Ergebnis der Prädiktorenselektion (siehe Tab. 42) war folgendes:

- Drei Variablen ergeben die beste Prädiktorenkombination: Die „Klassenstufe", das „Geschlecht" und die Variable „Projekt".
- Die Erwartungshaltung der Schüler („Thema der RCFP-Einheit wichtig?") hat *allein* (!) keinen signifikanten Einfluß darauf, ob die Schüler die Erprobung der RCFP-Einheit (nachträglich) anregend fanden oder nicht. Dies bestätigt nochmals das Ergebnis aus Abschnitt 8.3, Tab. 35. Dort hatten wir allerdings zeigen können, daß ein Zusammenhang „erscheint", wenn man das Geschlecht als Kontrollvariable einführt (Tab. 36). Bei unserer Variablenselektion bestätigt sich: Die Variable „Projekt" hat in der Kombination mit dem Geschlecht einen Einfluß auf die Zielvariable „Anregung". Im *Verhältnis* zu den Variablen Klassenstufe, Geschlecht und Projekt ist der Einfluß der Erwartungshaltung der Schüler verschwindend klein.
- In die gleiche Richtung wie das obige Ergebnis deutet die Tatsache, daß auch die Note und die Rollenspielerfahrung praktisch *keinen* Einfluß auf die Anregung der Schüler durch die Erprobung haben. Kontrolliert man die Variable „Projekt", dann verlieren auch die Variablen „Vortragserfahrung" und „Gruppenarbeit" völlig ihre Bedeutung (alle T_a-Werte sind insignifikant).
- Wir können daraus den Schluß ziehen, daß individuelle Schülervoraussetzungen wie Unterricht mit Gruppenarbeit und Rollenspiel praktisch keinen Einfluß darauf haben, ob ein bestimmter (Erprobungs-) Unterricht bei der RCFP-Erprobung als anregend empfunden wurde oder nicht.

(2) Erstellung eines Baumdiagrammes für die mehrdimensionalen Kontingenztabelle zur Zielvariable ANREGUNG:

Mit den drei Prädiktoren Klassenstufe, Geschlecht und Projekt ergibt sich folgende Gesamttabelle, die wir gleich in Form der Baumdarstellung präsentieren:

Variablenselektion für das Modell zur Erklärung von "SCORE21" (ANREGUNG aus der Erprobungseinheit)

1. Runde der Variablenselektion (bivariate Zusammenhänge)

Variable	χ^2	df	χ^2/df	T_a	df	Sign.
Klassenstufe	162,131	2	81,066			
Schultyp	62,422	2	31,211			
* Geschlecht	15,132	1	15,132 *			
Note	3,257	2	1,629			n.s.
Für später wichtig	1,773	1	1,773			n.s.
Rollenspiel	1,542	1	1,542			n.s.
Vortragserfahrung	34,274	1	34,274			
Gruppenarbeit	8,804	1	8,804			
Projekt	492,137	6	82,023			

2. Runde der Variablenselektion (zweidimensionale Tabellen)

Klassenstufe Projekt	531,931	15	35,462	43,003	9	
Schultyp Projekt	559,301	19	29,436	73,889	13	
Geschlecht Projekt	513,202	13	39,477	29,598	7	
Note Projekt	503,123	20	25,156	12,852	14	n.s.
Für später wichtig Projekt	495,387	13	38,107	6,939	7	n.s.
Rollenspiel Projekt	498,480	13	38,345	10,267	7	n.s.
Vortragserfahrung Projekt	499,125	13	38,394	9,402	7	n.s.
Gruppenarbeit Projekt	505,826	13	38,910	15,450	7	n.s.

3. Runde der Variablenselektion (dreidimensionale Tabellen)

Klassenstufe Geschlecht Projekt	558,142	31	18,005	48,204	18	
Schultyp Geschlecht Projekt	583,445	39	14,960	77,213	26	
Note Geschlecht Projekt	537,862	41	13,119	28,638	28	n.s.
Für später wichtig Geschlecht Projekt	529,421	27	19,608	20,358	14	n.s.
Rollenspiel Geschlecht Projekt	525,684	27	19,370	17,094	14	n.s.
Vortragserfahrung Geschlecht Projekt	529,604	27	19,615	21,236	14	n.s.
Gruppenarbeit Geschlecht Projekt	529,241	27	19,602	19,498	14	n.s.

χ^2-Wert: nicht signifikant ($p<0,01$): n.s., alles andere ist signifikant oder höchstsignifikant

Tab. 42: Drei Runden der Variablenselektion zur Bestimmung der besten Prädiktoren für die Zielvariable „score 21"

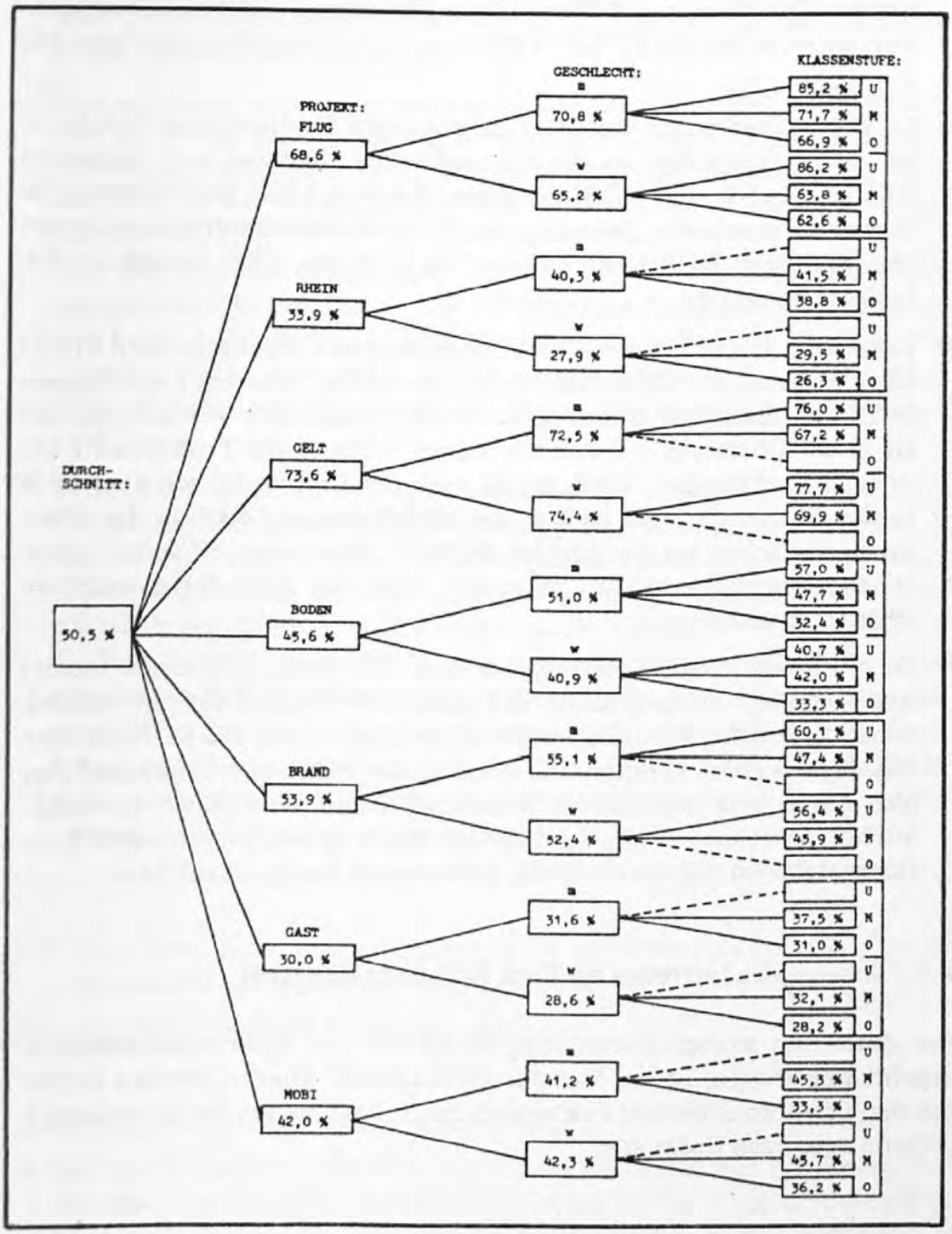


Abb. 26: Baumdiagramm der mehrdimensionalen Kontingenztabelle mit den Prädiktoren Projekt (P), Geschlecht (G) und Klassenstufe (K)

Betrachten wir die Baumdarstellung (aus Abb. 26) für die Zielvariable „ANREGUNG durch die RCFP-Einheit", so können wir folgendes feststellen:

- Der wichtigste Einfluß auf den Erfolg einer Erprobung wird erwartungsgemäß von der jeweiligen Einheit selbst ausgeübt: Während bei der

Einheit GELT sich ca. 74 % der Schüler hinterher als „angeregt" empfanden, waren es bei der Einheit GAST nur 30 % der Schüler, also fast 44 % weniger.

- Im allgemeinen waren die RCFP-Erprobungen für die männlichen Schüler etwas anregender als für die weiblichen. Nur bei den Einheiten GELT, BRAND und MOBI bestehen praktisch keine geschlechtsspezifischen Unterschiede. Besonders große Geschlechtsunterschiede traten bei der Einheit RHEIN auf (männl.: 40 %, weibl.: 28 % fanden die Erprobung anregend).
- Einen sehr deutlichen, bei allen Subpopulationen durchgehenden Effekt hat schließlich die Klassenstufe: Die „Anregung" aus den Erprobungen ist in der Unterstufe größer als in der Mittelstufe und da wieder größer als in der Oberstufe. Bei den weiblichen Schülern des Projektes FLUG sinkt der „angeregte" Schüleranteil nach der Erprobung von über 86 % in der Unterstufe (über 64 % in der Mittelstufe) auf 63 % in der Oberstufe. Ähnliches bei der Einheit BODEN: Hier waren 57 % der männlichen Unterstufenschüler „angeregt", aber nur 32 % der männlichen Oberstufenschüler.
- Zu erwähnen ist die Tatsache, daß nicht alle Subpopulationen besetzt sind. Das liegt einfach daran, daß die RCFP-Projekte für unterschiedliche Altersstufen konzipiert waren und deshalb nicht alle Klassenstufen abgedeckt werden konnten. Die vorhandenen Subpopulationen sind aber alle ausreichend besetzt. Im letzten Abschnitt werden wir ein GSK-Modell vorstellen, das in der Lage ist, solche mehrdimensionalen Kontingenztabellen mit unvollständig faktoriellem Design abzubilden.

8.4.3 Analyse des Interesses am Fach Erdkunde allgemein

Die dritte und letzten Zielvariable, für die wir eine Mehrvariablenanalyse durchführen wollen, ist die Variable INTERESSE: Hierbei handelt es sich um den (dichotomisierten) Faktorwert des 1. Faktors aus der Einstellungsbatterie zum Fach Erdkunde.

(1) Die Prädiktorenselektion:

Die Kandidaten-Liste für die Prädiktorenselektion ist wieder die gleiche wie vorhin. Das Ergebnis der Selektionsprozeduren zeigt die Tab. 43.
Der schrittweise Selektionsprozeß ergab folgendes:

- Erdkundenote, Geschlecht und Klassenstufe ergeben den erklärungskräftigsten Prädiktorensatz. Die „größte" *Gesamt*-Aufklärung der Zielvariable hatte der Prädiktorensatz Note, Geschlecht und Rollenspiel gehabt. Da der spezifische zusätzliche Beitrag der Variablen Rollenspiel

aber nicht mehr signifikant war ($T_a > 0{,}01$), wurde die zweitbeste Prädiktorkombination ausgewählt.

- Das Interesse der Schüler am Fach Erdkunde wird *nicht* beeinflußt dadurch, ob die Schüler Vortragserfahrung, Rollenspielerfahrung und Erfahrung mit Gruppenarbeit haben oder nicht. Das gilt auch dann, wenn der Einfluß der Erdkundenote, des Geschlechts und der Klassenstufe ausgeschaltet ist.
  Die individuellen Erfahrungen der Schüler mit verschiedenen Unterrichtsformen stehen also in *keinem* Zusammenhang mit ihrem Interesse am Fach. Dies dürfte ein bemerkenswertes Ergebnis für jene sein, die glauben, durch fortschrittliche Unterrichtsformen wie Gruppenarbeit oder Rollenspiele das Interesse der Schüler am Fach Erdkunde erhöhen zu können.
- Interessant ist auch, daß es für das Interesse der Schüler am Fach zunächst unwichtig ist, welchen Schultyp sie besuchen, wie wir dies ja bei der bivariaten Analyse schon festgestellt hatten (vergl. 7.3.1). Wenn jedoch der Einfluß der Note und des Geschlechts eliminiert wird, wird plötzlich der Effekt des Schultyps wieder signifikant (s. signifikanter T_a-Wert in der 3. Selektionsrunde). Die Prädiktorenkombination Note-Geschlecht-Schultyp liegt sogar in ihrer Gesamtaufklärung (mit einem χ^2/df-Wert von 20,3) nur knapp unter der von uns ausgewählten Kombination mit der Variablen „Klassenstufe" (χ^2/df-Wert = 21,4). Daran sieht man, daß die Clark-Higgins-Selektions-Prozedur in der Lage ist, „verschüttete" Variablenzusammenhänge aufzudecken.

(2) Erstellung eines Baumdiagrammes für die mehrdimensionale Kontingenztabelle zur Zielvariablen INTERESSE.

Es ergibt sich folgendes Baumdiagramm:

Zur Interpretation läßt sich folgendes sagen:

- Bei guten Schülern (Note 1 oder 2) ist der Anteil der (an Erdkunde) interessierten Schülern um fast 28% höher als bei schlechten Schülern (mit Note 4 oder 5). Aufgrund unserer bivariaten Analysen waren wir bereits zu dem Schluß gekommen, daß eher die Erdkundenote das Interesse am Fach bestimmt als umgekehrt das Interesse die Note. Dies sollte nochmals erwähnt werden, damit klar wird, warum wir die Note als Prädiktor betrachten. Wir sind uns aber bewußt, daß die Annahme einer solchen Kausalrichtung weiterer Überprüfung bedarf.
- Beim zweiten wichtigen Effekt ist die Kausalrichtung (oder wenn man will: „Bedingungsrichtung") dagegen eindeutig: Das „Geschlecht" verändert das Interesse am Fach Erdkunde relativ stark: Bei Durchschnitts-

Variablenselektion für das Modell zur Erklärung von "INTERES" (Faktor 1: Interesse am Fach Erdkunde)

1. Runde der Variablenselektion (bivariate Zusammenhänge)

Variable	χ^2	df	χ^2/df	T_a	df	Sign.
Klassenstufe	24,289	2	12,14			
Schultyp	2,316	2	1,16			n.s.
Geschlecht	72,145	1	72,15			
Note	258,913	2	129,46			
Rollenspiel	12,811	1	12,81			
Vortragserfahrung	6,459	1	6,46			n.s.
Gruppenarbeit	0,781	1	0,78			n.s.
Projekt	53,49	6	8,92			

2. Runde der Variablenselektion (zweidimensionale Tabellen)

Klassenstufe Note	290,514	8	36,31	33,42	6	
Schultyp Note	272,981	8	34,12	15,47	6	n.s.
Geschlecht Note	320,421	5	64,08	66,47	3	
Rollenspiel Note	268,588	5	53,72	8,09	3	n.s.
Vortragserfahrung Note	261,145	5	52,23	6,89	3	n.s.
Gruppenarbeit Note	259,203	5	51,84	0,23	3	n.s.
Projekt Note	316,787	20	15,84	59,86	18	

3. Runde der Variablenselektion (dreidimensionale Tabellen)

Klassenstufe Geschlecht Note	363,445	17	21,38	45,29	12	
Schultyp Geschlecht Note	344,795	17	20,28	26,39	12	
Rollenspiel Geschlecht Note	334,583	11	30,42	11,75	6	n.s.
Vortragserfahrung Geschlecht Note	326,482	11	29,42	10,39	6	n.s.
Gruppenarbeit Geschlecht Note	323,668	11	29,42	2,47	6	n.s.
Projekt Geschlecht Note	419,660	41	10,42	103,97	36	

²-Wert nicht signifikant (p 0,01): n.s., alles andere ist signifikant oder höchstsignifikant.

Tab. 43: Drei Runden der Variablenselektion zur Bestimmung der „besten" Prädiktoren für die Zielvariable INTERESSE am Fach Erdkunde.

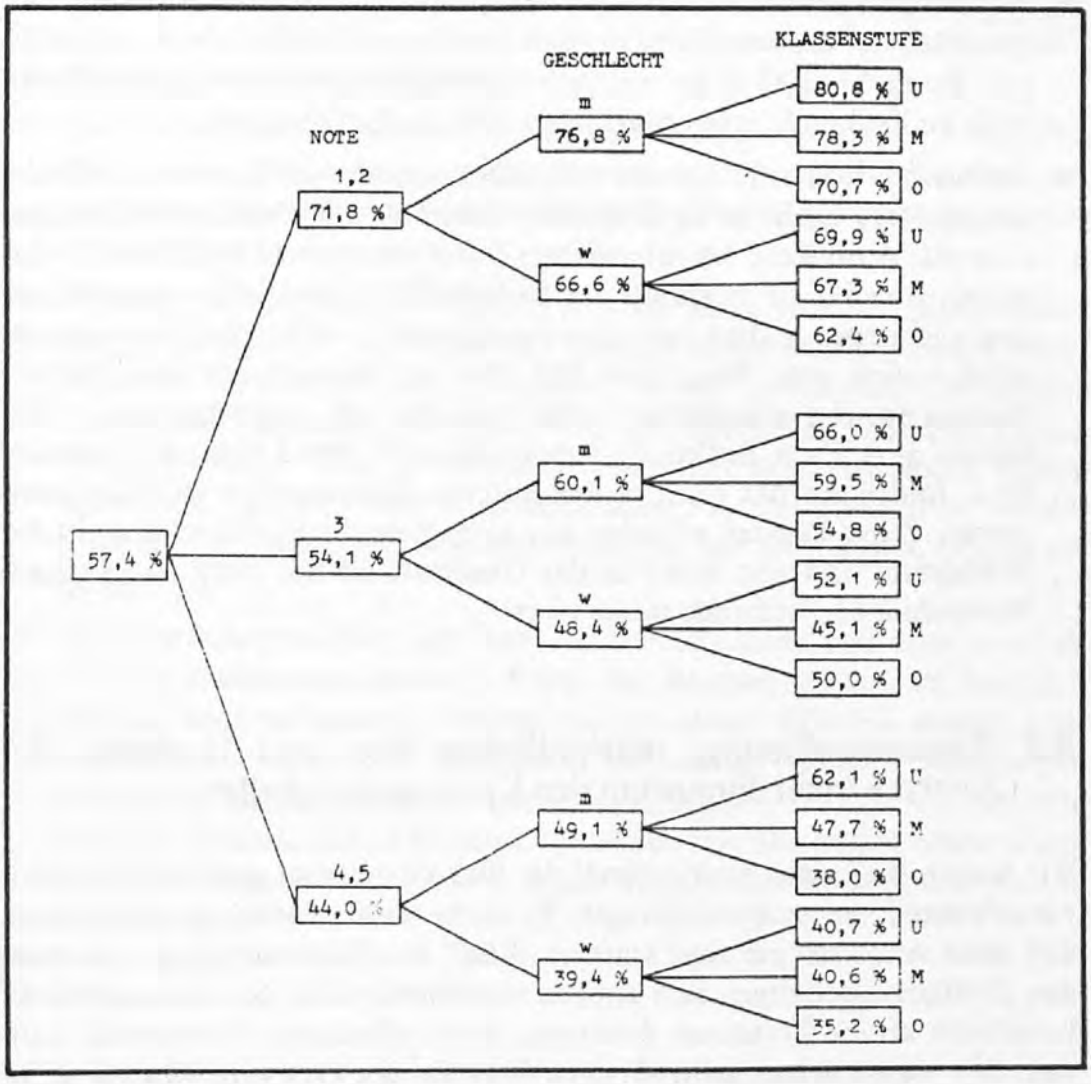


Abb. 27: Baumdiagramm der mehrdimensionalen Kontingenztabelle für die erklärenden Variablen Note (N), Geschlecht (G) und Klassenstufe (K)

schülern ist der Anteil der interessierten Mädchen um ca. 12% kleiner als der der interessierten Jungen. Auch bei guten und schlechten Schülern bewirkt der Effekt des Geschlechts einen Unterschied im Umfang der „Interessierten" von ca. 10%.

- Der Prädiktor „Klassenstufe" kompliziert die Erklärung des Schülerinteresses am Fach Erdkunde erheblich. Bei den guten und schlechten Schülern – gleich ob sie männlich oder weiblich sind – sinkt das Interesse am Fach in der Mittel- und Oberstufe allgemein stark gegenüber der Unterstufe ab. Besonders deutlich geschieht dies bei schlechten männlichen Schülern: Während sie in der Unterstufe sogar noch zu 62% interessiert sind, sackt der Anteil der Interessierten auf nur 38% in der Oberstufe ab (also um 24%).

- Betrachten wir jedoch die Durchschnittsschüler, dann dreht sich der oben beschriebene Klasseneffekt bei den weiblichen Schülern teilweise genau um: So sind nur 45 % der weiblichen Durchschnittsschüler in der Mittelstufe an Erdkunde interessiert, aber 50 % in der Oberstufe.
- Schließlich fällt auf, daß die schlechten männlichen Unterstufenschüler ein größeres Interesse an Erdkunde haben als *alle* Durchschnittsschüler (nur mit Ausnahme der männlichen Durchschnittsschüler aus der Unterstufe), obwohl sie ja wegen des Noteneffekts eigentlich *weniger* Interesse aufbringen sollten (wie dies die schlechten weiblichen Unterstufenschüler auch tun). Man kann sich dies nur dadurch erklären, daß leistungsschwache männliche Schüler zunächst ein „unrealistisches" Verhältnis zum Fach Erdkunde haben: Obwohl ihre Leistungen schwach sind, finden sie das Fach in den unteren Klassenstufen ziemlich interessant. Dann kommt offenbar um so stärker die Enttäuschung. In der Mittelstufe und erst recht in der Oberstufe ist nur noch eine (kleine) Minderheit an Erdkunde interessiert[21].

8.5 Zusammenfassung methodischer Vor- und Nachteile der Analyse mehrdimensionaler Kontingenztabellen

Wir wollen hier nicht noch einmal die Einzelergebnisse aus den drei Mehrvariablenanalysen zusammentragen. Es dürfte wohl deutlich geworden sein, daß diese Analysen ein weit klareres „Bild" der Zusammenhänge zwischen den Schülereinstellungen und einigen nichtmetrischen „Strukturvariablen" vermitteln als die bivariaten Analysen, die in Abschnitt 7 vorgestellt wurden. Die wesentlichen methodischen Vorteile der hier vorgeführten multidimensionalen Kontingenztabellen sind die folgenden:

- Das Verfahren ist außerordentlich vielseitig einsetzbar. Es können Daten auf *allen* Meßniveaus verwendet werden. Bei Daten auf ordinalem Niveau wird nur der kategoriale Meßanteil verwendet.
- Kombiniert man die mehrdimensionalen Kontingenztabellen mit einer „objektiven"[22] Selektionsprozedur für die Prädiktoren, dann erhält man immer ein Ergebnis, bei dem für die abhängige Variable (Zielvariable) eine maximale Varianzaufklärung garantiert ist (bei gegebenem Prädiktorensatz).
- Mehrdimensionale Kontingenztabellen sind nichts anderes als für bestimmte Populationen ausgezählte Häufigkeiten der Zielvariablenmerkmale, d. h. letztlich Prozentsätze. Dadurch sind die Ergebnisse für den interessierten Laien verständlich, was man bei den Parametern der metrischen Mehrvariablenmodelle kaum sagen kann. Diese Verständlichkeit gilt ganz besonders, wenn man die von uns vorgeschlagene Baumdarstellung verwendet.

Allerdings hat die Kontingenztabellenanalyse auch „systembedingte" Begrenzungen und Schwächen, auf die wir bereits hingewiesen haben. Hier noch einmal eine Zusammenfassung:

- Die mehrdimensionale Kontigenztabellenanalyse ist nur für 3 bis 6 Prädiktoren sinnvoll, weil sonst im allgemeinen zu kleine Subpopulationen entstehen. Als unterste Grenze gelten 20–30 Fälle pro Subpopulation (vgl. *Küchler* 1979, S. 169).
- Folglich ist eine solche Analyse erst für einen gut durchanalysierten Datensatz sinnvoll, bei dem sich bereits einige wenige, „strategisch" wichtige, Variablen herauskristallisiert haben. Die mehrdimensionalen Kontingenztabellenanalyse eignet sich also besonders dafür, eine bereits ziemlich „harte" Theorie zu testen. Sie eignet sich nicht dafür, Datensätze nach irgendwelchen Zusammenhängen zu durchkämmen, wie das z. B. mit der multiplen Regressionsanalyse auf metrischem Niveau oft getan wird.
- Da Kontingenztabellen mit mehr als 3 Prädiktoren und/oder mehr als 20 Subpopulationen selbst in Form der Baumdarstellung zu komplex werden, benötigt man zur Interpretation solcher Tabellen Modelle, die die wichtigen Strukturen der Tabelle auf wenige Parameter reduzieren. Die mehrdimensionale Kontingenztabellenanalyse ist also nur die erste Stufe für nichtmetrische Mehrvariablenanalysen. Die zweite Stufe bilden die verschiedenen Modelle, von denen wir jetzt das GSK-Modell vorstellen wollen.

9. Mehrvariablenanalyse: Modellbildung nach dem GSK-Ansatz

Im vorigen Abschnitt haben wir für drei ausgewählte Schülereinstellungen eine multidimensionale Kontingenztabellenanalyse einschließlich einer schrittweisen Prädiktorenselektion durchgeführt (Stufe 1 der Mehrvariablenanalyse). Im folgenden werden wir für die dort erhaltenen Kontingenztabellen GSK-Modelle entwickeln (Stufe 2 der Analyse), um ihre Interpretierbarkeit zu ermöglichen bzw. zu verbessern.

Das GSK-Modell ist zwar – verglichen mit seinen Konkurrenten, den log-linearen Modellen – von besonderer Klarheit und Einfachheit. Im Verhältnis zu dem bisher vorgestellten Analyseverfahren ist seine Logik aber doch relativ umfangreich und schwierig. Es wäre eine eigene Arbeit nötig, um den GSK-Ansatz in „extenso" darzustellen und mit ihm Analysen durchzuführen.

Wir müssen deshalb die theoretischen Ausführungen zu diesem Modell straffen. Aus diesem Grund wird es sich auch nicht vermeiden lassen, daß im folgenden eine gewisse Fachterminologie benutzt wird, die aus Platzmangel nicht im einzelnen hergeleitet werden kann.

9.1 Die Grundlagen des GSK-Modells

9.1.1 Die „Wurzeln" des GSK-Modells

Das GSK-Modell hat eine gewisse Ähnlichkeit mit dem schon länger bekannten Ansatz der üblichen Regressionsanalyse mit Dummy-Variablen. Bei der Dummy-Regression werden die unabhängigen Variablen nach ihren Merkmalen in dichotome Pseudovariablen, sog. „Dummys", zerlegt. Das bedeutet, daß alle Arten von Variablen, auch solche auf nominalem oder ordinalem Meßniveau, als unabhängige Variablen in die Regression eingehen können. Da dichotome unabhängige Variablen wie metrische behandelt werden dürfen[1], gibt es keine Schwierigkeiten mit einer solchen Analyse. Eine reine Dummy-Regression ist damit praktisch nichts anderes als eine multiple Varianzanalyse (vergl. *Fennessey* 1968).

Wie bei der Varianzanalyse ist es auch bei der Dummy-Regression allerdings unmöglich, eine *kategoriale abhängige* Variable zu benutzen. Angenommen, wir versehen die kategoriale Variable „Schultyp" mit folgender Kodierung: (1) Hauptschule, (2) Gymnasium und (3) Berufsschule und benutzten sie als *abhängige* Variable in einer Regression. Wie allgemein bekannt ist, liefert jede gewöhnliche Regression (auch die Dummy-Regression)

numerische Schätzwerte auf einer *metrischen* Skala. Wie soll man es aber interpretieren, wenn man einen Schätzwert für „Schultyp" von 2,5 bekommt? Eine Regression mit Dummy-Variablen benötigt also – wie jede andere Regression auch – eine *metrische abhängige Variable.*

Es liegt nun nahe, die abhängige Variable ebenfalls mit einer Dummy-Kodierung zu versehen, beispielsweise so: Hauptschule: (0) Ja, (1) Nein, und Gymnasium: (0) Ja, (1) Nein. Berechnet man mit einer so kodierten abhängigen Variablen eine Regression, wären Schätzwerte von zum Beispiel 0,75 ein durchaus sinnvolles Ergebnis. Angenommen, man erhielte für die abhängige Variable „Gymnasium" einen Schätzwert von 0,75, dann hieße dies nämlich, daß 75% der Befragten in ein Gymnasium gehen und 25% nicht[2]. Allerdings hätte man damit statt *einer* abhängigen Variablen (nämlich: Schultyp) *zwei abhängige Dummy-Variablen* (nämlich: Hauptschule und Gymnasium) und entsprechend auch zwei Schätzwerte.

Man könnte damit das Problem als gelöst betrachten, wenn dem nicht zwei statistische Hindernisse entgegenstünden: Erstens, das Problem fehlender Homoskedastizität einer dichotomen abhängigen Variablen und zweitens, das Problem fehlender Normalverteilung einer dichotomen Variablen. Das Problem fehlender Normalverteilung ist unmittelbar einsichtig, denn es kann etwas schlecht normalverteilt sein, wenn es lediglich aus zwei „Meßwerten" besteht[3].

Fehlende Homoskedastizität ist schwieriger zu erklären: Es handelt sich dabei um einen Aspekt der multivariaten Verallgemeinerung der Normalverteilungsforderung. Homoskedastizität bedeutet, daß die Werte der abhängigen Variablen für *alle* Werte (oder Ausprägungen) der unabhängigen Variable (n) *jeweils* normalverteilt sind. D. h. die sog. „Arrayverteilungen" müssen jeweils normalverteilt sein und gleiche Varianz haben (vergl. *Bortz* 1979 S. 224). Die Mittelwerte der Arrayverteilungen müssen außerdem auf einer Geraden liegen. Natürlich ist auch die Homoskedastizität bei dichotomen Variablen prinzipiell nicht gegeben.

Das bedeutet, daß das Verfahren der Dummy-Regression in eine „Sackgasse" führt, wenn die *abhängige* Variable eine *nicht*-metrische Variable ist. Dummy-Regressionen lösen nur das Problem für *unabhängige* nichtmetrische Variablen (vgl. *Küchler*, 1980).

9.1.2 Die Logik des GSK-Modells

Das GSK-Modell setzt genau dort an, wo die Dummy-Regression wegen der Verletzung statistischer Voraussetzungen unbrauchbar wird: nämlich bei den „Eigenschaften" der *abhängigen* Variablen. Wir hatten festgestellt, daß „Dummys" zwar als unabhängige Variablen in eine Regression eingehen können, *nicht* aber als *abhängige* Variablen, da sie hier nicht den statistischen Anforderungen genügen.

Das Kernproblem besteht also in folgendem Dilemma: Einerseits wäre es am einfachsten, wenn man die Zusammenhänge zwischen *nicht*metrischen Variablen im Rahmen der üblichen Regressionsverfahren analysieren könnte, die in den Sozialwissenschaften die weiteste Verbreitung gefunden haben[4].

Andererseits basieren alle diese Verfahren darauf, daß (mindestens) die abhängige Variable metrisches Niveau hat. Will man also ein Verfahren im Bezugsrahmen der Regressionsanalyse verwenden, so muß man zu einer *metrischen* Zielvariablen kommen.

Der GSK-Ansatz löst dieses Dilemma durch folgenden „Kunstgriff": Er wechselt die Analyse-*Ebene*. Bei jeder normalen Regression sind die Individuen bzw. ihre Werte für die abhängige Variable die „Einheiten" für die Berechnung. Beim GSK-Ansatz werden dagegen die – durch die unabhängigen Variablen definierten – Subpopulationen die Analyseeinheit. Die Werte dieser Analyseeinheiten sind dann nichts anderes als die Häufigkeiten der Zielvariablen für diese Subpopulationen. Da man diese Häufigkeiten sehr leicht in „Wahrscheinlichkeiten" umwandeln kann (was nichts anderes bedeutet als die Berechnung der Prozentsätze für die Ausprägungen der Zielvariablen, dividiert durch 100), erhält man pro Subpopulation einen metrischen Wert für die abhängige Variable. Man bezeichnet dies als *„Metrisierung"* der Zielvariablen (*Küchler* 1979 S. 158). Sehen wir uns dazu folgende Skizze an:

„normale" Regression			Regression nach dem GSK-Ansatz				
	u.V.	a.V.	Subpopu-	VAR A	VAR B	Zielvariable	
Fälle	VAR1	VAR2	lationen	Vortragserf.	Geschlecht	ja	nein
VP 1	1	3	Subp. 1	1	1	0,30	0,70
VP 2	2	4					
VP 3	1	2	Subp. 2	1	2	0,45	0,55
VP 4	2	3	Subp. 3	2	1	0,50	0,50
VP 5	2	4	Subp. 4	2	2	0,70	0,30
.							
.				„Super"-Variable			
.							
VP n	1	2					

Tab. 44: Gegenüberstellung von normaler (bivariater) Regression und der Regression zwischen einer Supervariablen und einer Zielvariablen nach dem Ansatz von *G*rizzle, *S*tarmer und *K*och

Links wurde der übliche Regressionsansatz veranschaulicht: Für „n"-Fälle wird die unabhängige Variable VAR1 und die abhängige VAR2 betrachtet und eine Regressionsgleichung berechnet. Rechts findet sich die Regressionsberechnung nach dem GSK-Ansatz: Hier ist die „unabhängige"

Variable die sog. „Supervariable", die aus den Merkmalskombinationen für die benutzten Prädiktoren (VAR A, VAR B) besteht. *Die abhängige Variable* (oder Zielvariable) *sind die Häufigkeiten für die Subpopulationen*, umgewandelt in „Wahrscheinlichkeiten". Da die Wahrscheinlichkeiten der Zielvariablen beliebige Werte zwischen 0 und 1 annehmen können, haben wir damit eine metrisierte[5] abhängige Variable. Wir rechnen unsere Regression (in obiger Skizze) also gleichsam mit *vier* „Fällen" – nämlich den vier homogenen Subpopulationen, und nicht mit den n Fällen, wie bei der normalen Regression (dies gilt nur für den Fall einer dichotomen Zielvariablen; bei polytomen Zielvariablen steigt die Anzahl der Subpopulationen, also der „Fälle").

Damit haben wir den „Kern" des GSK-Ansatzes verdeutlicht. Natürlich ist der statistische Aufwand für den GSK-Ansatz bedeutend größer, als dies hier dargestellt werden kann. Nur einen Aspekt wollen wir hier noch kurz erwähnen: Die Fallzahl (also die Anzahl der Subpopulationen) beim GSK-Modell ist im allgemeinen wesentlich kleiner als die bei üblichen Regressionen. Jedem einzelnen „Fall", d. h. jeder Subpopulation, kommt damit eine sehr große Bedeutung für das Ergebnis zu. Während bei einer normalen Regression mit 500 Fällen jeder Fall 1/500 „beiträgt", ist es bei einer GSK-Regression mit 4 Subpopulationen 1/4. Dies würde zu großen Verzerrungen führen, wenn die einzelnen Subpopulationen unterschiedliche Fallzahlen[6] repräsentieren. Deshalb wird beim GSK-Ansatz meistens die sog. „weighted-least-squares" (also „gewichtete-kleinste-Quadrate")-Schätzung vorgenommen. D. h. der Beitrag jeder Subpopulation wird danach gewichtet, wieviel (ursprüngliche) Fälle sie enthält.

9.1.3 Schritte der GSK-Modellbildung

Die Modellbildung nach dem GSK-Ansatz umfaßt im einfachsten Fall die folgenden Schritte, die wir in Form eines Überblicks zusammengestellt haben.

1. Schritt: Kodierung der Prädiktoren. Hier geht es um die Spezifizierung der sog. „Design-Matrix".

2. Schritt: Berechnung eines (orthogonalen) saturierten Modells, mit Berechnung der Signifikanz aller Effekte.

3. Schritt: Eliminierung der nicht signifikanten Effekte aus dem saturierten Modell und Berechnung eines Modells mit den verbleibenden Effekten.

4. Schritt: Vergleich zwischen den tatsächlichen Wahrscheinlichkeiten und den mit dem Modell vorhergesagten Wahrscheinlichkeiten. Dies bezeichnet man auch als Berechnung der „Residuen". Überprüfung, ob die Residuen noch bedeutend, d. h. signifikant, nach einer χ^2-Statistik sind.

5. Schritt: Verbesserung des Modells bei signifikanten Residuen, so lange bis die Residuen insignifikant werden.

6. Schritt: Vereinfachung der Modelleffekte so lange, bis *alle Effekte* sowohl statistisch signifikant als auch theoretisch gehaltvoll sind. Dies geschieht zum Beispiel durch Umwandlung von Interaktionseffekten in sog. „konditionale" Effekte.

7. Schritt: Aufbereitung der Regressionskoeffizienten für die Interpretation.

Die Schritte 3 bis 5 bezeichnet man auch als „model-fitting". Diese sieben Schritte gelten allerdings nur für den einfachsten Fall eines orthogonalen Designs mit *einer* Zielvariablen. Bei einer multivariaten Analyse – also einer mit zwei oder mehr Zielvariablen – kompliziert sich die Prozedur natürlich erheblich. Ebenso bei nicht orthogonalem Design, d. h. bei Nichtbesetzung einzelner Subpopulationen. Es ist hier unmöglich, diese Schritte im einzelnen darzustellen und zu begründen; (näheres findet sich bei *Küchler* 1976, 1979, S. 154 ff/ *Kritzer* 1978/ *Bedall* 1974/ *Lehnen, Koch* 1974). Im folgenden werden wir diese Analyseschritte exemplarisch vorführen, wenn wir das Modell für die erste von uns ausgewählte Zielvariable entwickeln.

9.2 Ein GSK-Modell für die Erwartungen der Schüler in Bezug auf das „Thema" der RCFP-Einheit" vor Erprobung

Wir knüpfen hier an die erste der drei Kontingenztabellenanalysen in Abschnitt 8. an. Dort hatten wir nach unserer Prozedur zur Prädiktorenselektion zwei in etwa gleichwertige Lösungen erhalten, die die unabhängige Variable „Thema der RCFP-Einheit wichtig?" mit den uns zur Verfügung stehenden Variablen relativ optimal erklären. Die erste Lösung mit *drei* Prädiktoren war noch leicht zu interpretieren, obwohl bereits dort Interaktionen auftraten, die man in ihrer Bedeutung nicht klar abschätzen konnte. Bei der zweiten Lösung dagegen ergaben sich – durch die vier Prädiktoren – so viele Subpopulationen, daß die Zellenbesetzung der Tabelle nicht mehr ausreichend war. Wir wollen nun für die erste Lösung ein GSK-Modell entwickeln, um zu demonstrieren, daß sich dadurch die Interpretierbarkeit wesentlich verbessert.

Die Basis der Modellbildung ist – wie schon erwähnt – eine multidimensionale Kontingenztabelle. Es handelt sich um die Tabelle mit den Prädiktoren V („Vortragserfahrung"), G („Gruppenarbeit") und P („Projekt"). Die abhängige Variable oder Zielvariable war T („Thema der RCFP-Einheit für später wichtig?").

Wir präsentieren diese Tabelle hier nochmals in der üblichen Form (also nicht als Baumdiagramm), denn wir benötigen sie für den 1. Schritt der Modellbildung, nämlich der Erstellung der Design-Matrix (siehe Tab. 45).

			T		
V	G	P	wichtig	nicht wichtig	
m	h	FLUG	72,7%	27,3%	m=mit Vortragserf.
m	h	RHEIN	85,2%	14,8%	o=ohne Vortragserf.
m	h	GELT	73,9%	26,1%	h=häufig Gruppenar.
m	h	BODEN	55,0%	45,0%	s=selten Gruppenar.
m	h	BRAND	76,7%	23,3%	
m	h	GAST	81,3%	18,7%	
m	h	MOBI	68,4%	13,6%	
m	s	FLUG	70.9%	29,1%	
m	s	RHEIN	86,6%	13,4%	
m	s	GELT	46,4%	35,6%	
m	s	BODEN	44,7%	55,3%	
m	s	BRAND	69,8%	30,2%	
m	s	GAST	73,7%	26,3%	
m	s	MOBI	81,8%	18,2%	
o	h	FLUG	71,5%	28,5%	
o	h	RHEIN	81,5%	18,5%	
o	h	GELT	61,8%	38,2%	χ^2=622,71; df=27;
o	h	BODEN	45,7%	54,3%	p=0,000;
o	h	BRAND	74,2%	25,8%	
o	h	GAST	70,7%	29,3%	
o	h	MOBI	70,5%	29,5%	
o	s	FLUG	71,1%	28,9%	
o	s	RHEIN	79,7%	20,3%	
o	s	GELT	36,9%	63,1%	
o	s	BODEN	38,4%	61,6%	
o	s	BRAND	66,1%	33,9%	
o	s	GAST	69,4%	30,6%	
o	s	MOBI	86,1%	13,9%	

Tab. 45: Mehrdimensionale Kontingenztabelle für die Prädiktoren V, G und P

(1) Erstellung der Design-Matrix

Der erste Schritt zur Modellbildung besteht darin, die Prädiktoren V, G und P in einer Form zu kodieren, mit der sie in die Regressionsrechnung des GSK-Modells eingehen können. Im Prinzip handelt es sich dabei um eine Dummy-Kodierung aller Effekte dieser drei Prädiktoren. Allerdings verwenden wir nicht die bei Dummys übliche 0/1-Kodierung (die man als „cornered" bezeichnet), sondern die −1/1-Kodierung (die man auch „centered" nennt). Der Grund dafür ist die leichtere Interpretierbarkeit der sog. „implizierten" Effekte[7]. Die Kodierung der beiden dichotomen Prädiktoren V und G ist dabei problemlos. Schwierig wird es bei dem siebenstufigen Prädiktor P und bei den Interaktionseffekten aus den drei Prädiktoren. Sehen wir uns zuerst die vollständig kodierte Design-Matrix an: (siehe Tab. 46)

	MEAN	V	G	P1	P2	P3	P4	P5	P6	VG
1	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00
2	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
3	1.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
4	1.00	1.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00
5	1.00	1.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00
6	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00
7	1.00	1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1.00
8	1.00	1.00	-1.00	1.00	0.00	0.00	0.00	0.00	0.00	-1.00
9	1.00	1.00	-1.00	0.00	1.00	0.00	0.00	0.00	0.00	-1.00
10	1.00	1.00	-1.00	0.00	0.00	1.00	0.00	0.00	0.00	-1.00
11	1.00	1.00	-1.00	0.00	0.00	0.00	1.00	0.00	0.00	-1.00
12	1.00	1.00	-1.00	0.00	0.00	0.00	0.00	1.00	0.00	-1.00
13	1.00	1.00	-1.00	0.00	0.00	0.00	0.00	0.00	1.00	-1.00
14	1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
15	1.00	-1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	-1.00
16	1.00	-1.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	-1.00
17	1.00	-1.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	-1.00
18	1.00	-1.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	-1.00
19	1.00	-1.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	-1.00
20	1.00	-1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	-1.00
21	1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
22	1.00	-1.00	-1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00
23	1.00	-1.00	-1.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
24	1.00	-1.00	-1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
25	1.00	-1.00	-1.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00
26	1.00	-1.00	-1.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00
27	1.00	-1.00	-1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00
28	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1.00

	VP1	VP2	VP3	VP4	VP5	VP6	GP1	GP2	GP3	GP4
1	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
4	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00
5	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
7	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
8	1.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	0.00
9	0.00	1.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.00
10	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00
11	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	-1.00
12	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
14	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1.00	1.00	1.00	1.00
15	-1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
16	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
17	0.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
18	0.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	1.00
19	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00
20	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00
21	1.00	1.00	1.00	1.00	1.00	1.00	-1.00	-1.00	-1.00	-1.00
22	-1.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	0.00
23	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.00
24	0.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00
25	0.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	-1.00
26	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00
27	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00
28	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

	GP5	GP6	VGP1	VGP2	VGP3	VGP4	VGP5	VGP6
1	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
5	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
6	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00
7	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
8	0.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.00
12	-1.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00
13	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	-1.00
14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15	0.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	0.00
18	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.00
19	1.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00
20	0.00	1.00	0.00	0.00	0.00	0.00	0.00	-1.00
21	-1.00	-1.00	1.00	1.00	1.00	1.00	1.00	1.00
22	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
23	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
24	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
25	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
26	-1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
27	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	1.00
28	1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00

Tab 46: Design-Matrix für die mehrdimensionale Kontingenztabelle aus Tab. 45

Auf den ersten Blick ist diese Design-Matrix sicher verwirrend. Es wurden deshalb einige „Hilfsmittel" eingezeichnet, um sie besser erklären zu können. Wir hatten 28 verschiedene Subpopulationen, d. h. Zeilen der Design-Matrix. Der Durchschnittseffekt (also der Mittelwert aller Wahrscheinlichkeiten) wird durchgehend mit 1 kodiert. Die Vortragserfahrung V wird −1/1 kodiert, ebenso die Gruppenarbeit G. Da wir 7 Projekte haben, müssen 6 unterschiedliche Dummy-Variablen für P kodiert werden. (Die siebte Dummy-Variable ergibt sich implizit, ebenso wie die *zweite* Dummy-Variable bei V und G.) Dies geschieht dadurch, daß eine − 1/1-Kodierung so für die 6 Dummys angelegt wird, daß die 1-Kodierung diagonal durch alle 6 Subpopulationen „wandert". Soviel zur Kodierung der sog. „Haupteffekte": V, G und P1 bis P6.

Wie wir aus unserer Interpretation der Kontingenztabelle aber bereits wissen, bestehen nicht nur isolierte, additive Effekte der Prädiktoren auf die Zielvariable, sondern auch *interaktive* Effekte zwischen den Prädiktoren. Bei unseren 3 Prädiktoren gibt es 13 Interaktionseffekte 1. Ordnung und 6 Interaktionen 2. Ordnung, also 19 mögliche Interaktionseffekte. Sie alle gehen zunächst in das Modell ein und müssen deshalb kodiert werden. Die Kodierung sieht zwar kompliziert aus, ist aber im Grunde sehr einfach. Nehmen wir als Beispiel den Effekt der Interaktion zwischen der Gruppenarbeit und dem Projekt 2, also die Spalte GP2 in Tab. 46. Die Kodierung dieser Spalte ergibt sich, indem man alle Kodierungen von G mit denen von P2 multipliziert. Also für die 1. Zeile: (1) . (0) = 0; oder für die 2. Zeile: (1) · (1) = 1; usw.

Im einfachsten Fall wird die Design-Matrix bei Benützung des Programms NONMET automatisch erstellt. Bei komplizierteren Designs (z. B. bei nicht orthogonalen Ansätzen oder mehreren Zielvariablen) muß diese Design-Matrix explizit spezifiziert werden.

(2) Berechnung des sog. „saturierten" Modells und der Signifikanzen für die Einzeleffekte

Im nächsten Schritt wird mit der Design-Matrix und dem Vektor der Wahrscheinlichkeiten der Zielvariablen die „kleinste Quadrateschätzung" der Modellparameter durchgeführt (die Einzelheiten der Berechnung finden sich bei *Kritzer* 1978).

Für den Beitrag jedes Parameters zum Modell wird außerdem ein χ^2-Wert berechnet und auf Signifikanz geprüft. Da wir hier beim saturierten Modell *alle* Effekte berücksichtigen, kann die ursprüngliche Wahrscheinlichkeitsverteilung der Zielvariablen *vollständig* reproduziert werden. Hier das Ergebnis des saturierten Modells für unser Beispiel:[8]

Die Tabelle zeigt für jeden der 28 Parameter (und für den Mittelwert, der dem „konstanten Glied" in der Regressionsgleichung entspricht) die b-Koeffizienten, deren Varianzen sowie einen χ^2-Wert für jeden b-Koeffi-

TESTS OF INDIVIDUAL PARAMETERS:

PARAMETER	B	VARIANCE	CHI SQUARE	P
1 MEAN	.6952E+00	.4684E-04	10318.11	.00000
2 V	.3553E-01	.4684E-04	26.95	.00000
3 G	.2415E-01	.4684E-04	12.46	.00042
4 P1	.2025E-01	.2654E-03	1.54	.21394*
5 P2	.1372E+00	.1881E-03	100.13	.00000
6 P3	-.1027E+00	.2700E-03	39.11	.00000
7 P4	-.2358E+00	.2226E-03	249.76	.00000
8 P5	.2177E-01	.2027E-03	2.34	.12624*
9 P6	.4256E-01	.5374E-03	3.37	.06636
10 VG	.3889E-02	.4684E-04	.32	.56987*
11 VP1	-.3317E-01	.2654E-03	4.15	.04174
12 VP2	-.8991E-02	.1881E-03	.43	.51214*
13 VP3	.6381E-01	.2700E-03	15.08	.00010
14 VP4	.3379E-02	.2226E-03	.05	.82081*
15 VP5	-.2010E-01	.2027E-03	1.99	.15798*
16 VP6	.1638E-02	.5374E-03	.00	.94366*
17 GP1	-.1845E-01	.2654E-03	1.28	.25758*
18 GP2	-.2335E-01	.1881E-03	2.90	.08869*
19 GP3	.6178E-01	.2700E-03	14.14	.00017
20 GP4	.1980E-01	.2226E-03	1.76	.18458*
21 GP5	.1362E-01	.2027E-03	.92	.33865*
22 GP6	-.1886E-02	.5374E-03	.01	.93516*
23 VGP1	-.5313E-03	.2654E-03	.00	.97398*
24 VGP2	-.1169E-01	.1881E-03	.73	.39394*
25 VGP3	-.4248E-01	.2700E-03	6.68	.00972
26 VGP4	.3425E-02	.2226E-03	.05	.81844*
27 VGP5	-.6942E-02	.2027E-03	.24	.62585*
28 VGP6	.1166E-01	.5374E-03	.25	.61499*

CHI-SQUARE = 26.9484 DF = 1 P = .0000 TEST FÜR V
CHI-SQUARE = 12.4561 DF = 1 P = .0004 TEST FÜR G
CHI-SQUARE = 402.6628 DF = 6 P = .0000 TEST FÜR P

* bedeutet: nicht signifikant.

Tab. 47: Saturiertes Modell für die mehrdimensionale Kontingenztabelle aus Tab. 45 (mit den erklärenden Variablen: Vortragserfahrung (V), Gruppenarbeit (G) und Projekt (P), sowie der Zielvariablen: „Thema wichtig")

zienten und dessen Signifikanz. Mit Hilfe der b-Koeffizienten und der Design-Matrix kann man die Wahrscheinlichkeiten der Zielvariablen für alle 28 Subpopulationen berechnen. Sie entsprechen (hier beim saturierten Fall) *genau den ursprünglichen Häufigkeiten* (geteilt durch Hundert) der Zielvariablen in den Subpopulationen.

Die Berechnung des saturierten Modells ist also „lediglich eine verlustfreie Umformung der Daten" (*Küchler* 1979, S. 168). Es gibt *keine* Residuen. Der Sinn dieser Umformung liegt darin, daß jetzt die *relativen* Beiträge der einzelnen Prädiktor-Effekte (also der „b-Gewichte") sichtbar werden: Aus den 28 Subpopulationen erhält man 28 Effekte mit unterschiedlichen b-Koeffizienten. Bei der Berechnung weiterer Modelle werden die unwichtigen Effekte, also die mit den *nicht-signifikanten* b-Koeffizienten, *ausgeschlossen*. Das saturierte Modell ist also lediglich eine Zwischenstufe für die Modellbildung. Wie wir aus Tab. 47 entnehmen, sind von den 28 prinzipiell möglichen Effekten nur wenige „wichtig". Nimmt man ein Signifikanzniveau von $p < 0{,}05$ an, so sind für die weitere Betrachtung fol-

gende Effekte auszuwählen: Die direkten Effekte: V, G, P2, P3, P4 und die Interaktionseffekte; VP1, VP3, GP3, VGP3.

(3) „Model-Fitting": statistisch-technische Bewertung des Modells

Mit diesen oben genannten Effekten werden nun weitere Modelle berechnet. Damit beginnt die sog. Modellanpassung.

Ihr Ziel ist es, mit so wenig wie möglich Effekten die Wahrscheinlichkeitsverteilung der Zielvariablen zu reproduzieren, und zwar so, daß die Abweichungen (Residuen) nicht mehr signifikant sind. Dazu wird für jedes der folgenden Modelle eine χ^2-Statistik für die *Residuen* berechnet. Sind die χ^2-Werte signifikant, heißt das, daß das Modell (noch) nicht optimal ist, weil es die Wahrscheinlichkeiten der Zielvariablen zu ungenau vorhersagt[9]. Sehen wir uns das Modell an, das sich aus den oben ausgesuchten (signifikanten) Effekten ergibt:

Modell: Haupteffekte: V,G,P2,P3,P4
Interaktionseffekte: VP3, GP3, VGP3

Parameter		b - Koeff.	χ^2	P
1	MEAN	o,70	12860,3	0,000
2	V	0,04	34,1	0,000
3	G	·0,03	21,1	0,000
4	P2	0,15	156,2	0,000
5	P3	- 0,09	34,4	0,000
6	P4	- 0,22	259,8	0,000
7	VP3	0,05	13,7	0,000
8	GP3	0,05	21,6	0,000
9	VGP3	- 0,05	14,8	0,000

χ^2 - der Residuen = 37,63 , DF = 19 , P = 0,007

Tab. 48: Modell der (nach dem saturierten Modell) signifikanten Effekte (1. Schritt des „model-fitting")

Wir haben nur noch 9 der ursprünglichen 28 Effekte. Jeder einzelne liefert jetzt einen signifikanten Beitrag, d. h. einen signifikanten b-Koeffizienten. Aber das *Gesamtmodell* ist dafür jetzt zu *ungenau*: Die Residuen sind bei einem χ^2-Wert von 37,6 und bei 19 Freiheitsgraden (28–9=19) höchstsignifikant.

Für diese sukzessive Modellanpassung gibt es keinen eindeutigen Lösungsweg. Die Methodenspezialisten betonen ausdrücklich, daß für das „model-fitting" ein gewisses Fingerspitzengefühl notwendig ist. Es müssen dabei nämlich mehrere – teilweise unvereinbare – Gesichtspunkte gleichzeitig beachtet werden:

- Das Modell soll nach statistischen Kriterien (also dem Effekte-Test und dem Residuen-Test) optimal sein.
- Es soll inhaltlich plausibel sein.
- Und es soll möglichst sparsam sein, d. h. möglichst wenige Parameter enthalten.

Es gibt deshalb oft mehrere gleich gute Modelle, die entweder besonders genau oder inhaltlich sehr plausibel oder besonders sparsam sind (vgl. *Küchler* 1979, S. 193).

Wir haben eine Reihe weiterer Modelle durchgerechnet. Sehr hilfreich ist es dabei, wenn man jedesmal die Residuen genau inspiziert. An ihnen ist zu ersehen, welche Subpopulationen am schlechtesten abgebildet werden. Daraus kann man Rückschlüsse für die Spezifizierung der Modelleffekte beim nächsten Modell ziehen.

Bei unserer Modellanpassung fanden wir folgendes „bestes" Modell.

Modell: Haupteffekte: V, G, P2, P3, P4, P6
konditionale Effekte: P7 V2, P3 V2 G2

	Modellparameter	b-Koeffizient	χ-Wert	P
1	MEAN	0,71	11685,1	0,000
2	V (Vortragserf.)	0,03	13,6	0,000
3	G (Gruppenarb.)	0,02	16,0	0,000
4	P2 (RHEIN)	0,13	119,0	0,000
5	P3 (GELT)	– 0,06	8,6	0,003
6	P4 (BODEN)	– 0,25	284,2	0,000
7	P4 (GAST)	0,05	9,8	0,002
8	P7 < V2	0,13	9,7	0,002
9	P3 < V2 < G2	– 0,23	56,7	0,000

χ^2 der Residuen = 16,259 DF = 10, P = 0,64
(P7* implizit: b-Koeffizient = 0,12)

Tab. 49: Bestes Modell zur Erklärung der mehrdimensionalen Kontingenztabelle von Tab. 45 (letzter Schritt des „model-fitting")

(4) Inhaltliche Auffüllung des Modells:

Zweifellos ist das Modell in der oben präsentierten (technischen) Form ziemlich unverständlich. Wir wollen es deshalb mit allgemeinverständlichem Inhalt füllen:

Dazu muß man zunächst wissen, daß bei dem hier vorgestellten GSK-Modell die b-Koeffizienten eine sehr anschauliche Bedeutung haben: Es sind nämlich einfache *Prozentsätze*.

Nun wird das Modell schon klarer: Der erste Modellparameter – der „MEAN" genannt wird – bedeutet inhaltlich nichts anderes, als daß im Durchschnitt 71 % der Schüler die RCFP-Einheiten für „wichtig" gehalten haben – jedenfalls nach der Schätzung des Modells. Das Modell sagt uns weiterhin, daß z. B. bei Schülern mit Vortragserfahrung (v) dieser Wert um 3 % steigt und bei Schülern ohne Vortragserfahrung um 3 % fällt– wobei der Einfluß aller anderen Variablen eliminiert ist.

Der konditionale Effekt P7 < V2 bedeutet, daß beim Projekt 7 (MOBI) die Schüler *ohne Vortragserfahrung* (V2) keine so negative Einstellung zum Thema der Einheit hatten, wie man erwarten würde: Der Anteil der Schüler, die diese Einheit schon vorher wichtig fanden, liegt um 13 % höher, als für diese Schülergruppe und dieses RCFP-Projekt zu erwarten gewesen wäre.

Das Modell deckt eine weitere Besonderheit auf. Bei Schülern, bei denen das Projekt P3 (GELT) erprobt wurde, machte sich fehlende Vortragserfahrung (V2) und fehlende Erfahrung mit Gruppenarbeit (G2) ganz besonbemerkbar: Hier lag nämlich der Anteil der Schüler, die das Projekt bereits vor der Erprobung für wichtig gehalten haben um 23 % unter dem Durchschnitt. Dies sieht man an dem konditionalen Modellparameter
P3 < V2 < G2.

(5) Graphische Aufbereitung des Ergebnisses in einem Flußdiagramm:

Wir können nun alle Modellparameter auf die obige Weise mit Inhalt füllen. Es ist jedoch unsere Überzeugung, daß eine Graphik oft mehr sagt als tausend Worte. Dies zeigt die folgende Abbildung 28.

Diese, einem Flußdiagramm nachempfundene, Graphik veranschaulicht durch die Breite der „Ströme" die Bedeutung der jeweiligen Modell-Parameter.

Um den Anteil der Schüler zu erklären, die die RCFP-Einheit schon vor der Erprobung für wichtig hielten, sind vor allem zwei Faktoren entscheidend:

- das Projekt als solches (vor allem P4 = BODEN)
- und die Tatsache, daß bei einem bestimmten Projekt (nämlich bei GELT) jene Schüler, die selten Gruppenarbeit gemacht hatten und auch keine Vortragserfahrung mitbrachten, eine ungewöhnlich starke Abneigung schon vor der Durchführung entwickelten (P3 < V2 < G2).

Zur Schätzung des Schüleranteils, der bereits vor der Erprobung das RCFP-Projekt für wichtig hielt, tragen die Variablen Vortragserfahrung und Gruppenarbeit, für sich genommen (!), also vergleichsweise wenig bei. Nur in Kombination miteinander und bei einem bestimmten Projekt (nämlich bei P3 auch allein, erhalten sie ihre ungewöhnlich große Bedeutung.

Auffallend ist weiterhin, daß beim Projekt „MOBI" (P7) fehlende Vortragserfahrung keine negative Ausgangssituation mit sich bringt – wie man

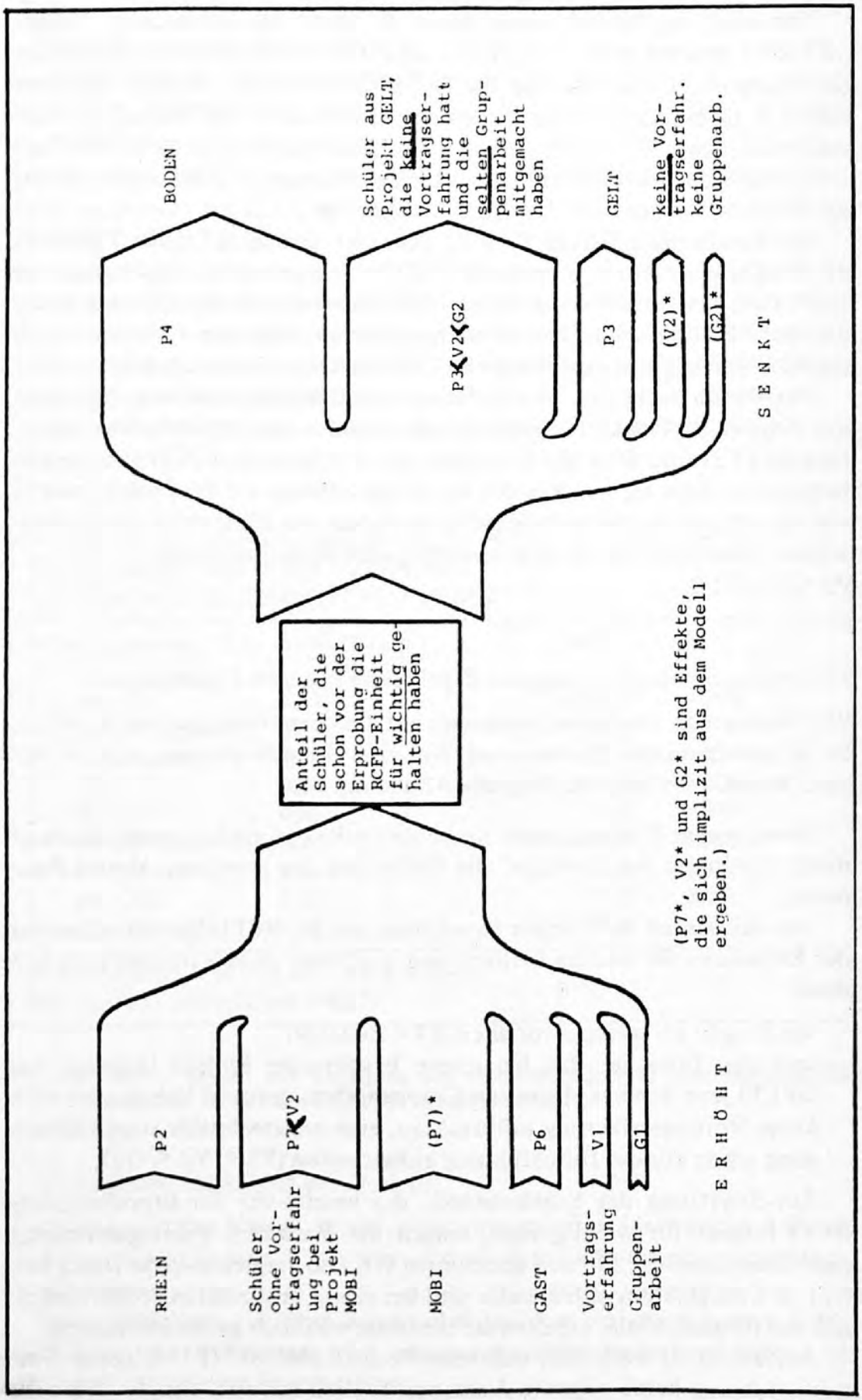


Abb. 28: Flußdiagramm für das GSK-Modell aus Tab. 49 zur Erklärung der Schülererwartungen in Bezug auf das Thema der RCFP-Einheit

erwarten würde. Im Gegenteil: Der Anteil von Schülern, die die Thematik schon vorher wichtig fanden, liegt deutlich *über* dem zu erwartenden Prozentsatz, d. h. der Effekt „fehlender Vortragserfahrung" wirkt beim Projekt MOBI *erhöhend* auf die Motivation vor der Erprobung.

9.3 Ein GSK-Modell für ein nicht-orthogonales Design: Die Anregung der Schüler durch den RCFP-Erprobungsunterricht

In dem vorangegangenen Abschnitt hatten wir die Erwartungshaltung der Schüler in Bezug auf die RCFP-Erhebung zu erklären versucht.

Hier geht es jetzt um ein Modell für die Variable ANREGUNG, mit der erfaßt werden soll, wie gut die RCFP-Einheiten diese Erwartungen erfüllt haben. ANREGUNG – wir erinnern uns – ist der erste Faktor-Score aus unserer Faktorenanalyse des Polaritätsprofils zur RCFP-Einheit. Die (dichotomisierten) Werte dieser Variablen repräsentieren die wichtigste Dimension in den Schülereinstellungen zur vorangegangenen Erprobung der Einheit. Es sei daran erinnert, daß dieser erste Faktor gleichzeitig die beste Reliabilität (Homogenität) des gesamten Polaritätsprofils aufweist. Der Faktor – der hier als Zielvariable fungiert – erfaßt, ob die RCFP-Erprobungen z. B. eher „anregend" oder „stumpfsinnig", eher „erfreuend" oder „bedrückend", eher „interessant" oder „langweilig" empfunden wurden, und ob sie für die Schüler eher „befriedigend" oder „unbefriedigend" waren.

Während unserer multidimensionalen Kontingenztabellenanalyse hat die schrittweise Prädiktorenselektion ergeben, daß drei Variablen zusammengenommen die „Anregung" aus der RCFP-Erprobung optimal erklären: Es handelt sich um das Projekt (P), das Geschlecht (G) und die Klassenstufe (K).

Das entsprechende Baumdigramm ist nochmals wiedergegeben in Abbildung 29.

(1) Vorbemerkung zum Problem eines nicht-orthogonalen Designs:

Die Berechnung eines GSK-Modells für diese Tabelle ist bereits eine relativ schwierige Aufgabe. Während die beiden vorigen Modelle von den Standardoptionen des NONMET-Programms Gebrauch machen konnten, mußten hier die Designmatrizen eigens per Hand erstellt und in das Programm eingegeben werden. Die hier zugrunde liegende Kontingenztabelle weist nämlich eine Besonderheit auf: sie ist nicht orthogonal. Das bedeutet, daß nicht alle Subpopulationen besetzt sind (siehe dazu Abb. 29).

Im Normalfall wird man sich natürlich bemühen, solche nichtorthogonalen Designs zu vermeiden. Man kann z. B. Ausprägungen der Prädiktorvariablen so zusammenfassen, daß sich immer alle logisch notwendigen

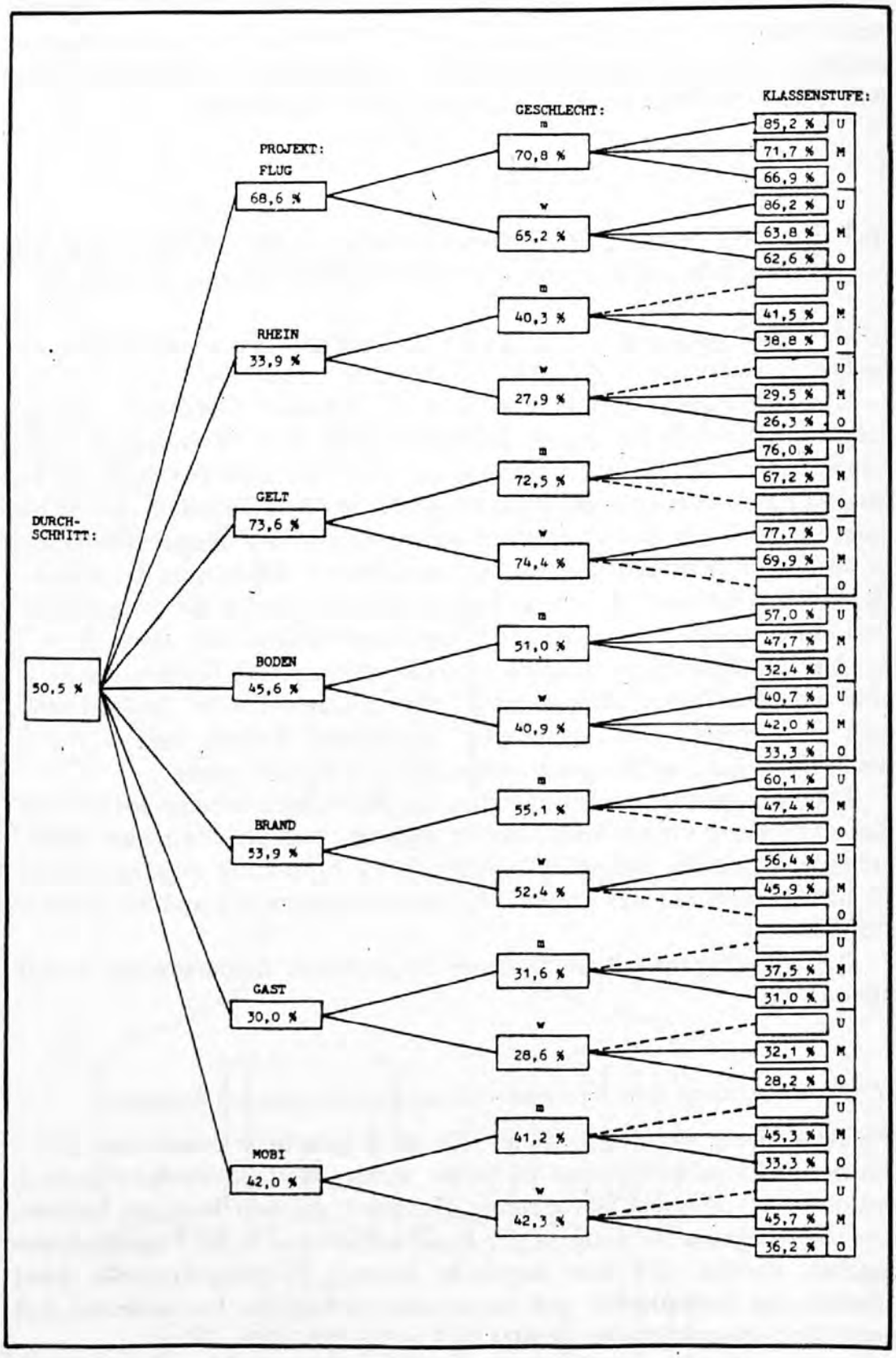


Abb. 29: Baumdiagramm der mehrdimensionalen Kontingenztabelle zur Erklärung der ANREGUNG der Schüler durch den RCFP-Erprobungsunterricht

Subpopulationen (oder Zellen) besetzen lassen. Dies ist freilich nicht immer möglich. In unserem Fall ergibt sich das nichtorthogonale Design gleichsam natürlich. Es folgt einfach aus der Tatsache, daß die RCFP-Einheiten für verschiedene Alterstufen entwickelt wurden und infolgedessen (zum Teil) nicht in der Unterstufe oder der Oberstufe erprobt wurden. Wollte man hier durch Zusammenfassen die Besetzung aller Subpopulationen erreichen, verbliebe nur die Kategorie Mittelstufe, in der alle Einheiten erprobt wurden. Eine erklärende Variable „Klassenstufe" mit dann nur einer Ausprägung (nämlich „Mittelstufe") wäre natürlich Unsinn. Um es nochmals zu verdeutlichen: Die fehlende Orthogonalität ist hier nicht eine Folge sich verkleinernder Fallzahlen durch die Kombination der unabhängigen Variablen, sondern eine *Eigenschaft des zu untersuchenden Gegenstandes.*

Der GSK-Ansatz zeichnet sich nun gerade dadurch besonders aus, daß solche nicht-orthogonalen Designs (und noch andere, sehr viel kompliziertere Tabellen) analysiert werden können, wobei der Aufwand noch einigermaßen vertretbar und auch die Durchschaubarkeit der Analyseschritte noch relativ groß ist. Wir haben die Zielvariable ANREGUNG auch deshalb ausgesucht, um die GSK-Modellbildung bei einem nichtorthogonalen Design demonstrieren zu können.

(2) Erstellung der Design-Matrix bei nicht-orthogonalem Design

Die Erstellung eines GSK-Modells beginnt normalerweise mit der Spezifizierung einer Design-Matrix für den saturierten Fall. Diese Design-Matrix dient der Berechnung aller b-Parameter (oder „Effekte") in der WLS-Schätzung der Wahrscheinlichkeiten der Zielvariablen. Wir hätten in unserem Fall 32 besetzte Subpopulationen, also 32 zu schätzende Haufigkeiten. Andererseits müßten wir eine siebenstufige Variable (Projekt), eine zweistufige (Geschlecht) und eine dreistufige Variable (Klassenstufe) in dieser Design-Matrix kodieren, das bedeutet im Normalfall also 7 mal 2 mal 3 = 42 Subpopulationen. Wir können also *nicht* die Standardkodierung für das saturierte Modell verwenden.

Wir schlagen deshalb einen anderen Weg ein und spezifizieren zunächst eine Design-Matrix, die nur die Haupteffekte des Modells berechnet. Wir beginnen also mit dem sog. Haupteffektmodell und nicht wie üblich mit dem saturierten Modell[10].

Sehen wir uns zunächst die sich ergebende Designmatrix an: (siehe Tab. 50)

An den eingezeichneten Hilfslienen kann man das Konstruktionsprinzip erkennen.

```
GENERATE X MATRIX:
   LE=P(7),G(2),K(3);SUBPOP=0/
DESIGN MATRIX (T X V)
```

	MEAN	P1	P2	P3	P4	P5
1	1.00	1.00	0.00	0.00	0.00	0.00
2	1.00	1.00	0.00	0.00	0.00	0.00
3	1.00	1.00	0.00	0.00	0.00	0.00
4	1.00	1.00	0.00	0.00	0.00	0.00
5	1.00	1.00	0.00	0.00	0.00	0.00
6	1.00	1.00	0.00	0.00	0.00	0.00
7	1.00	0.00	1.00	0.00	0.00	0.00
8	1.00	0.00	1.00	0.00	0.00	0.00
9	1.00	0.00	1.00	0.00	0.00	0.00
10	1.00	0.00	1.00	0.00	0.00	0.00
11	1.00	0.00	0.00	1.00	0.00	0.00
12	1.00	0.00	0.00	1.00	0.00	0.00
13	1.00	0.00	0.00	1.00	0.00	0.00
14	1.00	0.00	0.00	1.00	0.00	0.00
15	1.00	0.00	0.00	0.00	1.00	0.00
16	1.00	0.00	0.00	0.00	1.00	0.00
17	1.00	0.00	0.00	0.00	1.00	0.00
18	1.00	0.00	0.00	0.00	1.00	0.00
19	1.00	0.00	0.00	0.00	1.00	0.00
20	1.00	0.00	0.00	0.00	1.00	0.00
21	1.00	0.00	0.00	0.00	0.00	1.00
22	1.00	0.00	0.00	0.00	0.00	1.00
23	1.00	0.00	0.00	0.00	0.00	1.00
24	1.00	0.00	0.00	0.00	0.00	1.00
25	1.00	0.00	0.00	0.00	0.00	0.00
26	1.00	0.00	0.00	0.00	0.00	0.00
27	1.00	0.00	0.00	0.00	0.00	0.00
28	1.00	0.00	0.00	0.00	0.00	0.00
29	1.00	-1.00	-1.00	-1.00	-1.00	-1.00
30	1.00	-1.00	-1.00	-1.00	-1.00	-1.00
31	1.00	-1.00	-1.00	-1.00	-1.00	-1.00
32	1.00	-1.00	-1.00	-1.00	-1.00	-1.00

	P6	G	K1	K2	
1	0.00	1.00	1.00	0.00	
2	0.00	1.00	0.00	1.00	
3	0.00	1.00	-1.00	-1.00	
4	0.00	-1.00	1.00	0.00	
5	0.00	-1.00	0.00	1.00	
6	0.00	-1.00	-1.00	-1.00	
7	0.00	1.00	0.00	1.00	P2 RHEIN
8	0.00	1.00	-1.00	-1.00	
9	0.00	-1.00	0.00	1.00	
10	0.00	-1.00	-1.00	-1.00	
11	0.00	1.00	1.00	0.00	P3 GELT
12	0.00	1.00	0.00	1.00	
13	0.00	-1.00	1.00	0.00	
14	0.00	-1.00	0.00	1.00	
15	0.00	1.00	1.00	0.00	
16	0.00	1.00	0.00	1.00	
17	0.00	1.00	-1.00	-1.00	
18	0.00	-1.00	1.00	0.00	
19	0.00	-1.00	0.00	1.00	
20	0.00	-1.00	-1.00	-1.00	
21	0.00	1.00	1.00	0.00	
22	0.00	1.00	0.00	1.00	
23	0.00	-1.00	1.00	0.00	
24	0.00	-1.00	0.00	1.00	
25	1.00	1.00	0.00	1.00	
26	1.00	1.00	-1.00	-1.00	
27	1.00	-1.00	0.00	1.00	
28	1.00	-1.00	-1.00	-1.00	
29	-1.00	1.00	0.00	1.00	
30	-1.00	1.00	-1.00	-1.00	
31	-1.00	-1.00	0.00	1.00	
32	-1.00	-1.00	-1.00	-1.00	

Tab. 50: Design-Matrix für die nicht-orthogonale Kontingenztabelle mit den Prädiktoren: Projekt (P), Geschlecht (G) und Klassenstufe (K), sowie der Zielvariablen ANREGUNG durch die RCFP-Einheit

Sehen wir uns zunächst die Kodierung von P (Projekt) an: P1 (also FLUG) wurde an allen Klassenstufen und bei allen Geschlechtern erprobt. Seine sechs Subpopulationen sind besetzt und werden mit 1 kodiert. P2 (RHEIN) wurde *nicht* an einer Unterstufe erprobt; folglich sind nur 4 Subpopulationen besetzt und diese werden mit vier Einsen kodiert. Dies geschieht für alle 6 Projekte, das 7. Projekt (MOBI) ergibt sich implizit und wird nicht kodiert.

Die Kodierung des Geschlechtes ist unproblematisch und erfolgt sinngemäß für die 6 bzw. 4 Subpopulationen je Projekt. Der kritische Punkt ist die Kodierung der Klassenstufe: Sehen wir uns gleich das Projekt RHEIN an (7. bis 10. Zeile in Spalte K1 und K2). Die hier fehlende Unterstufe wird dadurch berücksichtigt, daß K1 keine „1"-Kodierung enthält, sondern nur K2 (Mittelstufe). K3, also die Oberstufe, ergibt sich implizit.

Beim Projekt GELT (11. bis 14. Zeile in Spalte K1 und K2) fehlt dagegen die *Oberstufe*. K1 und K2 enthalten nur die „1" (und die „0"-)Kodierungen. Die „-1"-Kodierung, die implizit K3 (Oberstufe) als vorhanden kennzeichnen würde, fehlt bei K1 und K2. Also ist die dritte, implizite Kategorie von K nicht besetzt.

(3) Berechnung des Haupteffektmodells bei nicht-orthogonalem Design

Die mit der oben beschriebenen Design-Matrix berechnete WLS-Schätzung der 32 vorhandenen Subpopulationen ergibt folgendes Haupteffektmodell:

Modell: Haupteffekte: P1, P2, P3, P4, P5, P6, G, K1, K2

	B-Koeffizienten	χ^2	P
Mittelwert	0,502	5948,49	0,000
P1	0,208	176,75	0,000
P2	-0,127	63,12	0,000
P3	0,194	133,55	0,000
P4	-0,076	23,68	0,000
P5	-0,010	0,35	0,552*
P6	-0,136	41,25	0,000
G	0,023	13,49	0,000
K1	0,080	40,45	0,000
K2	-0,009	1,11	0,291*

2-Fehler = 31,47, DF = 22, P = 0,087
*: nicht signifikant

Tab. 51: Haupteffekt-Modell für die nicht-orthogonale Kontingenztabelle entsprechend der Abb. 29 (1. Schritt des „model-fitting")

Das Modell ist auf Anhieb ganz gut gelungen:

Die Residuen sind (mit P = 0,087) zwar noch viel zu groß, aber auf dem 5 %-Niveau bereits nicht mehr signifikant. Bis auf P5 und K2 sind alle Haupteffekte höchstsignifikant.

(4) Modell-fitting bei nicht-orthogonalem Design

Der nächste Schritt ergibt sich zwangsläufig: Wir müssen ein Haupteffektmodell spezifizieren, daß die nicht-signifikanten Effekte P5 (BRAND) und K2 (Mittelstufe) *nicht* enthält. Dazu müssen wir wieder eine Design-Matrix konstruieren, bei der P5 und K2 eliminiert sind:

Wir können hier auf die Konstruktion dieser Matrix aus Platzgründen nicht näher eingehen.

Berechnet man aber mit dieser Design-Matrix erneut ein Modell, erhält man schon eine wesentlich bessere Anpassung:

Modell: Haupteffekte: P1, P2, P3, P4, P6, G, K1

	B-Koeffizienten	χ^2	P
Mittelwert	0,501	6126,10	0,000
P1 (FLUG)	0,205	179,84	0,000
P2 (RHEIN)	-0,130	72,67	0,000
P3 (GELT)	0,195	137,76	0,000
P4 (BODEN)	-0,076	23,83	0,000
P6 (GAST)	-0,136	50,79	0,000
G (MAENNLICH)	0,023	13,11	0,000
K1 (UNTERSTUFE)	0,073	47,02	0,000

2-Fehler (Residuen): 32,94, DF = 24; P = 0,11

Tab. 52: Modell der signifikanten Haupteffekte (2. Schritt des „model-fitting")

Alle in dem Modell enthaltenen Effekte sind nun höchstsignifikant. Die Fehler, die das Modell bei der Schätzung macht, sind etwa gleich groß wie vorhin (χ^2-Residuen 32,9). Da wir jetzt aber nur 7 Effekte (+ Mittelwert) haben, also 32−8=24 Freiheitsgrade für die Schätzung, sind die Residuen nun statistisch *nicht* mehr signifikant (p = 0,11). Allerdings sollten die Residuen bei etwa $p > 0{,}20$ nicht-signifikant sein, d. h. unser jetziges Modell ist zwar besser als das vorige, aber sicher noch nicht optimal.

(5) Berechnung des besten Modells

Die weitere Verfeinerung des Modells ist hier beim nicht-orthogonalen Fall mit großem Aufwand verbunden, da ja für jedes neue Modell per Hand eine Design-Matrix konstruiert und in das Programm eingegeben werden muß.

Wir stießen erst nach einer langwierigen Analyse der Residuen auf ein statistisch und inhaltlich sehr gutes Modell. Es enthält neben den Haupteffekten zusätzlich nur einen einzigen konditionalen Effekt, nämlich K1 < G2 < P4. Damit konnten wir das „model-fitting" bei folgendem Modell abbrechen:

Modell: Haupteffekte: P1, P2, P3, P4, P6, G, K1 konditionaler Effekt: K1 < G2 < P4			
Modellparameter	B-Koeffizienten	χ^2	P
MEAN	0,50	5994,4	0,000
P1 (FLUG)	0,20	176,0	0,000
P2 (RHEIN)	-0,13	74,1	0,000
P3 (GELT)	0,19	119,7	0,000
P4 (BODEN)	-0,06	14,2	0,000
P6 (GAST)	-0,14	49,9	0,000
G (Männlich)	0,02	9,0	0,003
K1 (Unterstufe)	0,08	53,0	0,000
K1 < G2 < P4	-0,09	6,8	0,009
χ^2-Residuen = 26,1897, DF = 23, P = 0,29		G2 implizit = -0,02	
P7 implizit = -(P1+P2+P3+P4+P6) ≈ -0,06		K3 implizit ≈ 0,07,	

Tab. 53: Bestes Modell zur Erklärung der nicht-orthogonalen Kontingenztabelle entsprechend der Abb. 29 (letzter Schritt des „model-fitting")

Dieses Modell hat eine ausreichende Anpassung an die empirischen Daten (P = 0,29 für den χ^2-Wert der Residuen) und kommt mit nur acht Effekten (plus dem MEAN) aus. Das heißt: Wir sind, nach der statistischen Logik des Modells, berechtigt, die *32* vorhandenen Subpopulationen durch die *acht* Effekte (plus den Mittelwert) zu ersetzen.

(6) Die inhaltliche Auffüllung des Modells

Im Durchschnitt schätzt das Modell den Anteil „angeregter" Schüler auf 50%. Bei der Einheit FLUG erhöht sich der Anteil um 20%, bei der Einheit GELT um 19%. Wenig angeregt waren die Schüler bei den Einheiten RHEIN (−13%), GAST(−14%), BODEN (−6%) und MOBI (die sich implizit ergibt: −6%).

Die Variable „Klassenstufe" hat nur insofern einen Einfluß auf die „Anregung" der Schüler durch die Erprobung, als bei Unterstufenschülern der Anteil angeregter Schüler um 8% größer ist, und bei Oberstufenschülern um ca. 8% kleiner. Die Mittelstufenschüler fühlten sich dagegen durchschnittlich angeregt.

Das Modell deckt eine auffallende Besonderheit bei Projekt P4 (BODEN) auf: Die *weiblichen Unterstufenschüler* wurden von dieser RCFP-Einheit zu 8% weniger stark angeregt, als dies bei dieser Einheit typisch war. (K1 < G2 < P4).

(7) Graphische Aufbereitung des Ergebnisses

Durch die obige Modellbildung sind wir nun in der Lage, die komplizierteren Zusammenhänge der mehrdimensionalen Kontingenztabelle durch ein einfaches Flußdiagramm zu veranschaulichen: (siehe Abb. 30)

Das Flußdigramm zeigt, daß vor allem natürlich die jeweilige RCFP-Unterrichtseinheit dafür verantwortlich ist, ob die Schüler angeregt waren oder nicht. Einen durchschlagend *positiven* Effekt hatten die Einheiten FLUG und GELT: ca. 20% mehr Schüler als man erwarten würde fühlten sich durch diese Einheiten angeregt. Oder anders formuliert: Aufgrund der Geschlechterproportion ihrer Erprobungspopulation und aufgrund der Klassenstufen, in denen die Einheiten erprobt wurden, müßte der Anteil angeregter Schüler um ca. 20% kleiner sein, wenn diese beiden Einheiten genauso erfolgreich sein sollten, wie die anderen RCFP-Einheiten im Durchschnitt tatsächlich waren.

Man muß sich – um diese Aussage richtig bewerten zu können – nochmals klarmachen, daß dies der *bereinigte* Effekt der jeweiligen RCFP-Einheiten ist. Nicht die Erdkundenoten, nicht Erfahrung mit Rollenspielen, nicht fehlende Vortragserfahrung oder Erfahrung mit Gruppenarbeit, nicht eine besondere Geschlechterproportion bei der Schülerstichprobe und auch nicht Besonderheiten in der Klassenstufe bei den jeweiligen Erprobungen könen für das gute Abschneiden dieser beiden RCFP-Einheiten verantwortlich gemacht werden. Zur Erinnerung: Die ersten vier Variablen wurden bei der *Clark-Higgins-Koch*-Variablenselektion als irrelevant ausgeschlossen: Das „Geschlecht" und die „Klassenstufe" erwiesen sich jetzt hier beim GSK-Modell als zwar signifikant, aber quantitativ relativ unbedeutend in ihrem Einfluß.

Beim Projekt P4 (BODEN) kommt es zu einer Interaktion zwischen den Variablen „Geschlecht", „Klassenstufe" und „Projekt": Bei diesem Projekt fühlt sich die Gruppe der *Mädchen aus der Unterstufe seltener angeregt*, als man es eigentlich erwarten würde (–9%). Der Anteil angeregter Schüler war bei den Projekten P6 (GAST) und P2 (RHEIN) deutlich niedriger als im Durchschnitt der Projekte.

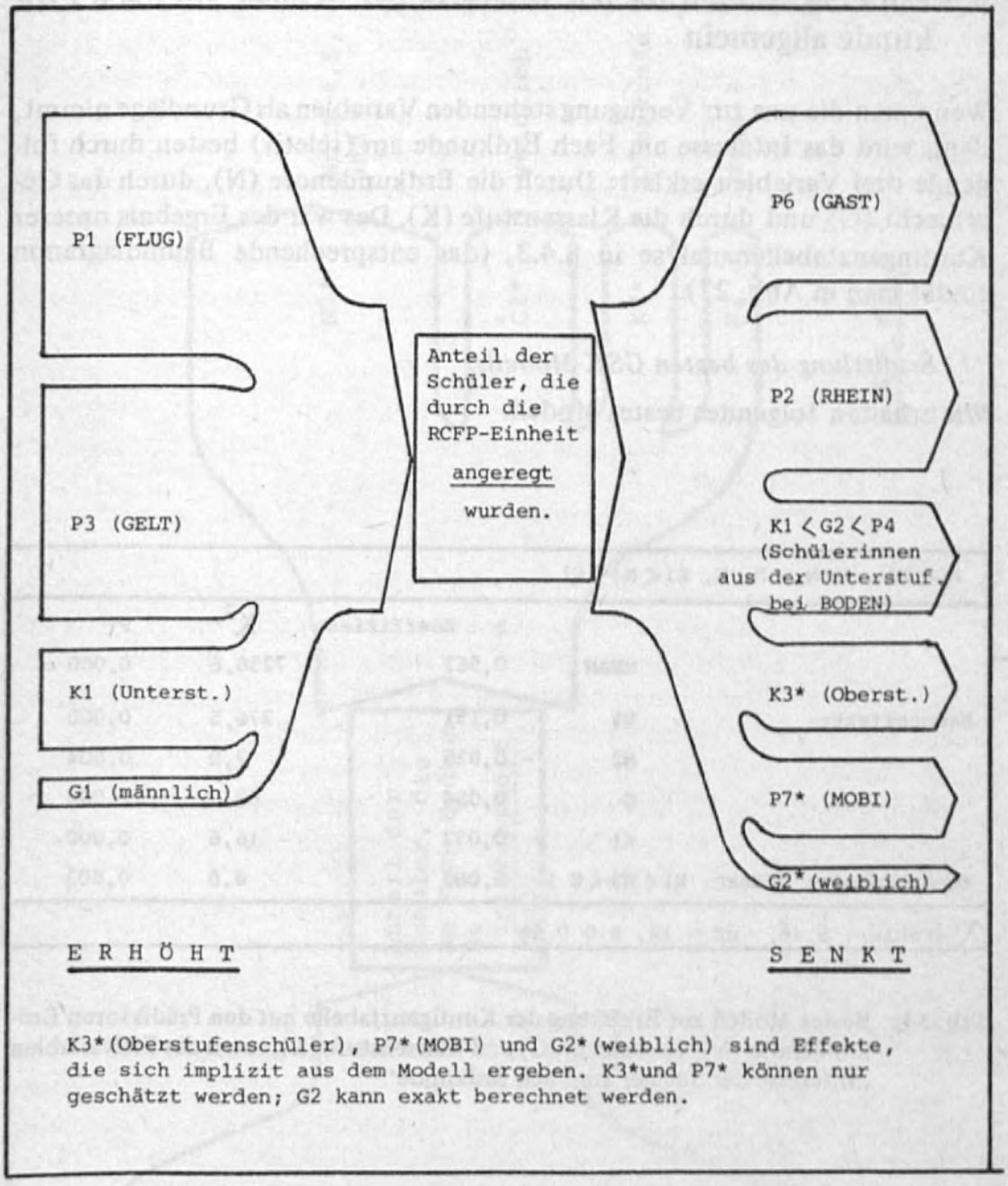


Abb. 30: Flußdiagramm zur Veranschaulichung des besten Modells nach Tab. 53

9.4 Ein GSK-Modell für das Interesse der Schüler am Fach Erdkunde allgemein

Wenn man die uns zur Verfügung stehenden Variablen als Grundlage nimmt, dann wird das Interesse am Fach Erdkunde am (reletiv) besten durch folgende drei Variablen erklärt: Durch die Erdkundenote (N), durch das Geschlecht (G) und durch die Klassenstufe (K). Das war das Ergebnis unserer Kontingenztabellenanalyse in 8.4.3, (das entsprechende Baumdiagramm findet man in Abb. 27).

(1) Ermittlung des besten GSK-Modells:

Wir erhalten folgendes bestes Modell.

MODELL: MAIN = N, G, K1<N3<G1				
		b - Koeffizient	χ^2	P
	MEAN	0,567	7236,6	0,000
Haupteffekte:	N1	0,151	276,5	0,000
	N2	- 0,025	7,9	0,004
	G	0,054	69,2	0,000
	K1	0,037	16,6	0,000
Konditionaler Effekt: K1 < N3 < G 1		0,083	8,8	0,003
χ^2-Fehler: 9,46, df = 12, P 0 0,66				

Tab. 54: Bestes Modell zur Erklärung der Kontigenztabelle mit den Prädiktoren Erdkundenote (N), Geschlecht (G) und Klassenstufe (K), sowie der Zielvariablen „Interesse der Schüler am Fach Erdkunde"

Dieses Modell ist – sowohl nach statistischen wie inhaltlichen Kriterien – ganz hervorragend: Es reduziert die Zahl der Effekte von 18 auf 5 explizite und 2 implizite Effekte, d. h. wir können die multidimensionale Kontingenztabelle zum Interesse der Schüler am Fach Erdkunde statt mit 18 Subpopulationen (bzw., was prinzipiell das gleiche ist: mit 18 Effekten) fast genauso gut mit nur 5 (+ 2) Effekten beschreiben. Der Fehler, den wir dabei machen, ist (mit χ^2-Fehler: 9,46) bei weitem nicht mehr signifikant. Außerdem sind die Effekte auch inhaltlich sehr plausibel. Dies sieht man dann besonders deutlich, wenn wir das Modell wieder im Form des Flußdiagramms darstellen:

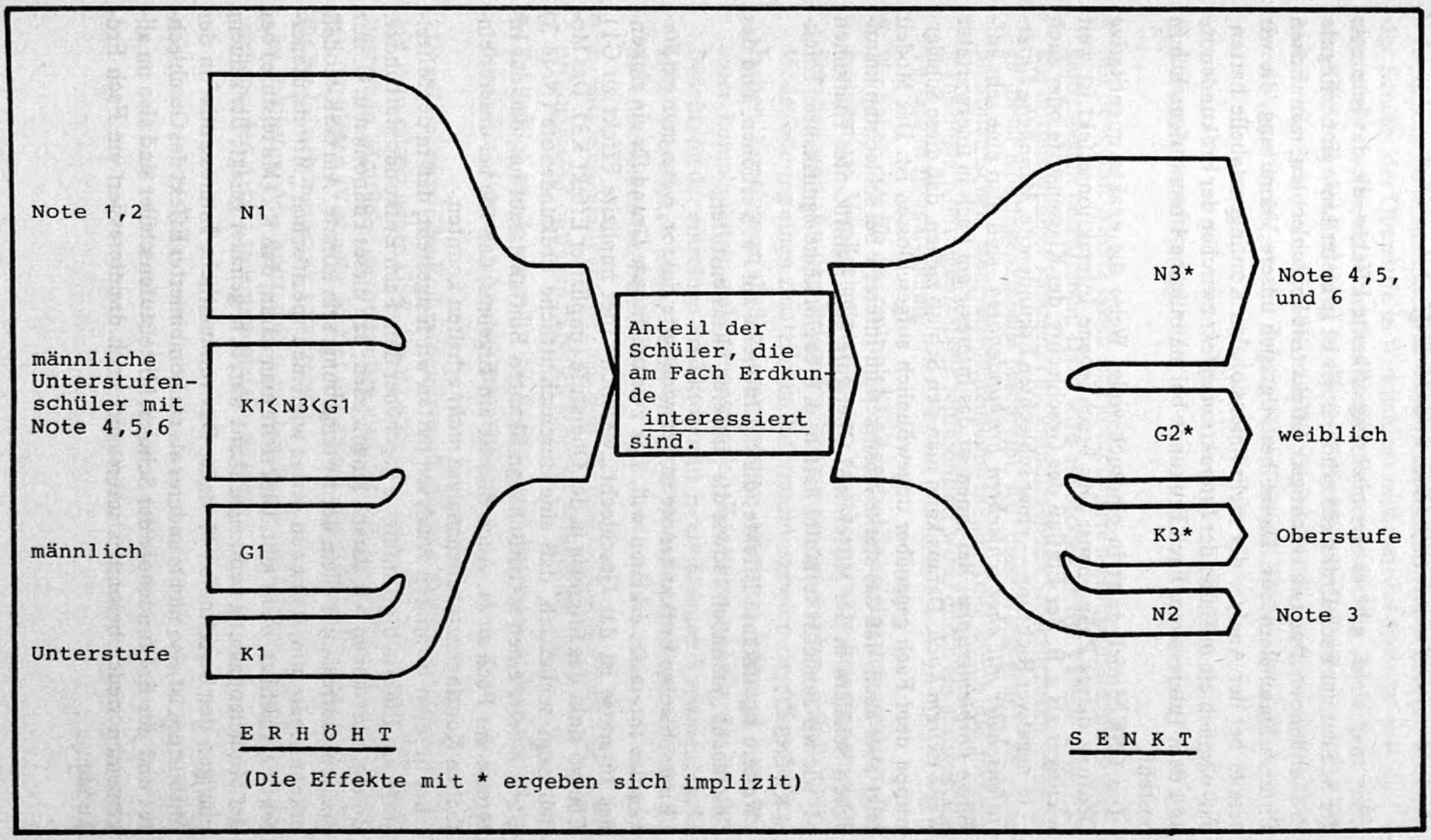


Abb. 31: Flußdiagramm zur Veranschaulichung des besten Modells nach Tab. 54

(2) Aufbereitung und Interpretation des Modells:

Wie man sieht, gibt es vier höchstsignifikante Effekte, die das Interesse der Schüler am Fach Erdkunde erhöhen: Es ist in erster Linie eine sehr gute Erdkundenote. Der zweitwichtigste Effekt tritt bei schlechten, männlichen Unterstufenschülern auf. Damit bestätigt sich unsere Vermutung, die wir bereits bei der Analyse der mehrdimensionalen Kontingenztabelle hatten, daß nämlich ein auffallender *Interaktionseffekt* zwischen der Erdkundenote und dem Interesse am Fach Erdkunde bei männlichen Unterstufenschülern besteht.

Das GSK-Modell zeigt in eindrucksvoller Weise die exakte quantitative Deutung dieses Phänomens: Der beschriebene Interaktionseffekt ist weit wichtiger als z. B. der Einfluß des Geschlechts, der Klassenstufe oder auch – in negativer Richtung – einer schlechten Erkundenote. *Männliche Unterstufenschüler mit eher schlechten Erdkundenoten* sind also eine sehr auffällige *Problemgruppe*. Bei ihnen ist das Interesse am Fach in unerwarteter Weise *extrem hoch.* Daraus kann man den Schluß ziehen, daß diese Schülergruppe dem Fach gegenüber ungewöhnlich aufgeschlossen ist. Das Modell zeigt aber auch, daß das starke Absinken im Interesse bei schlechten männlichen Schülern in der Mittel- und Oberstufe *kein* signifikantes Phänomen ist, wie wir zunächst vermutet hatten, d. h. es gibt keine signifikanten Effekte K3<N3<G1!

Weitere signifikante Effekte, die das Interesse am Fach erhöhen, sind das Geschlecht „männlich" sowie die „unteren" Klassenstufen.

Eine schlechte Erdkundenote ist der wichtigste Faktor, wenn man ein *Sinken* des Interesses erklären will. Der zweitwichtigste Grund für ein sinkendes Interesse ist das Geschlecht „weiblich" (der implizite Effekt zu G1). Ebenso sinkt das Interesse in der Oberstufe (impliziter Effekt K3). Das Modell zeigt schließlich, daß eine durchschnittliche Erdkundenote (Note 3) *allein* bereits einen statistisch signifikanten Einfluß darauf hat, daß das Interesse am Fach *sinkt.* Auch dies ist ein Ergebnis, daß wie bei unserer einfachen Kontingenztabellenanalyse *nicht* erhalten konnten.

Bei unseren bivariaten Analysen hatten wir festgestellt, daß in der Mittelstufe ein Einbruch im Interesse gegenüber dem Fach Erdkunde stattfindet. Bereits dort haben wir darauf hingewiesen, daß dieses Phänomen u. U. nur ein Scheineffekt, vor allem der Notengebung, sein könnte. Am GSK-Modell erkennt man nun, daß es so etwas wie einen spezifischen „Mittelstufeneffekt" tatsächlich *nicht* gibt. Dies sieht man daran, daß K2 (Mittelstufe) bei der Modellanpassung nicht signifikant war; d. h. genauer gesagt: Bei keinem einzigen der berechneten Modelle. Das vermeintliche Interessentief in der Mittelstufe ist also nichts anderes als ein kombinierter Effekt des Geschlechtes und der Erdkundenote der Schüler. Mittelstufenschüler sind also im allgemeinen weder besonders interessiert noch desinteressiert am Fach Erdkunde.

In diesem Zusammenhang ist auch darauf hinzuweisen, daß das Interesse der Schüler der Oberstufe in Wirklichkeit *nicht* wieder ansteigt, wie wir bei den oben erwähnten bivariaten Analysen gemeint hatten. Auch dies ist nur ein Effekt der Veränderung in den Geschlechterproportionen, sowie in den Noten der Schülerstichproben aus der Oberstufe. Tatsächlich fällt das Interesse sogar leicht, wie der negative (implizite) K3-Effekt anzeigt.

9.5 Zusammenfassung

- Mit dem GSK-Ansatz lassen sich komplizierte mehrdimensionale Kontingenztabellen beschreiben. Dazu wird ein Modell an die Tabelle angepaßt, das mittels weniger Parameter die grundlegenden Strukturzusammenhänge dieser Kontingenztabelle beschreibt. Auf diese Weise gelingt es, einen nichtmetrischen Mehrvariablenzusammenhang (der durch die mehrdimensionale Kontingenztabelle repräsentiert wird) durch ein statistisches Verfahren zu analysieren.
- Wir haben dies für den multiplen Fall an drei Beispielen demonstriert. Die multivariate Verallgemeinerung des Verfahrens wäre keine prinzipielle Schwierigkeit. An den Beispielen konnte gezeigt werden, daß die GSK-Modellbildung einen zusätzlichen Informationsgewinn gegenüber der einfachen Kontingenztabellen-Interpretation erbringt: Während man bei einer komplizierten mehrdimensionalen Kontingenztabelle die relative Bedeutung der einzelnen Prädiktoren (d. h. erklärenden Variablen) nicht mehr überblicken kann, liefert ein GSK-Modell quantitative Größen für die entscheidenden Effekte.
- Es wurde ebenfalls gezeigt, daß sich mit dem GSK-Ansatz auch für nichtorthogonale Kontingenztabellen Modelle entwickeln lassen.
- Schließlich wurde versucht, die Rechenergebnisse aus der Modellbildung durch Flußdiagramme zu veranschaulichen.

Damit haben wir die Reihe unserer methodischen Verbesserungsvorschläge mit dem relativ neuen GSK-Ansatz abgeschlossen. Dabei war es nicht unser Bestreben, sämtliche methodischen und statistischen Feinheiten der Verfahren dazulegen. Vielmehr ging es uns darum, die Anwendung der Methoden für die Forschungspraxis aufzuzeigen.

Im nächsten Abschnitt sollen nochmals die wichtigsten Ergebnisse der vorliegenden Arbeit zusammengefaßt werden.

10. Die wichtigsten Ergebnisse

Die Arbeit hatte das Ziel zur methodischen Weiterentwicklung der empirischen Forschung in der Didaktik der Geographie, und der sozialwissen-
sich diese Mängel – zumindest zum Teil – vermeiden lassen.
versucht:

Erstens analysierten wir vorliegende empirische Untersuchungen zur Geographiedidaktik nach methodischen Gesichtspunkten.

Zweitens führten wir eine Sekundäranalyse von Erhebungsdaten des Raumwissenschaftlichen-Curriculum Forschungsprojektes durch.

Der erste Schritt diente dazu, methodische Mängel in der bisherigen Forschung aufzudecken. In einem zweiten Schritt sollten Verfahren und Techniken anhand einer konkreten Datenanalyse vorgestellt werden, mit denen sich diese Mängel – zumindest zum Teil – vermeiden lassen.

Dabei handelt es sich unter anderem um ein statistisches Modell zur multivariaten Zusammenhangsanalyse bei nichtmetrischen Variablen.

Entsprechend diesen beiden Schritten kann man auch die Ergebnisse der vorliegenden Arbeit nach zwei Gesichtspunkten unterteilen: Zum einen handelt es sich um die Schlußfolgerungen unserer methodischen Bestandsaufnahme. Zum anderen lassen sich die inhaltlichen Ergebnisse zusammenfassen, die sich aus der Analyse der RCFP-Erhebungen mittels verschiedener, zum Teil neuerer Methoden ergeben haben.

10.1 Ergebnisse der methodischen Analyse vorliegender empirischer Arbeiten zur Geographiedidaktik

Der methodische Beitrag der vorliegenden Arbeit umfaßt drei Aspekte: Erstens, die Analyse empirischer Untersuchungen auf dem Hintergrund einer „ganzheitlichen" Betrachtungsweise des empirischen Forschungsprozesses. Zweitens, die Zusammenstellung *typischer Mängel* und Schwachstellen bisheriger Untersuchungen. Und drittens, die Erarbeitung von *Verbesserungsvorschlägen* in Form verschiedener Auswertungsverfahren.

(1) Unausgewogenheit

Das wichtigste Ergebnis unserer methodischen Bestandsaufnahme und Analyse war die Feststellung, daß fast alle empirischen Untersuchungen eine sehr starke methodische Unausgewogenheit aufwiesen.

Während einzelne Schritte dieser Arbeiten auf, zum Teil, höchstem methodischen Niveau durchgeführt wurden, erreichten andere Teile nicht einmal die Mindestanforderungen: So wurden höchst anspruchsvolle statistische Analyseverfahren auf völlig unzulänglicher Datenbasis durchgeführt.

Oder man erstellte Kreuztabellen, die wegen einer extremen Verzerrung der Grundgesamtheit, nur Scheinzusammenhänge darstellten. Ein letztes Beispiel: Es wurden außerordentlich umfangreiche Variablensätze erhoben, ohne daß man die Reliabilität und Validität dieser Variablen irgendwie überprüft und abgesichert hätte.

Da wir von der Annahme ausgingen, daß eine empirische Untersuchung nur so „stark" sein kann, wie „ihr schwächstes Glied", zogen wir aus obiger Feststellung folgende Konsequenz: Der wichtigste Beitrag für eine methodische Verbesserung der empirischen Forschung wäre eine *Angleichung* der Untersuchungskomponenten in ihrem methodischen Niveau.

(2) Typische Mängel

Ein zweites Ergebnis der vorliegenden Arbeit ist die Identifizierung typischer Schwachstellen im empirischen Forschungsprozeß. Die grundlegende methodische Unausgewogenheit ist das allgemeine Kennzeichen dieser häufig auftretenden Mängel. Im einzelnen handelt es sich um folgendes:

- Der zu geringe Umfang und/oder die unzulängliche innere Struktur der Stichproben,
- die fehlende Überprüfung der Meßqualität von Variablen,
- die Nichtbeachtung der sehr großen Gefahr von Scheinzusammenhängen bei der bivariaten Analyse,
- die Überbewertung, ja Verselbständigung komplexer statistischer Analyseverfahren und
- die Vernachlässigung nichtmetrischer Auswertungsmethoden, bei gleichzeitiger übertriebener Verwendung metrischer Verfahren.

(3) Verbesserungsvorschläge

Für die oben genannten, häufig auftretenden Mängel wurden von uns Verbesserungsvorschläge erarbeitet und anhand der Sekundäranalyse von RCFP-Daten vorgestellt.

- Es wurden Hinweise zur Zusammenführung von Datensätzen aus Einzelerhebungen zu einem Gesamt-Datensatz gegeben, bzw. Hinweise zur Aggregierung von Daten auf unterschiedlichen Erhebungsniveaus (z. B. Lehrer-Schüler-Daten).
  Dies könnte beitragen zu einer Vergrößerung und Strukturverbesserung von Stichproben.
- Der zweite Verbesserungsvorschlag betrifft eine relativ einfache Vorgehensweise zur (nachträglichen) Überprüfung und Erhöhung der Meßqualität von Variablen mit Mehrfachindikatoren. Eine Kombination aus traditioneller Kennwertanalyse und Faktorenanalyse wurde dafür vorge-

schlagen. An den RCFP-Daten wurde gezeigt, wie sich damit eine Reliabilitäts- und Dimensionalitätsanalyse praktisch durchführen läßt.

- Um die Gefahr von Scheinzusammenhängen bei einer bivariaten Analyse auch bei nichtmetrischen Variablen zu verringern, wurde an eine Methode erinnert, die seit Jahrzehnten bekannt ist, aber dennoch in keiner der von uns durchgesehenen geographiedidaktischen Untersuchungen benutzt wurde. Es handelt sich um die Methode der Einführung einer Kontrollvariablen, mit der Partialtabellen erstellt und partielle Zusammenhangskoeffizienten berechnet werden können.
- Ein letzter Verbesserungsvorschlag wandte sich gegen die Überbewertung metrischer Analyseverfahren. Es wurden dazu zwei neuere statistische Methoden vorgestellt, mit denen sich multivariate Zusammenhangsanalysen auch bei nichtmetrischen Variablen durchführen lassen. Bei den Verfahren handelt es sich um die sog. „Clark-Higgins-Selektion Procedure" und den „GSK-Ansatz".

10.2 Ergebnisse aus der inhaltlichen Analyse der RCFP-Daten

Neben verschiedenen Einzelergebnissen ergab die Sekundäranalyse der RCFP-Daten folgende allgemeinere Resultate:

(1) Meßqualität der Variablen

Die wichtigsten, vom RCFP bei allen Erhebungen eingesetzten Meßinstrumente – eine Einstellungsbatterie und ein Polaritätsprofil – konnten ihrer Aufgabe nur teilweise gerecht werden. Die Reliabilitäts- und Dimensionalitätsanalyse ergab, daß die Itembatterie fünf verschiedene Einstellungsdimensionen bei den Schülern ansprach, wobei zwei Dimensionen, mit jeweils nur zwei Items, unzulängliche Meßqualität aufwiesen.

Beim Polaritätsprofil zeigten die Analysen, daß jeweils vier Einstellungsdimensionen der Schüler angesprochen wurden. Auch hier hatten zwei Dimensionen ungenügende Meßqualität (d. h. zu wenig Items und zu geringe Homogenität).

Außerdem ergab sich, daß das Polaritätsprofil besser dazu geeignet war, die Einstellungen der Schüler in bezug auf die RCFP-Erprobungseinheiten zu erfassen, als die Meinungen der Schüler zum Fach Erdkunde allgemein.

Es wurde erörtert, wie man das Polaritätsprofil zu Beginn der Untersuchung hätte verbessern können.

(2) Aufdeckung von Scheinzusammenhängen

Die bivariate Analyse einiger Variablenzusammenhänge konnte zeigen, daß nahezu alle Einstellungsvariablen im RCFP-Datensatz durch grundlegende Strukturvariablen (wie: Geschlecht, Klassenstufe, Erprobungsprojekt, usw) verzerrt werden, weil die Erprobungspopulationen nicht danach kontrolliert worden waren.

An mehreren Beispielen wurde demonstriert, daß eine bivariate Betrachtungsweise bei survey-ähnlichen Erhebungen, wie der des RCFP, sehr schnell an ihre Grenzen stößt, da sich die Bedeutung direkter und indirekter Variablenzusammenhänge nicht mehr abschätzen läßt. Jede bivariate Betrachtung eines Variablenzusammenhanges bei den RCFP-Daten führt deshalb nahezu zwangsläufig zu falschen Ergebnissen.

Aus den Ergebnissen der bivariaten Analysen mußte der Schluß gezogen werden, daß nur eine Mehrvariablenanalyse Klarheit in die vielfältig verflochtenen Zusammenhänge bringen kann. In einem ersten Schritt wurden dazu zwei bivariate Variablenzusammenhänge durch Einführung einer Kontrollvariablen analysiert: Es konnte gezeigt werden, daß der Zusammenhang zwischen der Klassenstufe und der Erwartungshaltung der Schüler in Bezug auf die Thematik der bevorstehenden RCFP-Erprobungseinheit (je höher die Klassenstufe, desto höher die Erwartungen) teilweise ein Scheinzusammenhang ist. Er wird zum Teil „verursacht" durch die schulische Erfahrung der Schüler: Schüler mit Vortragserfahrung haben höhere Erwartungen an die RCFP-Einheiten *und* befinden sich häufiger in höheren Klassenstufen. Bei zwei anderen Variablen konnte zunächst kein Zusammenhang festgestellt werden: Schüler die *vor* Erprobung hohe Erwartungen an die RCFP-Einheit hatten, fanden die Einheit *nach* Erprobung genauso anregend (oder nicht anregend) wie Schüler, die niedrige Erwartungen hatten. Erst nach Einführung der Kontrollvariablen „Geschlecht" zeigten sich signifikante Unterschiede: Mädchen, die vor Erprobung das Thema wichtig gefunden hatten, waren nach Erprobung häufiger „nicht angeregt", d. h. fühlten sich in ihren Erwartungen enttäuscht.

(3) Ergebnisse der Mehrvariablenanalyse bei nichtmetrischen Variablen

Die Durchführung von Mehrvariablenanalysen mit zwei relativ neuen Verfahren zur Kontingenztabellenanalyse erbrachte folgendes:

- Die *Erwartungshaltung* der Schüler in Bezug auf das Thema der RCFP-Einheit wird am stärksten beeinflußt durch die Attraktivität oder „Unattraktivität" des Themas. Dies gilt aber *nur* für die Projekte BODEN, RHEIN, MOBI und GELT. Bei den Projekten FLUG und GAST dagegen spielte dies keine Rolle.
  Beim Projekt GELT hatte die Unattraktivität des Themas aber nur bei jenen Schülern einen Einfluß auf ihr Urteil, die keine Vortragserfahrung

und auch keine Gruppenarbeit mitgemacht hatten. Die Mehrvariablenanalyse deckte auch auf, daß beim Projekt MOBI gerade Schüler *ohne* Vortragserfahrung häufiger hohe Erwartungen an die Einheit hatten. Alle anderen betrachteten Variablen, wie Geschlecht, Klassenstufe, Erdkundenote, standen in *keinem* Zusammenhang mit den Erwartungen der Schüler.

- Der Anteil der Schüler, die sich durch die jeweilige RCFP-Erprobungseinheit angeregt gefühlt haben (nach der Erprobung erhoben), wird primär durch die Einheit selbst bestimmt. Da bei bestimmten Projekten häufiger Oberstufenschüler, bei anderen Projekten häufiger Schüler mit guten Erdkundenoten und bei wieder einem anderen Projekt hauptsächlich Schüler beteiligt waren, die besonders stark an Erdkunde interessiert waren, hätte man vermuten können, daß der „Erfolg" eines Projektes bei den Schülern auch mit diesen Variablen zusammenhängt. Dies konnte widerlegt werden. Neben dem Projekt selbst hat *nur* das Geschlecht und die Klassenstufe einen (kleinen) Einfluß auf den „Erfolg". Mädchen waren generell – also bei allen Projekten und allen Klassenstufen – seltener angeregt als Jungen. Unterstufenschüler waren (unabhängig vom Projekt und vom Geschlecht) häufiger „angeregt", Oberstufenschüler dagegen seltener.
  Beim Projekt BODEN trat jedoch eine Besonderheit auf: Hier waren Schülerinnen aus der Unterstufe weitaus seltener durch das Projekt angeregt, als man nach den generellen Tendenzen erwartet hätte.
  Unabhängig vom Geschlecht und der Klassenstufe fühlten sich die Schüler der FLUG und der GELT-Einheit weitaus häufiger durch die Einheit angeregt als dies sonst der Fall war. Für die Schüler der Einheit GAST und RHEIN gilt das Gegenteil: bei ihnen war der Anteil angeregter Schüler deutlich kleiner.
- Unter den von uns untersuchten Variablen hatte das *Interesse am Fach* Erdkunde nur Zusammenhänge mit dem Geschlecht, der Klassenstufe und der Erdkundenote. Die multiple Betrachtungsweise zeigte, daß primär die Erdkundenote mit dem Interesse am Fach zusammenhängt. Bei guten und sehr guten Schülern ist der Anteil interessierter Schüler – unabhängig vom Geschlecht und der Klassenstufe – deutlich höher, bei schlechten und auch bei durchschnittlichen (!) Schülern ist er deutlich geringer. Mädchen sind (unabhängig von der Erdkundenote und unabhängig von der Klasstenstufe) weniger häufig interessiert als Jungen. Eine auffällige Sondergruppe sind männliche Unterstufenschüler mit schlechten Erdkundenoten: Sie sind weitaus häufiger am Fach interessiert, als man aufgrund der obigen Tendenzen vermuten würde.
  Unabhängig vom Geschlecht und unabhängig von der Erdkundenote sind Oberstufenschüler häufiger am Fach Erdkunde interessiert, Unterstufenschüler dagegen seltener.

ZUSAMMENFASSUNG

Ziel der vorliegenden Arbeit ist es, zur methodischen Weiterentwicklung der empirischen Forschung in der Geographiedidaktik beizutragen.

Dazu werden Daten aus der umfangreichsten Serie von Schulversuchen im Fach Geographie ausgewertet, die je in Europa durchgeführt wurde. Es handelt sich dabei um das „Raumwissenschaftliche Curriculum Forschungsprojekt" (RCFP), das in seinem Umfang mit dem amerikanischen High School Geography Projekt (HSGP) vergleichbar ist.

Anhand der RCFP-Erhebung werden die Möglichkeiten verschiedener quantitativer Techniken zur Überprüfung der Meßqualität von Variablen, zur Erstellung eindimensionaler Fragebatterien und zur bivariaten Analyse von Variablenzusammenhängen dargestellt. Der Schwerpunkt der Arbeit liegt bei der Diskussion multivariater Verfahren zur Analyse nichtmetrischer Daten: Es wird gezeigt, daß sich durch ein neueres statistisches Verfahren – den sog. GSK-Ansatz – Zusammenhänge zwischen mehreren qualitativen Variablen aufdecken lassen, die bei bivariater Betrachtungsweise verborgen blieben. Der Zusammenhang zwischen Einstellungen von Schülern zur Unterrichtseinheit, der Unterrichtssituation und der schulischen Erfahrung von Schülern wurden mit diesem Verfahren untersucht.

SUMMARY

The purpose of this book is to promote empirical research in the education of geography by improving some methodological aspects. We analyse data from a large series of experiments in geography curriculum in Germany – the so-called „Raumwissenschaftliches Curriculum Forschungsprojekt (RCFP)", which is comparable to the American High School Geography Project (HSGP).

In particular we are concerned with the following quantitative research techniques: analysis of reliability, exploratory factor-analysis, and techniques of bivariate analysis. Special focus is given to the discussion of multivariate analysis of nommetric data. A relatively new statistical method – the design of a GSK-Model – is demonstrated with data from the RCFP. This model makes it possible to analyse in detail the complex direct and indirect effects of multiple independent variables. Attitudes towards the RCFP-Curriculum, the situation in the classroom, and students experience and motivation are analyzed by GSK-Models.

RÉSUMÉ

Le but du présent ouvrage est de contribuer au développement méthodique de la recherche empirique dans le domaine de la didactique en géographie.

On a analysé des données venant de la série d'essais scolaires en géographie la plus étendue qui ait jamais été effectuée en Europe. Il s'agit du RCFP (Raumwissenschaftliches Curriculum Forschungsprojekt; projet de recherche pour un curriculum en géographie), comparable dens son étendue au HSGP (High School Geography Project) américain.

Au moyen des résultats de l'enquête faite avec le RCFP, on a exposé la capacité de différéntes techniques quantitatives à vérifier la qualité de mesure de certaines variables, à produire des sets de questionnaires unidimensionels et à faire une analyse bivariate des rapports entre variables.

Le point capital du travail consiste à discuter l'aptitude des méthodes multivariates à analyse des donnees non-métriques: on a démontré qu'à l'aide d'une nouvelle méthode statistique – le GSK – il est possible de découvrir entre plusieurs variables qualitatives des rapports qui n'apparaissent pas quand on emploie un procédé bivariate. On a etudie par cette méthode le rapport entre l'attitude des élèves face à l'unitéde cours, la situation de cours et l'experiénce scolaire des élèves.

Anmerkungen

Kap. 1

1 vergl. *Fitt* 1956/; *Slater* 1976
2 Die vorliegende Arbeit kann diesen geschlechtsspezifischen Unterschied im Interesse bestätigen (vergl. dazu die Tab. 30).
3 Auch wir konnten in der vorliegenden Arbeit diesen geschlechtsspezifischen Alterseffekt in bezug auf das Interesse am Fach Erdkunde feststellen: Allerdings tritt dieser Interaktionseffekt unabhängig von der jeweiligen Schulart auf. (vergl. dazu die Kap. 8.4.3 und Kap. 9.4).

Kap. 3

1 Da die Namen der Unterrichtseinheiten in der vorliegenden Arbeit sehr häufig vorkommen, werden einheitliche Abkürzungen benutzt (*Flug, Tabi, Rhein* usw).
2 In der Graphik wird sowohl die Struktur der Datensätze als auch ihr Umfang (Anzahl der Fälle) durch die Größe der „Kästchen" veranschaulicht. Bei der Einheit INDIOS war nur ein sehr kleiner Anteil der Variablen mit den anderen Einheiten identisch (s. kleines gestrichteltes Kästchen). Noch weniger gemeinsame Variablen – mit den anderen Variablen – hatte die Einheit EGBE. Außerdem ist zu erkennen (an den gestrichelten Kästchen), daß insgesamt gesehen bei *allen* Einheiten nur ein (kleiner) Teil der Variablen für einen Gesamtdatensatz verwendbar war.
3 Die Einstellungsbatterie und die Polaritätsprofile werden in 5.1 und 6.1 ausführlich besprochen.
4 Die Variablen V055 bis V062 („Hat dir (Ihnen) die Arbeit am Projekt spaß gemacht oder eher nicht?" – für die einzelnen Teile der Einheit) und die Variablen V063 bis V070 („Glaubst Du (Sie), daß Du (Sie) durch die Mitarbeit an dem Projekt nützliche Dinge gelernt hast (haben)") wurden von uns – wegen der variierenden Anzahl der Projektabschnitte – zu je einem Indexwert SPASS bzw. NUTZEN zusammengefaßt.

Kap. 4

1 Die gründlichste Diskussion dieser Problematik der Perspektivität lieferte der Begründer der „verstehenden Soziologie" *Alfred Schütz (Schütz* 1932 und *Schütz/Luckmann* 1975).
2 Wir mußten in diesem Kapitel leider einige Begriffe (wie z. B. Koeffizient α) verwenden, die erst in den folgenden Kapiteln erklärt werden. Es sei deshalb auf das Kap. 5.1 verwiesen.

Kap. 5

1 Es wurde eine Hauptkomponentenanalyse gerechnet, die von tetrachorischen Korrelationskoeffizienten ausging, um das dichotome Datenniveau zu berücksichtigen.

Kap. 7

1 vgl. dazu: *Noonan/Wold* 1980; *Reynolds* 1977 b; und vor allem *Blalock* 1971.
2 Um korrekt zu sein: Bei *Rhein* und bei *Mobi* wird die Berufsschule nicht ausdrücklich erwähnt. Aber ausgerechnet hier waren Berufsschüler in der Stichprobe.
3 Genauer müßte man eigentlich sagen: „ Die Gruppe, die – nach unserer Definition – dem Faktor INTERESSE mehr als der Durchschnitt aller Schüler zustimmt, ist bei den männlichen Schülern signifikant größer." Wir belassen es im folgenden jedoch bei der obigen, verkürzten Formulierung.
4 Der Ausdruck „korreliert" wird hier im nicht-technischen Sinn von: „es besteht ein Zusammenhang" gebraucht.
5 Allerdings sind diese Ergebnisse mit Vorsicht zu bewerten, da die vollkommen uneinheitliche Kodierung der Variablen SPASS und NUTZEN durch den RCFP-Forschungsstab zu sehr großen Ausfällen durch „Missing Values" führte. Insgesamt konnten nur 3270 Fälle bei der Kreuztabellierung berücksichtigt werden, was gerade bei der Berufsschule zu sehr kleiner Zellenbesetzung führte.

Kap. 8

1 Bei dieser Variablen handelt es sich um eine Einzelfrage aus dem Schülerfragebogen, der *vor* der Erprobung der jeweiligen RCFP-Unterrichtseinheiten ausgefüllt wurde. Die Variablen erfaßt die Erwartungshaltung der Schüler in bezug auf die Thematik der bevorstehenden Unterrichtseinheit.
2 u. V. = unabhängige Variable, a. V. = abhängige Variablen
3 Die Berechnungsweise von Gamma wird ausführlich hergeleitet im Anhang 1
4 Zum Konzept der Auspartialisierung bei Kategorialdaten vergl. *Reynolds* 1977, S. 90 ff.
5 Sämtliche Berechnungen zu dem besprochenen Beispiel finden sich ausführlich dargestellt in Anhang
6 Zur graphischen Darstellung dieses „Interaktions-Effektes" vergl. *Blalock* 1968, S. 224.
7 Es gibt noch andere Möglichkeiten, z. B. die sog. Standardisierung in bezug auf die Kontrollvariable.
8 Dies ist eine dichotome Variable: häufig Gruppenarbeit, selten Gruppenarbeit.
9 Bei diesen Partialtabellen war der χ^2-Test nicht mehr signifikant ($p > 0{,}01$).
10 Die Variable A („Anregung") ist der dichotomisierte „factor-score" des 1. Faktors aus dem Polaritätsprofil zur RCFP-Einheit.
Man wird bemerken, daß in diesem Beispiel die Variable T („Thema wichtig") als *un*abhängige Variable betrachtet wird, während sie im vorangegangenen Beispiel als abhängige Variable behandelt wurde. Man sieht daran, daß allein der (theoretische) Bezugsrahmen bestimmt, welche Funktion eine Varaible hat: In bezug auf die Variable A liegt T zweifellos sowohl zeitlich als auch logisch *vor* A.
11 Selbstverständlich wäre es prinzipiell möglich, auch mehrdimensionale Kontingenztabellen in der „zellenweisen" Kreuztabellenform anzulegen. Dies erweist sich aber dann als sehr unpraktisch, wenn an diese Tabelle ein „Modell" (wie z. B. ein Modell nach dem GSK-Ansatz) angepaßt werden soll.

12 Um eine einfache Unterscheidung zu ermöglichen, sprechen wir von „Verfahren", wenn wir die Methode der Stufe 1 meinen, und von „Modellen", wenn es um die Verfahren der Stufe 2 geht.
13 vergl. *Bishop, Fienberg, Holland* 1975/ *Reynolds* 1977b/ *Gillespie* 1977/ *Habermann* 1978, 1979/ *Goodman* 1970, 1971, 1972, 1978/ *Langenheine* 1979.
14 vergl. *Kritzer* 1979a/ *Grizzle, Starmer, Koch* 1969/ *Küchler* 1976, 1978 a,
15 Zur Terminologie vergl. *Küchler* 1979, S. 154 ff.
16 Ideal wäre es, wenn man diese Auswahl *vor* der Erhebung bereits treffen würde, um dann die entsprechenden Variablen zu erheben. Da wir bereits vorliegende Daten analysieren, bleibt nur übrig, aus den überhaupt vorhandenen Variablen sinnvolle auszuwählen.
17 Es handelt sich um den Datensatz, bei dem die Einstellungsbatterie sowie die Polaritätsprofile bereits in Form der Faktor-Werte (bzw. factor-scores) verdichtet wurden.
18 Die Reihenfolge der 4 Prädiktoren ist gleichgültig (deshalb „Kombinationen ohne Wiederholung"), da zwei Sätze von Prädiktoren mit *gleichen* Prädiktoren aber unterschiedlicher Reihenfolge die gleiche *insgesamte* Varianzaufklärung haben. Bei der schrittweisen Variablenselektion nach *Clark-Higgins-Koch* spielt die Reihenfolge dagegen eine Rolle, da die *Anteile* in der Varianzaufklärung bei den selegierten Variablen je nach Reihenfolge unterschiedlich ausfallen. Auch hier ist aber der *insgesamte* Effekt der Prädiktoren von ihren Reihenfolge unabhängig.
19 Die χ^2-Werte werden durch die Freiheitsgrade dividiert, um eine Normierung über die Größe der Tabelle zu erreichen (vergl. *Iversen*, 1979).
20 Bei der *Clark-Higgins-Koch-Procedure* ist darüber hinaus noch eine weitere „termination-statistic" T vorgesehen. Auf sie wurde wegen der Schwierigkeit der Berechnung verzichtet. Sie liefert jedoch im allgemeinen für die ersten 3-4 Selektionsschritte kein substantiell anderes Ergebnis als T. Erst wenn mehr als 4 Prädiktoren selegiert werden, können sich andere Abbruchskriterien ergeben (vergl. Chi 1979, S. 10).
21 Es wurde bereits darauf hingewiesen, daß in Ermangelung besserer Daten hier „Querschnittdaten" wie „Längsschnittdaten" analysiert wurden.
22 Mit „objektiv" meinen wir, daß der Selektionsalgorithmus standardisiert und damit reproduktionsfähig ist.

Kap. 9

1 Bei dichotomer Kodierung ist es völlig gleichgültig, welche Zahlenwerte man für die Kodierung verwendet.
2 Dieses Beispiel wurde sinngemäß übernommen von *Swafford* 1980, S. 665.
3 Wobei – wie schon erwähnt – der Zahlenwert dieser „Meßwerte" völlig beliebig ist.
4 Damit ist gemeint, daß möglichst ein OLS – („ordinary-least-squares") oder ein WLS („weight-least-squares") – Ansatz verwendet werden sollte. Es gibt auch völlig andere Analyseverfahren für Zusammenhangsanalysen, z. B. den der Informationstheorie (vergl. *Ku/Kullback* 1968), den der Graphentheorie (*Davis* 1975) oder den der „Bayesien Analysis" (*Lindley* 1964).

Einen ganz besonderen Lösungsweg legen *Young, Leeuw* und *Takane* vor: Sie ermöglichen Regressionen mit nichtmetrischen Variablen durch einen Algorithmus wechselweiser Datentransformationen und Modellanpassungen. So erreichen sie eine „optimale Skalierung" der nichtmetrischen Variablen für ein normales OLG-Modell. Die „optimale Skalierung" wird durch ein Verfahren der multidimensionalen Skalierung bewerkstelligt (s. *Young/Leeuw/Takane* 1976).

5 Diese Metrisierung ist nicht ganz perfekt, da eine wirklich metrische Variable einen Definitionsbereich von $\pm \infty$ haben müßte, Berechnet man den Logarithmus der Wahrscheinlichkeiten, erreicht man auch noch, daß dieses Problem verschwindet, denn der Logarithmus ist nicht auf +1, 0 begrenzt. Die log-linearen Modelle basieren auf diesem Vorgehen (s. *Goodmann* 1978b).

6 Hier ist mit „Fallzahl" natürlich die *ursprüngliche* Fallzahl gemeint, aus der die Tabelle ausgezählt wurde.

7 Es ist völlig belanglos, welche Kodierung man wählt, da sich die Regressionskoeffizienten nur durch eine lineare Transformation voneinander unterscheiden. Es ist übrigens richtig, daß sich die Signifikanzen der *Effekte* je nach Kodierung unterscheiden, aber nicht die Signifikanz des Gesamtmodells, da diese sich aus den Residuen ergibt.

8 Diese Tabelle wurde direkt aus dem NONMET-Programm entnommen, um einen Eindruck davon zu vermitteln, wie die Ergebnisse im Ausdruck vorliegen. Im Folgenden werden wir jedoch (statt der exponentiellen) die übliche Zahlendarstellung benutzen und auch die Varianzen der Parameter weglassen.

9 Es ist zunächst etwas ungewohnt, daß ein *signifikanter* χ^2-Wert ein *schlechtes* Modell signalisiert. Aber „signifikant" heißt hier eben „signifikant" große Residuen.

10 Dies soll nur dazu dienen, die Konstruktion der Design-Matrix beim nicht-orthogonalen Fall möglichst einfach zu gestalten. Man könnte im Prinzip auch gleich eine Design-Matrix für den saturierten Fall konstruieren (mittels konditionaler Effekte), nur wäre das sehr kompliziert.

LITERATUR:

Aibauer, R. B. – 1954: Die Lehrerpersönlichkeit in der Vorstellung des Schülers, Regensburg.
Aiken, L. R. – 1970: Attitudes toward mathematics, in: Review of Educational Research, 40, 551–596.
Allerbeck, K. – 1978: Meßniveau und Analyseverfahren – Das Problem strittiger Intervallskalen, in: Zeitschrift für Soziologie, 7, 199–214.
Anwander, G. – 1974: Geschichtliches Interesse und politische Bildung Jugendlicher. Eine psychologisch-soziologische Untersuchung in Münchner Schulen, München.
Armor, D. J. – 1974: Theta Reliability and Factor Scaling. In: Sociological Methodology – 1973–1974, Hg.: *Costner, H. L.*, San Francisco.

Bachmair, G. – 1969: Einstellungen von Schülern zum Lehrer und zum Unterrichtsfach. Dissertation, Erlangen-Nürnberg.
Bauer, L. – 1969: Das geographische Interesse der Gymnasiasten, in: Geographische Rundschau, 21, 106–108.
Bauer, L. – 1968: Erdkunde im Gymnasium. (Wege der Forschung, Band XLVII) Darmstadt.
Baumgärtner, A. M. – 1969: Wie Schüler sich heute ihre Lehrer wünschen, München.
Bedall, F. K. – 1974: Zur Analyse mehrdimensionaler Häufigkeitstabellen. In: Zeitschrift für Sozialpsychologie, 5, (1974), S. 108–114.
Bergan, J. R./Towstopiat, Olga M. u. a. – 1980: A Computer Program for Multidimensional Contingency Table Construction, in: Educational and Psychological Measurement, Vol. 40.
Bergler, R. – 1975: Das Eindrucksdifferenzial – Theorie und Technik, Bern.
Bintig, A. – 1980: The Efficiency of Various Estimations of Reliability of Rating Scales, in: Educational and Psychological Measurement, Vol. 40, 619–643.
Bishop, Y. M. M./Fienberg, S./Holland, P. W. – 1975: Discrete Multivariate Analysis, Cambridge.
Blalock, H. M./Blalock, A. B. (Ed.) – 1968: Methodology in Social Research, New York, u. a..
Blalock, H. M. (Ed.) – 1971: Causal Models in the Social Sciences. Chicago.
Bohrnstedt, G. W./Borgatta, E. F. – 1980: Foreword – Special Issue on Measurement. In: Sociological Methods & Research, Vol. 9, No. 2, (1980), S. 134–146.
Borgatta, E. F./Bohrnstedt, G. W. – 1980: Levels of Measurement. In: Sociological Methods & Research, Vol. 9, No. 2, (1980), S. 147–160.
Bortz, J. – 1979: Lehrbuch der Statistik. Berlin, Heidelberg, New York, 1979 2. Aufl.
Brennan, J. – 1981: Rotation of Two Factor Matrices to a Common Resolution, in: Educational and Psychological Measurement, Vol. 41 S. 237 f.
Busz, M./R. Cohen,/U. Poser, u. a. – 1972: Die soziale Bewertung von 880 Eigenschaftsbegriffen sowie die Analyse der Ähnlichkeitsbezeichnungen zwischen einigen dieser Begriffe, Zeitsch. f. experim. u. angew. Psychologie, 19.

Campbell, D. T./Stanley, J. C. – 1963: Experimental and Quasi-Experimental Design for Research, Chicago, ursprünglich erschienen in: *Gage, N. L.* (Ed.): Handbook of Research on Teaching, Chicago 1966, als deutsche Übersetzung unter dem Namen von Elisabeth Schwarz in: Handbuch der Unterrichtsforschung, Teil 1, Weinheim, 1970. S. 448–631.

Chi, G. Y. – 1979: The Clark-Higgins-Koch Variable Selection Procedure, Mimeo.

Cho, D. W. – 1981: Inter-Rater Reliability – Intraclass Correlation Coefficients, in: Educational and Psychological Measurement, Vol. 41, S. 223 ff.

Christmann, H. – 1978: Beispiel einer Anwendung der Faktorenanalyse Q-Technik im erziehungswissenschaftlichen Bereich, in: Zeitschrift für erziehungswissenschaftliche Forschung, 12, Nr. 4, 221–233.

Christoffersson, A. – 1975: Factor Analysis of Dichotomized Variables. In: Psychometrica, Vol. 40, No. 1, (1975), S. 5–32.

Cicchette, D. V./Heavens, R. – 1981: A Computer Program for Determining the Significance of the Difference between Pairs of Independently Derived Values of KAPPA or Weighted KAPPA. In: Educational and Psychological Measurement, Vol. 41, (1981), S. 189 ff.

Cicourel, A. – 1974: Methode und Messung in der Soziologie. Frankfurt 1974.

Cloß, H./Sperling, W. – 1977: Quantitative Analyse einer Lehrerbefragung über Curricula, Medieneinsatz und berufsbezogene Ausbildung. In: Haubrich, u. a. (Hg.): Quantitative Didaktik der Geographie – Geographiedidaktische Forschung Band 1, Braunschweig, 1977, S. 120–136.

Coleman, J. S. – 1970: The Methods of Sociology. In: *Forcese, D. P./Richer, S.* (Ed.): Stages of Social Research – Contemporary Perspectives, Englewood Cliffs, 1970, S. 399–419.

Costner, H. L/Wagner, W. L. – 1964/1965: The Multivariate Analysis of Dichotomized Variables. In: American Journal of Sociology, Vol. 70, (1964/65), S. 455–466.

Cronbach, L. J. – 1951: Coefficient Alpha and the Internal Structure of Tests. In: Psychometrika, Vol. XVI, (1951), S. 297–334.

Davis, J. A. – 1975: Analyzing Contingency Tables with Linear Flow Graphs – D-Systems. In: Sociological Methodology 1976, *Heise, D. R.* (Ed.), San Francisco, 1975, S. 111–145.

Davis, P. B. – 1977: Conjoint Measurement and the Canonical Analysis of Contingency Tables. In: Sociological Methods & Reserach, Vol. 5, No. 3, (1977), S. 347–365.

Dawes, R. M. – 1977: Grundlagen der Einstellungsmessung. 1. übersetzte Auflage, Weinheim und Basel, 1977.

Dieterich, R. – 1977: Psychodiagnostik – Grundlagen und Probleme. Basel, 1977, 2. Aufl.

Dolansky, R. – 1980: Die Innovationsdiffusion des RCFP bei Erdkundelehrern in Bayern, München, Zulassungsarbeit März 1980.

Dreesmann, H. – 1979: Zusammenhänge zwischen Unterrichtsklima, kognitiven Prozessen bei Schülern und deren Leistungsverhalten, in: Zeitschrift für empirische Pädagogik, 3, 121–133.

Dueck, K. G./Zwirner, W. – 1978: Diskriminanzanalytische Untersuchung zur Raumvorstellung von Kindern, in: Der Erdkundeunterricht, Heft, 28, (Quantitative Didaktik der Geographie Teil II, 5–19.

Duncan, O. D. – 1979: Constrained Parameters in a Model for Categorical Data, in: Sociological Methods & Research, Vol. 8, No. 1, 57–68.

Edwards, A. L./Kenney K. C. – 1946: A Comparison of the Thurstone and Likert Techniques of Attitude Scale Construction, in: Journal of Applied Psychology, 30, 72–83.

Edwards, A. L. – 1975: Techniques of Attitude Scale Construction, New York.

Eigler, G. – 1970: Methoden – Empirische Verfahren in der Erziehungswissenschaft. In: *Speck, J./Wehle G.* (Hg.): Handbuch pädagogischer Grund-Begriffe. München 1970.

Engel, J. – 1978: Dezentrale Curriculumentwicklung – Erfahrungen und Ergebnisse der Entwicklungsphase des RCFP 1974–1976. In: Deutscher Geographentag Mainz, Tagungsberichte und wissenschaftliche Abhandlungen. Wiesbaden, 1978, S. 423–431.

Engelhardt, W. D. – 1973: Zur Entwicklung des kindlichen Raumerfassungsvermögens und der Einführung in das Kartenverständnis. In: *Engelhardt, W. D./Glöckel, H.* (Hg.): Einführung in das Kartenverständnis. Bad Heilbrunn, 1973, S. 103–113.

Engelhardt, M. – 1979: Qualifikation und Selektion in der Schule – Pädagogische Arbeitsorientierung und gesellschaftliches Bewußtsein von Lehrern. In: Zeitschrift für Soziologie, Vol. 8, (1979), S. 111–128.

Federico, P. A./Figliozzi, P. W. – 1981: Computer Simulation of Social Systems. In: Sociological Methods & Research, Vol. 9, No. 4, (1981), S. 513–533.

Fechner, G. H. – 1860: Elemente der Psychophysik, 1860. Auszug in: *Dennis, W.* (Ed.): Readings in the History of Psychology, New York 1948.

Fennessey, J. – 1968: The General Linear Model – A new Perspective on some Familiar Topics. In: American Journal of Sociology, Vol. 74, (1968), S. 1–27.

Fichtinger, R./Geipel, F./Schrettenbrunner, H. – 1969: Studien zu einer Geographie der Wahrnehmung. In: Der Erdkundeunterricht, Heft 19, (1969).

Fienberg, S. – 1977: The Analysis of Cross-Classified Categorical Data, Cambridge, Mass.

Fitt, A. B. – 1956: An Experimental Study of Children's Attitude to School in Auckland, N. Z., in: British Journal of Educational Psychology, Vol. 26, 25–30.

Fittkau, B. – 1969: Dimensionen des Lehrerverhaltens und ihre Bedeutung für die Auslösung von Angst und Sympathie bei Schülern, in: Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie, S. 77–92.

Flechsig, K. H. – 1967: Die Funktion des Experiments in der Unterrichtsforschung, in: Die Deutsche Schule, 59, S. 397–413.

Fricke, R. – 1972: Über Meßmodelle in der Schulleistungsdiagnostik, Düsseldorf.

Fuchs, A. – 1974: Untersuchungen zu metrischen Problemen der Technik der Bedeutungsdifferenzierung. In: Archiv für Psychologie, Bd. 126, (1974), S. 114–124.

Fürstenberg, M./Jungfer, H. – 1979: Evaluation und Revision der RCFP-Unterrichtseinheiten – Bericht über die Erprobungsphase des Raumwissenschaftlichen Curriculum Forschungsprojektes (RCFP) 1976–1978. In: Materialien zu einer neuen Didaktik der Geographie, Bd. 19, München 1979.

Fürstenberg, M./Jungfer, H. – 1980: Evaluation und Revision der RCFP-Unterrichtseinheiten. In: Der Erdkundeunterricht, Heft 34, (1980).

Gaensslen, H./May, F./Wölpert, F. – 1976: Altersgruppenvergleiche von politisch-weltanschaulichen Einstellungen und Persönlichkeitsmerkmalen bei Gymnasiasten. In: Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie, Bd. 8, (1976), S. 24–36.

Geiger, M. – 1977: Die Problematik des Programmierten Erdkundeunterrichts im Vergleich zum Lehrerunterricht – Eine empirische Untersuchung in der Hauptschule, in der Realschule und am Gymnasium. In: Quantitative Didaktik der Geographie, Hg.: *Haubrich*, u. a. Braunschweig, 1977, S. 355–380.

Gillespie, M. W. – 1977: Log-Linear Techniques and the Regression Analysis of Dummy Dependent Variables. Further Bases for Comparison. In: Sociological Methods & Research, Vol. 6, No. 1, (1977), S. 103–122.

Gocka, D. F. – 1973: Stepwise Regression for Mixed Mode Predictor Variables. In: Educational and Psychological Measurement, Vol. 33, (1973), S. 319–325.

Goodman, L. A. – 1970: The Multivariate Analysis of Qualitative Data – Interactions among Multiple Classifications. In: Journal of the American Statistical Association, Vol. 65, (1970), S. 226–256.

Goodman, L. A. – 1971: The Analysis of Multidimensional Contingency Tables. Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications. In: Technometrics, Vol. 13, (1971), S. 33–61.

Goodman, L. A. –1972: A General Model for the Analysis of Surveys. In: American Journal of Sociology, Vol. 77, (1972), S. 1035–1085.

Goodman, L. A. – 1978: Analyzing Qualitative/Categorical Data – Log-linear Models on Latent Structure Analysis. Cambridge, Mass. 1978.

Gorsuch, R. L. – 1980: Factor Score Reliabilities and Domain Validities, in: Educational and Psychological Measurement, Vol. 40, S. 895 ff.

Green, S. B./Lissitz, R. W./Mulaik, S. A. – 1977: Limitations of coefficient alpha as an index of test unidimensionally, in: Educational and Psychological Measurement, 37, S. 827–837.

Greene, V. L./Carmines, E. G. – 1980: Assessing the Reliability of Linear Composites, in: Sociological Methodology, S. 160–190.

Grizzle, J. E./Starmer, C. F./Koch, G. – 1969: Analysis of Categorical Data by Linear Models, in: Biometrics, Vol. 25, 489–504.

Habermann, S. J. – 1978, 1979: Analysis of Qualitative Data, Vol. 1 und 2, New York.

Haldyna, T./Thomas, G. – 1979: The Attitudes of Elementary School Children toward School and Subject Matters, in: Journal of Experimental Education, Vol. 48 No. 1, 18–22.

Hard, G./Wenzel, H. J. – 1979: Wer denkt eigentlich schlecht von der Geographie, in: Geographische Rundschau, 31, 262–267.

Hard, G. – 1977: Zur Inhaltsanalyse fachdidaktischer Texte- Vorbericht über eine Lehrplananalyse, in: *Haubrich, H.* u. a.: Quantitative Didaktrik der Geographie, S. 92–109, Braunschweig.

Harman, H. H. – 1976: Modern Factor Analysis, Chicago, London, 1976, 3. Ausg.

Hartmann, H./Wakenhut, R. – 1972: Zur Dimensionalität gesellschaftlich-politischer Attitüden bei unterschiedlichen Gruppen. In: Zeitschrift für Sozialpsychologie, Bd. 3, (1972), S. 96–115.

Haubrich, H. u. a. – 1977: Konkrete Didaktik der Geographie. Braunschweig, 1977.

Haubrich, H. – 1977 a: Einführungsreferat – Situation und Perspektive geographiedidaktischer Forschung. In: *Haubrich, H.* (Hg.): Quantitative Didaktik der Geographie, Braunschweig, 1977.

Haubrich, H. – 1977 b: Thesen zur Geographiedidaktischen Forschung. In: Geographische Rundschau, Bd. 29, (1977), S. 26–28.

Haubrich, H./Nebel, J. – 1977: Entwicklung eines fachdidaktischen Kategorienprofils zur Analyse von Unterrichtsprozessen in der Geographie – Interaktionsprozeßanalyse mit Hilfe von Videoaufzeichnungen über Aktionsrekorder. In: *Haubrich H.* u. a. (Hg.): Quantitative Didaktik der Geographie. Braunschweig, 1977, S. 225–287.

Havers, N. – 1972: Der Religionsunterricht – Analyse eines unbeliebten Faches. München, 1972.

Heilig, G. – 1981: Evaluationsbericht Hallenbad. In: Materialien zu einer neuen Didaktik der Geographie, Bd. 20, (1981), S. 51–130.

Heilig, G. – 1980: Die einzelnen Schritte einer empirischen Untersuchung. In: Der Erdkundeunterricht, Heft 35, (1980), S. 8–30.

Heilig, G. – 1980: Die Faktorenanalyse als ein vorbereitendes Verfahren zur Bildung homogener Regionen mehrdimensionaler Definition. In: Karlsruher Manuskripte zur Mathematischen und Theoretischen Wirtschafts- und Sozialgeographie, Heft 39, (1980).

Heinrich, H. Ch. – 1974: Verbale Attitüden und tatsächliches Verhalten – Ein Überblick über empirische Befunde zum Problem eines Validitätsaspektes von Einstellungsskalen. In: Zeitschrift für experimentelle und angewandte Psychologie. Bd. XXXI, Heft 1, (1974), S. 62–79.

Heipcke, K. – 1967: Die methodologische Bedeutung des programmierten Lernens für die Unterrichtsforschung. In: Die Deutsche Schule, Bd. 59, (1967), S. 396–413.

Heise, D. R. – 1971: Seperating Reliability and Stability in Test-Retest Correlation. In: *Blalock, H. M.* (Ed.): Causal Models in the Social Sciences, Chicago, 1971, S. 348–363.

Hendrix, L. J./Carter, M. W,/Hintze, J. L. – 1978/79: A Comparison of Five Statistical Methods for Analyzing Pretest-Posttest Designs. In: Journal of Experimental Education, Vo. 47, No. 2, (1978/1979), S. 92–102.

Henrysson, S. – 1973: Methoden der Konstruktion und Analyse von Testaufgaben. In: *Ingenkamp, K./Marsolek, T.* (Hg.): Möglichkeiten und Grenzen der Testanwendung in der Schule. Weinheim, Basel 1973, S. 79–84.

Herbig, M. – 1975: Zur Vortest-Nachtest-Validierung lernzielorientierter Tests. In: Zeitschrift für erziehungswissenschaftliche Forschung, Bd. 9, (1975), S. 112 ff.

Hermann, T./Stapf, A./Krohne, H. W. – 1971: Die Marburger Skalen zur Erfassung des elterlichen Erziehungsstils, Diagnostica, Vol. 17, (1971), S. 118–131.

Higgins, J. E./Koch, G. G. – 1977: Variable Selection and Generalized CHI-Square Analysis of Categorical Data applied to a Large Cross-sectional Occupational Health Survey. In: International Statistical Review, Vol. 45, (1977), S. 51–62.

Hoffmann, P. –1911: Das Interesse der Schüler an den Unterrichtsfächern. In: Zeitschrift für pädagogische Psychologie, 12. Jahrg., (1911), S. 458–470.

Holm, K. – 1974: Theorie der Frage. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie, Bd. 26, Nr. 1, (1974), S. 81–114.

Holm, K. – 1974: Theorie der Fragenbatterie. In: Kölner Zeitschrit für Soziologie und Sozialpsychologie, Bd. 26, Nr. 2, (1974), S. 316–341.

Irwin, H./Baumgart, N. – 1978: Quantitative Analyse der Auswirkungen eines Lernspiels. In: Der Erdkundeunterricht, Heft 28, (1978), S. 29–38.

Iversen, G. R. – 1979: Decomposing CHI-Square – A Forgotten Technique. In: Sociological Methods & Research, Vol. 8, No. 2, (1979), S. 143–157.

Jäger, F. – 1977: Quantitative Methoden zur Analyse von Unterrichtsprozessen auf der Basis audiovisueller Unterrichtsdokumentation. In: *Haubrich, H.* u. a. (Hg.): Quantitative Didaktik der Geographie, Braunschweig, 1977, S. 306–328.

Jiobu, R. M./Lundgren, T. D. – 1978: Catastrophe Theory. A Quasi-Quantitative Methodology. In: Sociological Methods & Research, Vol. 7, No. 1, (1978), S. 29–54.

Kaplitza, G. – 1975: Die Stichprobe. In: *Holm, K.* (Hg.): Die Befragung I, München, 1975.

Kim, J. O. – 1975: Factor Analysis. In: *Nie, N. H./Hull, C. H.* u. a. SPSS, Statistical Package for the Social Sciences, 2. Aufl. New York, 1975.

King, H. A./Karres, L. E. – 1981: INDEX- An Interactive Computer Program for Item Analysis of Objective Tests. In: Educational and Psychological Measurement, Vol. 41, (1981), S. 181 ff.

Kish, L. – 1970: Some Statistical Problems in Research Design. In: *Forcese, D./Richer,* S. (Ed.): Stages of Social Research – Contemporary Perspectives. Englewood Cliffs. 1970. S. 103–116.

Klauer, K. J./Fricke, u. a. – 1972: Lernzielorientierte Tests – Beiträge zur Theorie, Konstruktion und Anwendung. Düsseldorf 1972.

Klauer, K. J. – 1973: Das Experiment in der pädagogischen Forschung. Düsseldorf 1973.

Klauer, K. J. – 1970: Die Bedeutung des Klasseneffektes für die schulpädagogische Forschung. In: Programmiertes Lernen und programmierter Unterricht, Bd. 7, Nr. 1, (1970), S. 149–164.

Klauer, K. J. – 1971: Über Möglichkeiten der Experimentalforschung in Schulklassen. In: Programmiertes Lernen und programmierter Unterricht, Bd. 8, Nr. 1, (1971), S. 1–19.

Kleiter, E. F./Petermann, F. – 1978: Verfahren und Einsatzfelder der gerichteten Voraussetzungs-Cluster-Analyse. In: Zeitschrift für erziehungswissenschaftliche Forschung, Bd. 12, (1978), S. 95–126.

Knoche, W. – 1969: Jungen, Mädchen, Lehrer und Schulen im Zensurenvergleich. Weinheim, 1969.

Koch, G. G./Abernathy, J. R./Imrey, P. B. – 1975: On a Method for Studying Family Size Preferences. In: Demography, Vol. 12, (1975), S. 57–66 (Zu: GSK-Modellen).

Köck, H. – 1977: Zur Problematik von Auswahlantwortaufgaben in lernzielorientierten erdkundlichen Klassenarbeiten – Ergebnisse einer Vorstudie in der Sekundarstufe 1. In: *Haubrich, H.* u. a. (Hg.): Quantitative Didaktik der Geographie, Braunschweig, 1977, S. 339–354.

Kritzer, H. L. – 1977: Analyzing Measures of Association Derived from Contingency Tables. In: Sociological Methods & Research, Vol. 5, (1977), S. 387–418.

Kritzer, H. L. – 1978: An Introduction to Multivariate Contingency Table Analysis. In: American Journal of Political Science, Vol. 22, (1978), S. 187–226.

Kritzer, H. L. – 1979 a: Analyzing Contingency Tables by Weighted Least Squares. An Alternative to the Goodman approach. In: Political Methodology, (1979).

Kritzer, H. L. – 1979 b: Approaches to the Analysis of Complex Contingency Tables – A Guide for the Perplexed. In: Sociological Methods & Research, Vol. 7, (1979).

Kriz, J. – 1981: Methodenkritik empirischer Sozialforschung – Eine Problemanalyse sozialwissenschaftlicher Forschungspraxis. Stuttgart 1981.

Ku, H. H./Kullback, S. – 1968: Interaction in Multidimensional Contingency Tables – An Information Theoretic Approach. In: Journal of Research of the National Bureau of Standarts 72B, (1968), S. 159–199.

Küchler, M. – 1976: Multivariate Analyse nominalskalierter Variablen – Ein theoretischer und empirischer Vergleich unter dem Gesichtspunkt der Forschungspraxis. In: Zeitschrift für Soziologie, Bd. 5, (1976), S. 237–255.

Küchler, M. – 1978: How to use the SPSS regression procedure in the multivariate analysis of dichotomous data. Vortrag auf der ISSUE-Konferenz 1978, Chicago. Erscheint in den Proceedings.

Küchler, M. – 1978 a: Alternativen in der Kreuztabellenanalyse – Ein Vergleich zwischen Goodmans „General Modell" (ECTA) und dem Verfahren gewichteter Regression nach Grizzle et al. (NONMET II). In: Zeitschrift für Soziologie, Bd. 7, (1978), S. 347–365.

Küchler, M. – 1979: Multivariate Analyseverfahren, Stuttgart 1979.

Küchler, M. – 1980: The Analysis of Nonmetric Data. The Relation of Dummy Dependent Variable Regression Using an Additive-Saturated Grizzle-Starmer-Koch Model. In: Sociological Methods & Research, Vol. 8, No. 4, (1980), S. 369–388.

Küppers, W. – 1976: Zur Psychologie des Erdkundeunterrichts. In: Geographische Rundschau, Beiheft 1, (1976), S. 13–19.

Küppers, W. – 1961: Zur Psychologie des Geschichtsunterrichts. Eine Untersuchung über Geschichtswissen und Geschichtsverständnis bei Schülern. Stuttgart, 1961.

Labovitz, S. I. – 1970: Methods for Control with Small Sample Size. In: *Forcese, D./Richer S.* (Ed.): Stages of Social Research – Contemporary Perspectives. Englewood Cliffs, 1970, S. 273–282.

Landis, J. R./Heyman, E. R./Koch, G. G. – 1978: Average Partial Association in Three-Way Contingency Tables – A Review and Discussion of Alternative Tests. In: International Statistical Review, Vol. 46, (1978), S. 237–254.

Lang, S. – 1981: The File – Case Study in Correction 1977–1979, New York, Heidelberg, Berlin 1981.

Langeheine, R. – 1979: Multivariate Analyse nominalskalierter Daten via Goodmans Modell – Sehr wohl eine Alternative. In: Zeitschrift für Soziologie, Jg. 8, Heft 4, (1979), S. 380–390.

Lehnen, R./Koch, G. – 1974: A General Linear Approach to the Analysis of Nonmetric Data. Applications for Political Sciences. In: American Journal of Political Sciences, Vol. 18, (1974), S. 282–313.

Lehnen, R./Koch, G. – 1974: The Analysis of Categorical Data from Repeated Measurement Research Designs. In: Political Methodology, Vol. 1, (1974), S. 103–123.

Leusmann, C. – 1976 a: Die Bestimmung geographisch-inhaltsstruktureller Einstellungsdimensionen von Schülern an Gymnasien. In: Der Erdkundeunterricht, Heft 24, (1976), S. 87–98.

Leusmann, C. – 1976 b: Zur Bewertung des Schulfaches Erdkunde durch Schüler – Teilergebnisse einer Fragebogenaktion an vier Bonner Gymnasien. In: Geographie in Ausbildung und Planung, Nr. 4, S. 5–40, (1976).

Leusmann, C. – 1977: Schülereinstellungen zum Fach Erdkunde, zu Unterrichtsstoffen und zu fachspezifischen Erarbeitungsformen. In: *Haubrich, H.* u. a. (Hg.): Quantitative Didaktik der Geographie, Bd. 1, Braunschweig 1977, S. 145–180.

Leusmann, C. – 1978: Zur Bedingtheit der Einstellungsdimension von Schülern zum Fach Erdkunde. In: Der Erdkundeunterricht, Heft 28, 1978.

Lienert, G. A. – 1969: Testaufbau und Testanalyse. Weinheim, 1969.

Lindley, D. V. – 1964: The Bayesian Analysis of Contigency Tables. In: Annales of Mathematical Statistics, Vol. 35, (1964), S. 1622–1643.

Lobsien, M. – 1909: Beliebtheit und Unbeliebtheit der Schulfächer. In: Langensalza, Pädagogisches Magazin, herausgegeben von Fr. Mann, 361. Heft, (1909).

Lück, H. E./Regelmann, S./Schönbach, P. – 1976: Zur sozialen Erwünschtheit von Eigenschaftsbezeichnungen – Datenvergleiche: Köln 1966 – Bochum 1971 – Köln 1972. In: Zeitschrift für experimentelle und angewandte Psychologie, Bd. XXIII, (1976), Heft 2.

Lukesch, H./Kleiter, G. D. – 1974: Die Anwendung der Faktorenanalyse. Darstellung und Kritik der Praxis einer Methode. In: Archiv für Psychologie, Bd. 126, (1974), S. 265–370.

Lunk, G. – 1924: Münchner Erhebung über das Interesse der Schüler an den Lehrgegenständen. In: Zeitschrift für pädagogische Psychologie, Jahrg. 25, (1924), S. 34 ff.

Marsolek, Th./Ingenkamp, K. H. – 1968: Literatur über Tests im Bereich der Schule – Annotierte Bibliographie der deutschsprachigen Literatur über psychometrische Verfahren. Weinheim, Basel, Berlin 1968.

Miller, D. C. – 1977: Handbook of Research Design and Social Measurement. New York.

Mock, A. – 1965: Entwicklung, Korrelation und Faktorenanalyse der Zeugnisnoten von Oberschülern, Köln 1965.

Morris, J. D. – 1980: The Predictive Accuracy of Full-Rank Variables vs. Various Types of Factor Scores – Implications for Test Validation. In: Educational and Psychological Measurement, Vol. 40, (1980), S. 389–396.

Münzel, H. – 1969: Was sagen Schüler von ihrem Religionsunterricht? Ergebnisse einer Meinungsbefragung unter katholischen Schülern des Saarlandes. In: Religionsunterricht an höheren Schulen, Jg. 12, (1969), S. 8–11.

Namboodiri, K./Carter, L. P./Blalock, H. M. – 1975: Applied Multivariate Analysis and Experimental Designs. New York 1975.

Nie, N. u. a. – 1975: SPSS – Statistical Package for the Social Sciences, New York.

Nolzen, H. – 1977: Entwicklung von Computerprogrammen und Computertests und ihre Auswertung durch EDV. In: *Haubrich, H.* u. a. (Hg.): Quantitative Didaktik der Geographie, Band 1, Braunschweig 1977, S. 56–79.

Noonan, R./Wold, H. – 1980: PLS – Path Modelling with Latent Variables – Analysing School Survey Data Using Partial Least Squares – Part II. In: Scandinavian Journal of Educational Research, Vol. 24, No. 1, (1980), S. 1–23.

Novick, M. R./Lewis, R. – 1967: Coefficient alpha and the reliability of composite measurement. In: Psychometrika, Vol. 32, (1967), S. 1–13.

Orlik, P. – 1965: Eine Modellstudie zur Psychophysik des Polaritätsprofils. In: Zeitschrift für experimentelle und angewandte Psychologie, Vol. 12, (1965), S. 614–647.

Orlik, P. – 1967: Eine Technik zur erwartungstreuen Skalierung psychologischer Merkmalsräume aufgrund von Polaritätsprofilen. In: Zeitschrift für experimentelle und angewandte Psychologie, Vol. 14, (1967), S. 616–650.

Osburn, H. G. – 1968: Item sampling for achivment testing. In: Educational and Psycholigical Measurement, Vol. 28, (1968), S. 95–104.

Pearson, K. – 1901: On lines and planes of closest fit to systems of points in space. In: Philosophical Mag., Ser. 2, 6, (1901), S. 559–572.

Peppler, H. – 1977: Zum Problem inter- und intraindividueller Beurteilungsunterschiede bei der Benotung von Schulleistungen. Eine empirische Untersuchung zum Überprüfungsverfahren zur Aufnahme in die Sonderschule für Lernbehinderte. In: Zeitschrift für erziehungswissenschaftliche Forschung, Bd. 11, Nr. 2, (1977), S. 112–125.

Petermann, F./Knopf, M. – 1976: Probleme bei der Messung von Einstellungsänderungen II – Neuere Entwicklungen und Konzepte. In: Zeitschrift für Sozialpsychologie, Bd. 7, (1976), S. 217–230.

Popper, Karl R./Eccles, John C. – 1977: The Self and Its Brain. Berlin u. a. 1977.

Prawdzik, W. – 1971: Untersuchung zur Einstellung der 9. Klasse Hauptschule in München zum katholischen Religionsunterricht. Dissertation, München 1971.

RCFP, – 1974: RCFP-Lenkungsausschuß (Hg.): Zielsetzung und Schwerpunkte des RCFP. Materialien zu einer neuen Didaktik der Geographie, Bd. 1, (1974), München.

RCFP, – 1976: RCFP-Lenkungsausschuß (Hg.): Probleme und Verfahren der Curriculumentwicklung im RCFP. Arbeitsbericht über die Jahrestagung des Raumwissenschaftlichen Curriculum Forschungsprojektes vom 1. bis 5. März 1976 in Goslar. Materialien zu einer neuen Didaktik der Geographie, Bd. 3, München 1976.

RCFP, – 1978: RCFP-Lenkungsausschuß (Hg.): Das Raumwissenschaftliche Curriculum Forschungsprojekt. Erfahrungen und Ergebnisse der Entwicklungsphase 1973–1976. Braunschweig 1978.

Reichert, R. – 1981: Fallstudie zur Diffusion von Neuerungen. Zulassungsarbeit für das Lehramt, München 1981.

Remplein, H. – 1968: Die seelische Entwicklung des Zehn- bis Zwanzigjährigen und sein Verhältnis zur Schule. In: *Bauer, L.* (Hg.): Erdkunde am Gymasium. Darmstadt 1968, S. 309–334.

Revenstorf, D. – 1973: Über Profilähnlichkeit. In: Archiv für Psychologie, Bd. 125, (1973), S. 203–232.

Revenstorf, D. – 1976: Lehrbuch der Faktorenanalyse, Stuttgart 1976.

Revenstorf, D. – 1978: Vom unsinnigen Aufwand. In: Archiv für Psychologie, Bd. 130, (1978), S. 1–36.

Reynolds, H. T. – 1977 a, The Analysis of Cross-Classifications. New York 1977.

Reynolds, H. T. – 1977 b: Some Comments on the Causal Analysis of Surveys with Log-Linear Models. In: American Journal of Sociology, Vol. 83, (1977), S. 127–143.

Rhenius, D. – 1974: Bemerkungen über Verfahren zur Zielrotation von Ladungsmatrizen. In: Archiv für Psychologie, Bd. 126, (1974), S. 125–130.

Rollet, B. – 1969: Das Design in der empirischen Unterrichtsforschung. In: *Roth, L.* (Hg.): Beiträge zur empirischen Unterrichtsforschung, Hannover 1969.

Rollet, B. – 1978: Lernpsychologische Untersuchungen als Grundlage geographiedidaktischer Planung. In: Der Erdkundeunterricht, Heft 28, (1978), S. 39–55.

Rosenberg, M. – 1962: Test Factor Standardization as a Method of Interpretation. In: Social Forces, Vol. 41, (1962), S. 53–61.

Rosenshine, B. – 1970: The Stability of Teacher Effects upon Student Achivement. In: Review of Educational Research, Vol. 40, (1970), S. 647–662.

Rost, J. – 1977: Diagnostik des Lernzuwachses – Ein Beitrag zur Methodik von Lerntests. IPN-Arbeitsbericht 26, (1977).

Rütter, T. – 1971: Das didaktische Experiment. In: *Dohmen, G.* (Hg.): Forschungstechniken für die Hochschuldidaktik, München 1971.

Ruprecht, H. – 1974: Einführung in die empirische pädagogische Forschung. Bad Heilbrunn, 1974.

Schanz, G. – 1973: Tests im Erdkundeunterricht. In: Der Erdkundeunterricht, Heft 18, (1973).

Schedl, G. – 1981: Evaluationsbericht: „Unsere landesvätterliche Sorgfalt". In: Materialien zu einer neuen Didaktik der Geographie, Bd. 20, (1981), S. 131–172.

Schrettenbrunner, H. – 1969: Schülerbefragung zum Erdkundeunterricht. In: Geographische Rundschau, Bd. 21, (1969), S. 100–106.

Schrettenbrunner, H. – 1976: Zielsetzung für eine quantitative Didaktik der Geographie. In: Der Erdkundeunterricht, Heft 25, (1976), S. 3–11.

Schrettenbrunner, H. – 1976 a: Quantitative Didaktik der Geographie – Teil 1. In: Der Erdkundeunterricht, Heft 24, (1976).

Schrettenbrunner, H. – 1976 b: Die graphentheoretische Darstellung von Unterrichtsprogrammen. In: Der Erdkundeunterricht, Heft 24, (1976), S. 28–45.

Schrettenbrunner, H. – 1977: Methodische Anmerkungen zu Lernweganalysen bei verzweigten Unterrichtsprogrammen. In: *Haubrich H.* u. a. (Hg.): Quantitative Didaktik der Geographie, Braunschweig 1977, S. 41–47.

Schrettenbrunner, H. – 1978: Konstruktion und Ergebnisse eines Tests zum Kartenlesen – Kartentest KAT. In: Der Erdkundeunterricht, Heft 28, (1978), S. 56–75.

Schrettenbrunner, H. – 1980: Quantitative Didaktik der Geographie – Teil II. In: Der Erdkundeunterricht, Heft 28, (1980).

Schrettenbrunner, H. – 1980a: Untersuchungsplan zum Messen von Schülerreaktionen. In: *Jäger, F.* u. a. (Hg.): Prozeßanalysen geographischen Unterrichts. Geographiedidaktische Forschung, Bd. 6, (1980), S. 82–94.

Schrettenbrunner, H. – 1981: Evaluationsbericht Siedlungsspiel. In: Materialien zu einer neuen Didaktik der Geographie, Bd. 20, (1981), S. 7–34.

Schütz, A. – 1932: Der sinnhafte Aufbau der sozialen Welt. Wien 1932, Neuauflage: Frankfurt 1974.

Schütz, A./Luckmann, T. – 1975: Strukturen der Lebenswelt. (Aus dem Nachlaß), Neuwied, Darmstadt 1975.

Schumacher, J. – 1974: Befragung – Schülermeinungen zur „Dritten Welt". In: Geographie in Ausbildung und Planung, Nr. 3, (1974), S. 6–72.

Schwarzer, R. – 1979: Bezugsgruppeneffekte in schulischen Umwelten. In: Zeitschrift für empirische Pädagogik, Bd. 3, (1979), S. 153–166.

Seelig, G. – 1968: Beliebtheit von Schulfächern. Empirische Untersuchungen über psychologische Schulfachbevorzugungen. In: Theorie und Praxis der Schulpsychologie, Bd. 12, (1968).

Shepard, R./Romney, K./Nerlove, S. – 1972: Multidimensional Scaling – Theory and Applications in the Behavioral Sciences. Vol. 1 und 2.

Silverstein, A. B. – 1980: Item Intercorrelations, Item-Test Correlations, and Test Reliability. In: Educational and Psychological Measurement, Vol. 40, (1980), S. 353 ff.

Slater, F. – 1976: Student Perception of Teacher Style in Geography – An Exploratory Study in London. In: Stoltman, J.P. (Ed.): International Research in Geographical Education, Western Michigan University, 1976, S. 115 ff.

Smith, J. K. – 1980: On the Examination of Test Unidimensionality. In: Educational and Psychological Measurement, Vo. 40, (1980), S. 885 ff.

Spearman, C. – 1904: General Intelligence, objectively determined and measured. In: American Journal of Psychology, Vol. 15, (1904), S. 201–231.

Specht, D. A. – o. J.: SPSS – Users Guide to Subprogram Reliability and Repeated Measurement Analysis of Variance. Department of Sociology, Iowa State University, o. J.

Steinbrink, J. – 1976: Researching Instructional Style and Classroom Environments – A Survey of Techniques. In: *Stoltman, J. P.* (Ed.): International Research in Geographical Education. Research Reports Prepared in Conjunction with the 23rd Congress of the International Geographical Union. Western Michigan University, 1976, S. 89 ff.

Stern, W. – 1905: Über die Beliebtheit und Unbeliebtheit der Schulfächer. In: Zeitschrift für pädagogische Psychologie, 7. Jahrg., (1905), S. 267 ff.

Stock, W. A./Elliott, S. D. – 1980: A Coefficient Alpha Program Allowing Stepwise Item Deletion. In: Educational and Psychological Measurement, Vol. 40, (1980).

Summers, – 1971: Attitude Scaling.

Swafford, M. – 1980: Three Parametric Techniques for Contingency Table Analysis – A Nontechnical Commentary. In: American Sociological Review, Vol. 45, (1980), S. 664–690.

Tarnai, C. – 1978: Anwendung von Methoden der Transformationsanalyse als Form der konfirmatorischen Faktorenanalyse. In: Zeitschrift für Soziologie, Bd. 7, (1978), S. 366–379.

Tatsuoka, M. M./Tiedemann, D. V. – 1973: Statistics as an Aspect of Scientific Method in Research Design. In deutsch erschienen unter: *Holzkamp, C.:* Statistik als ein Aspekt der wissenschaftlichen Methoden in der Unterrichtsforschung, Teil 1, Hg.: *Ingenkamp, K.*, Weinheim, Basel 1973, 3. Aufl.

Thompson, B./Miller, A. H. – 1981: The Utility of Social Attitudes Theory. In: Journal of Experimental Education, Vol. 49, No. 3, (1981), S. 157–160.

Thompson, B./Stapleton, J. C. – 1979/1980: A Method for Validating Semantic Differential Referents. In: Journal of Experimental Education, Vol. 48, No. 2, (1979/80), S. 110–113.

Waddington, C. H. – 1977: Tools for Thought. Frogmore, St. Albans 1977.
Watzka, W. – 1977: Einflüsse von Unterrichtsformen auf Motivation und Lernerfolg. In: *Haubrich, H.* u. a. (Hg.): Quantitative Didaktik der Geographie. Braunschweig, 1977, S. 381–403.
Weber, B. – 1974: Schulbuchanalyse – „Dritte Welt". In: Geographie in Ausbildung und Planung, Nr. 3, (1974), S. 73–116.
Weisberg, H. F. – 1974: Dimensionland – An Excursion into Spaces. In: American Journal of Political Sciences, Vol. 18, (1974), S. 743–776.
Wendeler, J. – 1967: Konstruktion eines Schulinteresseninventars. In: Mitteilungen und Nachrichten des Deutschen Instituts für Internationale Pädagogische Forschung, Nr. 46/47, S. 35–45.
Wiederkehr, G. – 1907/1908: Statistische Untersuchungen über die Art und den Grad des Interesses bei Kindern in der Volksschule. In: Neue Bahnen, Zeitschrift für Erziehung und Unterricht, (1907/08), S. 251 ff.
Wolf, G./Cartwright, B. – 1974: Rules for Coding Dummy Variables in Multiple Regression. In: Psychological Bulletin, Vol. 81, (1974), S. 173–179.

Young, F. W./Leeuw, J./Takane, Y. – 1976: Regression with Qualitative and Quantitative Variables – An Alternating Least Squares Method with Optimal Scaling Features. In: Psychometrika, Vol. 41, No. 4, (1976), S. 505–529.

Zwirner, W. – 1978: Trennschärfenberechnung und Diskriminanzanalyse. In: Der Erdkundeunterricht, Heft 28, (1978), S. 20–28.

Anhang 1: Die Rechenschritte zur Berechnung von . . .

(1) Gamma,
(2) „konditionalem" Gamma für Partialtabellen.
(3) „partiellem" Gamma,
(4) Gamma der Gesamttabelle (mit **Supervariablen)**

(1) **Berechnung des Koeffizienten Gamma** für das Beispiel in 8.1.1, S. 128:

Formel: $\text{Gamma}_{KT} = \frac{P-Q}{P+Q}$; P: konkordantes Paar
Q: diskordantes Paar

Tabelle:

986 (60,2 %)	1681 (67,6 %)	1157 (77,7 %)
648 (39,8 %)	806 (32,4 %)	333 (22,3 %)

Rechengang: $P = 986 \cdot (806+333) + 1681 \cdot 333 = 1\,682\,827$
$Q = 1157 \cdot (806+648) + 1681 \cdot 648 = 2\,771\,566$

$P - Q = -1\,088\,739$
$P + Q = 4\,454\,393$
$\frac{P-Q}{P+Q} = -0{,}2444192$

Ergebnis: $\text{Gamma}_{KT} = -0{,}25$

(2) **Berechnung der „konditionalen" Gamma-Koeffizienten** für die Partialtabellen in 8.1.2, S. 130

Formel: $\text{Gamma}_{KT.V1} = \frac{P_1-Q_1}{P_1+Q_1}$; $\text{Gamma}_{KT.V2} = \frac{P_2-Q_2}{P_2+Q_2}$;

dabei ist $\text{Gamma}_{KT.V1}$: Gamma-Koeffizient bei Tabelle V_1 (d. h. **mit** Vortragserfahrung)

$\text{Gamma}_{KT.V2}$: Gamma-Koeffizient bei Tabelle V_2 (d. h. **ohne** Vortragserfahrung)

Partialtabellen:

375 (68,8 %)	1067 (72,0 %)	939 (78,1 %)	611 (56,1 %)	614 (61,1 %)	218 (76,0 %)
170 (31,2 %)	415 (28,0 %)	264 (21,9 %)	478 (43,9 %)	391 (38,9 %)	69 (24,0 %)

Rechengang: $P_1 = 375 \cdot (415 + 264) + 1067 \cdot 264 = 536313$

$Q_1 = 939 \cdot (415 + 170) + 1067 \cdot 170 = 730705$

$P_1 - Q_1 = -194392$
$P_1 + Q_1 = \cdot\ 1267018$

$$\frac{P_1 - Q_1}{P_1 + Q_1} = -0{,}153424$$

$P_2 = 611 \cdot (391 + 69) + 614 \cdot 69 = 323426$

$Q_2 = 218 \cdot (391 + 478) + 614 \cdot 478 = 482934$

$P_2 - Q_2 = -159508$
$P_2 + Q_2 = 806360$

$$\frac{P_2 - Q_2}{P_2 + Q_2} = -0{,}1978124$$

Ergebnis: $\text{Gamma}_{KT.V1} = -0{,}15$

$\text{Gamma}_{KT.V2} = -0{,}20$

(3) **Berechnung der „partialen" Gamma-Koeffizienten** bei Auspartialisierung der Variablen „Vortragserfahrung" aus dem Beispiel in: 8.1.2, S. 131

Formel: $$\text{Gamma}_{KT.V} = \frac{\sum_i (P_1 - Q_1)}{\sum_i (P_i + Q_i)} \qquad \frac{(P_1 - Q_1) + (P_2 - Q_2)}{(P_1 + Q_1) + (P_2 + Q_2)}$$

Rechengang: siehe unter (2).

$$\text{Gamma}_{KT.V} = \frac{(-194392) + (-159508)}{1267018 + 806360} = -0{,}170688$$

Ergebnis: $\text{Gamma}_{KT.V} = -0{,}17$

Meinungen und Vorstellungen der Schüler zum Schulfach Erdkunde prägen die emotionale Ausgangssituation für den Unterricht und sind wichtige Bausteine im Schüler - Lehrerverhältnis. Inhaltliches Anliegen der Arbeit ist es, solche Einstellungen mit empirischen Methoden zu erfassen, zu beschreiben und die nach Alter, Klassenstufe und Schulart unterschiedlichen Ergebnisse zu erklären.

Eine weitere Zielsetzung des Buches liegt im methodischen Bereich. Es soll aufgezeigt werden, wie sich mit Hilfe multivariater statistischer Verfahren geographiedidaktische Erhebungen verbessern lassen. Die methodischen Fragestellungen der Arbeit werden anhand einer Analyse RCFP-Erhebung (Raumwissenschaftliches Curriculum Forschungsprojekt) untersucht. Es handelt sich dabei um die erste Analyse des zusammengefaßten Datensatzes aus 7 RCFP-Einzelprojekten.

ISBN 3-496-00728-1