Forschungsmethoden der Psychologie

Hochschullehrbücher

Band 3

Wilhelm Kempf

Forschungsmethoden der Psychologie

Zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik

Band 2: Quantität und Qualität

verlag irena regener berlin

Das vorliegende Buch entstand im Rahmen des Projektes "Modulare Methodenausbildung im Fach Psychologie" und wurde aus Mitteln des Ausschusses für Forschungsfragen an der Universität Konstanz (AFF) gefördert.

1. Auflage
© 2008 by verlag irena regener
ostseestr.109, d – 10409 berlin
http://www.regener-online.de

Alle Rechte vorbehalten, auch die der fotomechanischen Wiedergabe und der Speicherung in elektronischen Medien.

Printed in Germany

ISBN 978-3-936014-15-0

Inhalt

Vorwort 9

Kapitel 1:
Textinterpretation und Inhaltsanalyse

1.1 Qualitative Sozialforschung 15

1.1.1 Grundlagen qualitativer Forschung 15
1.1.2 Qualitative Erhebungsinstrumente 19
1.1.3 Transkriptionsverfahren 27
1.1.4 Qualitative Analysemethoden 31

1.2 Grundlagen der sozialpsychologischen Rekonstruktion 37

1.3 Textinterpretation 44

1.3.1 Varianten der Textinterpretation 44
1.3.2 Exemplifikation der Vorgehensweise 46
1.3.3 Gegenstandsbedingte Differenzierungen 68

1.4 Inhaltsanalyse 74

1.4.1 Varianten der Inhaltsanalyse 74
1.4.2 Qualitative Inhaltsanalyse 80
1.4.3 Materialbezogene Kategorienbildung 86
1.4.4 Theoriebezogene Kategorienbildung 101
1.4.5 Quantitative Inhaltsanalyse 112

1.5 Integration quantitativer und qualitativer Analysemethoden 123

Kapitel 2:
Psychometrie

2.1 Grundlagen 139

2.1.1 Die Anfänge der Testpsychologie 139

2.1.2 Die Fragestellungen der Testtheorie 141
2.1.3 Die Beurteilung von Testleistungen 143

2.2 Testen und Messen 149

2.2.1 Grundannahmen der klassischen Testtheorie 149
2.2.2 Reliabilität und Validität 152
2.2.3 Stochastische Testmodelle 176
2.2.4 Das Konzept der lokalen stochastischen Unabhängigkeit 190

2.3 Klassische und stochastische Testtheorie 199

2.3.1 Ein allgemeines testtheoretisches Modell 199
2.3.2 Rechtfertigung der Scorebildung 204
2.3.3 Notwendige und hinreichende Bedingungen für die Suffizienz des Summenscores 218

2.4 Parameterschätzung, Modellkontrolle und Beurteilung der Testleistung der Probanden in der klassischen Testtheorie 234

2.4.1 Parameterschätzung 234
2.4.2 Überprüfung der Modellannahmen 234
2.4.3 Beurteilung der Testleistungen 238

2.5 Parameterschätzung und Modellkontrolle im saturierten Modell 243

2.5.1 Parameterschätzung 243
2.5.2 Modellkontrolle 244

2.6 Parameterschätzung und Modellkontrolle in Modellen mit personen- und/oder itemunabhängigen Antwortvariablen 247

2.6.1 Parameterschätzung 247
2.6.2 Überprüfung der Modellannahmen 253

2.7. Parameterschätzung in Latent-Trait-Modellen 258

2.7.1 Die unbedingte Maximum-Likelihood-Methode 258
2.7.2 Die bedingte Maximum-Likelihood-Methode 261
2.7.3 Die marginale Maximum-Likelihood-Methode 265
2.7.4 Die gewichtete Maximum-Likelihood-Methode 268

2.8. Modellkontrolle und Beurteilung der Testleistung der Probanden im Rasch-Modell 274

2.8.1 Eigenschaften des Rasch-Modells 274

2.8.2 Kontrolle der spezifischen Objektivität 276
2.8.3 Kontrolle der lokalen Unabhängigkeit 282
2.8.4 Beurteilung der Testleistungen 285

2.9. Verallgemeinerungen des Rasch-Modells 290

2.9.1 Mehrfaktorielle Rasch-Modelle 290
2.9.2 Das lineare logistische Testmodell 296
2.9.3 Versuchsplanung. 300
2.9.4 Hypothesentestung 304
2.9.5 Degenerierte Modellvarianten. 305
2.9.6 Grenzfälle des Rasch-Modells 310
2.9.7 Mischverteilungs-Rasch-Modelle 312

2.10 Identifikation der Klassenanzahl, Parameterschätzung und Beurteilung der Testleistung der Probanden in der Latent-Class-Analyse. . . 317

2.10.1 Beurteilung der Testleistung. 317
2.10.2 Parameterschätzung 318
2.10.3 Identifikation der Klassenanzahl 321
2.10.4 Modellvoraussetzungen 324
2.10.5 Weiterverrechnung der LCA-Ergebnisse. 327

2.11 Modellvergleiche 330

2.12 Validität 334

2.12.1 Augenscheinvalidität, theoretische Validität und Kriteriumsvalidität . 334
2.12.2 Konstruktvalidität 338

Literatur 350

Anhang

Über den Autor 364

Abkürzungen und Notationen 365

Das griechische Alphabet. 369

Terminologische Tabellen 370
Bewertung des Koeffizienten κ 370
Beschreibung von LCA-Ergebnissen 370

Statistische Tabellen 371
Binomialkoeffizienten 371
Binomialverteilung 373
Standardnormalverteilung. 374

t-Verteilung 374
χ^2-Quadrat Verteilung 376
Aufgabenlösungen 377
Sachregister 389

Formelsammlung

Kostenloser Download unter http:\\www.regener-online.de,
Eingangsbereich Bücher

Vorwort

Quantität und Qualität. Gegenstand des zweiten Bandes der Forschungsmethoden der Psychologie sind zwei Teilgebiete der Methodenlehre, die auf geradezu entgegengesetzten Polen des Methodenspektrums angesiedelt sind: *Textinterpretation und Inhaltsanalyse*, die dem Wissensideal der qualitativen Sozialforschung verpflichtet sind und auf die Rekonstruktion der subjektiven Welt des Menschen abzielen, welche sie als eine sinn- und bedeutungsstrukturierte Wirklichkeit begreifen, und *Psychometrie*, die von manifesten Beobachtungsdaten auf zugrunde liegende Eigenschaften der Probanden (z. B. Intelligenz, Depressivität oder Einstellungen) schließt und die wohl konsequenteste Umsetzung eines quantifizierenden Wissensideals in der Psychologie darstellt. Entgegen der ursprünglichen Planung werden die experimentalpsychologischen Methoden nicht in diesem Band der Forschungsmethoden vertieft, sondern in dem gemeinsam mit Markus Kiefer herausgegebenen dritten Band des Lehrbuchwerkes, der im Sommer 2009 vorliegen wird.

In ihren Grundzügen wurden die beiden Themen des Buches bereits im ersten Band vorgestellt. Hier sollen sie nun eine vertiefte Behandlung erfahren, die den Leser in die Lage versetzt, die einschlägige Fachliteratur kritisch zu rezipieren, die Methoden in der eigenen Forschungspraxis anzuwenden und sie in gegenstandsangemessener Weise miteinander zu verbinden.

Aus dem Bedürfnis nach Qualitätssicherung psychologischer Testverfahren heraus entwickelt, stellen die im zweiten Kapitel dargestellten psychometrischen Modelle und Methoden eine allgemeine (statistische) Theorie psychologischer Daten dar, deren forschungspraktische Bedeutung weit über die Testpsychologie hinausreicht und die auch für die experimentelle Psychologie und die Surveyforschung relevant ist.

Abgesehen von verhaltenswissenschaftlichen oder neurowissenschaftlichen Experimenten, in denen die abhängige Variable durch

physikalische Messungen erfasst werden kann – z. B. durch Reaktionszeitmessung oder die Registrierung von Augenbewegungen, durch EEG oder EMG etc. –, haben wir es in der Psychologie durchgehend mit qualitativen (d. h. mit nominalskalierten) Beobachtungsdaten zu tun. Und so gut wie überall müssen wir mit interindividuellen Unterschieden zwischen den Probanden rechnen, welche die empirischen Regelmäßigkeiten überlagern, die wir untersuchen. Die Grundfragen der Psychometrie – ob und ggf. wie von manifesten qualitativen Variablen auf latente Variablen geschlossen werden kann, welche die interindividuellen Unterschiede modellieren – sind in der psychologischen Forschung daher allgegenwärtig und finden nicht erst neuerdings auch in der Fragebogenkonstruktion gesteigerte Aufmerksamkeit. Im Gegenteil, die Probleme der Fragebogenkonstruktion waren bereits für die Pionierarbeiten konstitutiv, mit welchen Guttman (1950) und Lazarsfeld (1950) den Grundstein für die moderne Psychometrie legten.

Tatsächlich ist die Fragebogenmethode aus der experimentellen Psychologie nicht wegzudenken und in der Surveyforschung ist sie ohnedies die Methode schlechthin. Während die aus dem Behaviorismus hervorgegangene verhaltenswissenschaftliche Tradition der Psychologie sprachliche Äußerungen als "verbale Daten" thematisiert und die mit der Fragebogenkonstruktion einhergehenden Methodenprobleme – über die hier behandelten psychometrischen Fragen hinaus – hauptsächlich in Problemen der Itemformulierung, der Konstruktion der Antwortformate und den möglichen Fehlerquellen bei der Itembeantwortung sieht (z. B. Bühner, 2006; Jonkisz & Moosbrugger, 2007), pflegt die qualitative Sozialforschung (z. B. Flick et al., 2000) einen völligen anderen Umgang damit.

Obwohl die Anwendung quantifizierender und/oder statistischer Methoden auch von der qualitativen Sozialforschung nicht grundsätzlich abgelehnt wird und der einstige Antagonismus von quantitativen oder qualitativen Methoden inzwischen dem Prinzip der Methoden-Triangulation gewichen ist, sind Messung und Quantifizierung für den qualitativen Sozialforscher allenfalls von nachrangigem Stellenwert. Worum es in erster Linie geht, ist die methodisch fundierte Rekonstruktion von Bedeutungen, wie wir sie im ersten Kapitel dieses Buches am Beispiel der sozialpsychologischen Rekonstruktion ausarbeiten, die ein Spektrum von textanalytischen Methoden umfasst, das von der Textinterpretation über

die qualitative Inhaltsanalyse bis hin zur quantitativen Inhaltsanalyse reicht und über den Einsatz psychometrischer Verfahren (namentlich der Latent-Class-Analyse) eine Brücke zwischen den beiden Themen des Buches schlägt.

In didaktischer Hinsicht ist das Buch dem Prinzip des Lernens am konkreten Objekt verpflichtet, das in beiden Buchkapiteln allerdings eine je sehr unterschiedliche Ausformung erfährt.

Textanalytische Methoden sträuben sich geradezu gegen eine strikte Schematisierung. Wirklich verstanden hat man sie erst, wenn man die konkrete Erfahrung gemacht hat, wie die zugrunde liegenden methodologischen Prinzipien praktisch umgesetzt werden können, wofür wir auf Beispiele aus eigenen Forschungsprojekten der Projektgruppe Friedensforschung Konstanz zurückgreifen.

Ganz anders verhält es sich dagegen mit der Psychometrie, die man erst dann verstanden hat, wenn man die zugrunde liegende mathematische Logik begriffen und die mathematischen Beweisführungen (zumindest im Prinzip) nachvollzogen hat.

Da es inzwischen eine ganz Reihe von Lehrbüchern gibt, welche die psychometrischen Methoden anwendungsnah darstellen (z. B. Rost, 1996; Moosbrugger & Kelava, 2007), legen wir den Schwerpunkt daher auf deren mathematische Grundlagen und diskutieren die klassische und die probabilistische Testtheorie im Rahmen eines allgemeinen testtheoretischen Modells, welches die beiden Schulrichtungen integriert und auf eine gemeinsame konzeptionelle Basis stellt.

Die Beweise sind vom Text optisch abgesetzt, damit der Leser die Chance hat, sie beim ersten Lesen des Kapitels zu überspringen und sie sich in einem zweiten Durchgang zu erarbeiten, nachdem er die Stoßrichtung der Argumentation bereits verstanden hat.

Damit die Beweisführung für den Leser leichter nachvollziehbar wird und ihm ein lästiges Hin-und-Herblättern zwischen den verschiedenen Unterkapiteln erspart bleibt, wurde eine Formelsammlung erstellt, die sämtliche im Buch mit Nummern versehenen Formeln enthält und die unter http://www.regener-online.de, Eingangsbereich Bücher, zum kostenlosen Download bereitsteht. Um seine mathematischen Vorkenntnisse nicht zu überfordern, bedienen wir uns nicht immer der elegantesten, sondern der einfachsten und anschaulichsten Beweisführung und stellen diese

auch nicht in ihrer allgemeinsten Form, sondern anhand der einfachsten Modellvarianten dar. Dies bedeutet z. B. für die klassische Testtheorie, dass wir die Methoden der Testverlängerung nicht für den allgemeinen Fall der n-fachen Testlänge, sondern nur für die 2-fache Testlänge herleiten; für die Latent-Trait-Modelle, dass wir die Beweisführung nicht für polytome, sondern (nur) für dichotome Itemformate darstellen. Hat der Leser diese Beweise erst einmal nachvollzogen, so sollte es ihm ein Leichtes sein, sich anhand der angegebenen Literatur auch den allgemeineren Fall zu erarbeiten.

Zur Einübung der Methoden und Argumentationsweisen gibt es neben den Kontrollfragen auch Übungsaufgaben am Ende der entsprechenden Unterkapitel. Die Aufgabenlösungen sind im Anhang dokumentiert.

Kapitel 1

Textinterpretation und Inhaltsanalyse

1.1
Qualitative Sozialforschung

1.1.1 Grundlagen qualitativer Forschung

Ursprünge der qualitativen Sozialforschung

Lange Zeit hindurch als "unwissenschaftlich", "feuilletonistisch" oder "unseriös" abqualifiziert, hat die qualitative Sozialforschung seit den 60er Jahren des vergangenen Jahrhunderts in den USA und seit den 70er Jahren auch im deutschsprachigen Raum eine Renaissance erlebt (vgl. Flick, 1995) und findet seit den 80er Jahren auch in der Psychologie eine zunehmende Verbreitung. Ihre Entwicklung war jedoch seit jeher dadurch geprägt, dass sie sich in unterschiedlichen Disziplinen und Forschungszusammenhängen entwickelt hat, die jeweils durch einen eigenen Theoriehintergrund und ein eigenständiges methodisches Programm gekennzeichnet waren. Entsprechend ist qualitative Forschung ein Sammelname für z. T. sehr unterschiedliche Forschungsprogramme, denen allen jedoch der Anspruch gemeinsam ist, Lebenswelten gleichsam von innen heraus, aus der Sicht der Handelnden zu beschreiben und zu erklären und auf diesem Wege zu einem besseren Verständnis der subjektiven Welt des Menschen und der sozialen und kulturellen Regeln zu gelangen, aufgrund derer sie sich konstituiert (vgl. Flick et al., 2000).

In der Tradition des symbolischen Interaktionismus (vgl. Denzin, 2000), der Phänomenologie (vgl. Hitzler & Eberle, 2000), der Ethnomethodologie (vgl. Bergmann, 2000b) und des Konstruktivismus (vgl. Flick, 2000a; siehe auch Bd. 3, Kap. 2) stehend, zeichnet sich die qualitative Forschung nicht – oder zumindest nicht primär – durch die Ablehnung quantitativer Methoden, sondern durch ein gemeinsam geteiltes Gegenstandsverständnis aus, das die soziale Wirklichkeit als das Produkt von Bedeutungszuweisungen begreift, welche ihren Ursprung in der sozialen Interaktion haben und die Grundlage für das Handeln und die Handlungsentwürfe der Menschen bilden.

Gegenüberstellung von quantitativen und qualitativen Methoden

Daraus ergibt sich das Aufgabenverständnis der qualitativen Forschung, welches sich die Rekonstruktion der subjektiven Sichtweisen und Deutungsmuster der sozialen Akteure und deren Entstehung zum Ziel setzt. Erst relativ zu diesem Aufgabenverständnis können die jeweiligen Forschungsmethoden dann begründet werden, denn letztlich sind es die Methoden, die darüber bestimmen, welche Wirklichkeitsaspekte in den Blick geraten und welche Fragen gestellt und genauer untersucht werden können (Appelsmeyer et al., 1997). Ein methodologisches Programm, das die Forschungsmethoden bereits a priori auf Experiment, Statistik und standardisierte Erhebungsinstrumente festlegt, kann einen derart begründeten Wissensaufbau jedoch nicht leisten, weshalb sich die qualitative Forschung immer wieder gegen ein so eingeschränktes Methodenverständnis gewendet und z. T. auch ihr Selbstverständnis auf die Ablehnung quantitativer Methoden gegründet hat. So sieht z. B. Strauss (1991, 26) einen besonderen Vorzug der qualitativen Methoden darin, dass sie ohne die Anwendung mathematischer und vor allem statistischer Operationen auskommen, und Bogdan & Taylor (1975, 5) weisen darauf hin:

"When we reduce people to statistical aggregates, we lose sight of the subjective nature of human behaviour".

Methoden-Triangulation

Eine grundsätzliche Unvereinbarkeit von quantitativen und qualitativen Forschungsmethoden wird heute jedoch kaum noch vertreten und wurde inzwischen durch das Konzept der Methoden-Triangulation (vgl. Flick, 2000b; Kelle & Erzberger, 2000) abgelöst, das von der wechselseitigen Ergänzung quantitativer und qualitativer Methoden ausgeht.

Der Begriff Triangulation stammt aus der Landvermessung und bezeichnet dort die Bestimmung eines Orts durch Messungen von zwei verschiedenen Punkten aus. Auf Psychologie und Sozialwissenschaften übertragen, bedeutet er, dass derselbe Sachverhalt unter zwei verschiedenen empirischen, theoretischen und/oder methodologischen Standpunkten betrachtet wird.

Von Denzin (1978) in die sozialwissenschaftliche Methodenliteratur eingeführt, bezeichnete er zunächst eine Strategie zur Validierung von Forschungsergebnissen, wobei vier verschiedene Formen der Triangulation unterschieden werden können, deren erste beide verschiedene empirische Standpunkte miteinander verbinden:

- die Daten-Triangulation, die in der Kombination von Daten besteht, die verschiedenen Quellen entstammen und zu verschiedenen Zeitpunkten, an verschiedenen Orten oder an verschiedenen Probanden erhoben werden, und
- die Investigator-Triangulation, die zur Verbesserung der Objektivität der Datenerhebung und/oder Datenauswertung auf verschiedene Beobachter, Interviewer und/oder Kodierer zurückgreift, die das erhobene Datenmaterial auswerten.

Die anderen beiden Strategien verbinden verschiedene theoretische und/oder methodologische Herangehensweisen miteinander:

- die Theorien-Triangulation, indem sie sich dem Forschungsgegenstand von verschiedenen theoretischen Perspektiven und Hypothesen her annähert, und
- die Methoden-Triangulation, die verschiedene Forschungsmethoden zum Einsatz bringt, die – wie z. B. beim Einsatz verschiedener Fragebogen – entweder innerhalb derselben Methodologie beheimatet sind (within-method triangulation) oder – wie z. B. beim gleichzeitigen Einsatz quantitativer und qualitativer Forschungsmethoden – verschiedenen Methodologien entstammen (between-method triangulation).

Geht man davon aus, dass die Angemessenheit einer Methode nicht unabhängig von der jeweiligen theoretischen Position ist, so können die letztgenannten Spielarten der Triangulation jedoch kaum voneinander getrennt werden. Indem jede Methode den Gegenstand, den sie erforscht, auf spezifische Weise konstruiert (vgl. Bloor, 1997, 39), beinhaltet die Methoden-Triangulation mehr als nur das Gegeneinander-Ausspielen der Methoden (Denzin, 1978, 304). Sie geht immer auch mit dem Gegeneinander-Ausspielen verschiedener theoretischer Positionen einher (vgl. Blaikie, 1991). Entsprechend versteht auch Denzin die Triangulation heute nicht mehr bloß als eine Strategie zur Validierung von Forschungsergebnissen, sondern als Strategie zur Erlangung eines tieferen Verständnisses des untersuchten Gegenstandes (vgl. Denzin, 1989; Denzin & Lincoln, 1994), oder, anders ausgedrückt: als Strategie, um "Erkenntnisse durch die Gewinnung weiterer Erkenntnisse zu begründen und abzusichern" (Flick, 2000b, 311).

Damit ist die Auseinandersetzung um quantitative vs. qualitative Forschungsmethoden, die lange Zeit als geradezu erbitterter

Streit geführt worden war, schließlich wieder auf jenem methodologischen Reflexionsniveau angelangt, das bereits während der 30er Jahre für die berühmte und viel zitierte Studie über die Arbeitslosen von Marienthal (Jahoda et al., 1933) grundlegend gewesen ist und das Lazarsfeld (1980, 14) mit den Worten beschreibt:

> "Wir konnten uns nicht damit begnügen, Verhaltenseinheiten einfach zu 'zählen'; unser Ehrgeiz war es, komplexe Erlebniswelten empirisch zu erfassen. Der oft behauptete Widerspruch zwischen 'Statistik' und phänomenologischer Reichhaltigkeit war sozusagen von Anbeginn unserer Arbeiten 'aufgehoben', weil gerade die Synthese der beiden Ansatzpunkte uns als die eigentliche Aufgabe erschien".

Gegenstandsverständnis und Wissensideal

Tatsächlich unterscheiden sich quantitative und qualitative Sozialforschung ja nicht erst in den Methoden, welche sie favorisieren. Sie unterscheiden sich bereits hinsichtlich des Gegenstandsverständnisses[1] und des Wissensideals[2], relativ zu denen die Methoden als geeignet begründet oder als ungeeignet verworfen werden können. Während der quantitativen Sozialforschung ein nomologisches Theorienverständnis zugrunde liegt, wie es im deduktiv-nomologischen (vgl. Bd. 1, Kap. 2.3) und im induktiv-statistischen Erklärungsmodell (vgl. Bd. 1, Kap. 2.6) zum Tragen kommt, basiert die qualitative Sozialforschung auf einem interpretativen Theoriebegriff, wie er im intentionalen (vgl. Bd. 1, Kap. 3.2) und im narrativen Erklärungsmodell (vgl. Bd. 1, Kap. 3.3) ausgearbeitet wurde. Während die nomologische Psychologie darauf abzielt, menschliches Erleben und Verhalten relativ zu allgemeingültigen empirischen Gesetzmäßigkeiten zu erklären, geht es in der interpretativen Forschung primär darum, die subjektive Welt des Menschen aus seiner Innenperspektive heraus zu rekonstruieren und die sozialen und kulturellen Regeln aufzudecken, aufgrund derer sie sich konstituiert. Während sich die nomologische Psychologie vorrangig für das objektseitig definierte Verhalten des Menschen (vgl. Bd. 3, Kap. 4) und seine neurophysiologischen Korrelate (vgl. Bd. 3, Kap. 5) interessiert, wird die subjektive Welt des Menschen in der interpretativen Forschung als eine sinn- und bedeutungstrukturierte Wirklichkeit begriffen (vgl. Bd. 1, Kap. 3.3.3).

1. D. h. darin, welche Fragen sie für angemessen halten (vgl. Bd. 1, Kap. 1.4).
2. D. h. darin, welche Antworten sie als zufriedenstellend erachten (vgl. Bd. 1, Kap. 1.4).

Empirieverständnis

Wie wir bereits in Bd. 1 (Kap. 3.3.3) diskutiert haben, schlägt sich dieser methodologische Perspektivenwechsel auch in einem geänderten Empirieverständnis nieder, das nun nicht mehr (wie z. B. das psychologische Experiment) auf eine methodisch-technisch produzierte Empirie abzielt, sondern darauf, die ohnedies gemachten lebensweltlichen Erfahrungen, Handlungen und Interaktionen in ihrem Sinnzusammenhang zu verstehen. Während in der nomologischen Psychologie die Theorie den Daten vorgeordnet ist (vgl. Holzkamp, 1972), besteht in der interpretativen Forschung somit ein Primat der Empirie vor der Theorie, wie ihn z. B. Hoffmann-Riem (1980) auf das Konzept des Verzichts auf Hypothesenbildung ex ante zugespitzt hat.[1]

Offene Erhebungsinstrumente und Dokumentenanalyse

Konkret bedeutet dieser Primat der Empirie vor der Theorie erstens, dass den Mitgliedern des Forschungsfeldes, welches untersucht wird, der Expertenstatus bezüglich ihrer eigenen subjektiven Wirklichkeit zuzuerkennen ist (Schütze, 1977). Zweitens ergibt sich daraus die Bevorzugung offener Erhebungsinstrumente, bei denen die Probanden nicht von vornherein auf bestimmte Antwort- oder Reaktionsmöglichkeiten festgelegt sind (Witzel, 1982), und drittens bedeutet er, dass sich die interpretative Forschung darüber hinaus auch auf bereits vorfindliche und gleichsam naturwüchsig entstandene Objektivationen der subjektiven Welt des Menschen (z. B. Therapieprotokolle) stützen kann und dies über weite Strecken auch tut. Namentlich die aus der Propagandaforschung (Lasswell, 1927) hervorgegangene Analyse von Medienberichten zwecks Untersuchung der sozialen Konstruktion der Wirklichkeit kann hier auf eine lange Tradition – nicht nur in der Psychologie, sondern auch in anderen Sozialwissenschaften – zurückblicken.

1.1.2 Qualitative Erhebungsinstrumente

Methodologische Prinzipien

Die Bedeutung eines Sachverhaltes konstituiert sich durch Interpretation, d. h. sie resultiert daraus, dass er in einen bestimmten Kontext gestellt wird. Um die (subjektive) Bedeutung eines Sachverhaltes für einen Probanden zu verstehen, ist es daher erforderlich, die vom Probanden selbst vollzogenen Interpretationsleistungen nachzuvollziehen, d. h. jene Kontexte zu rekonstruieren, in denen er vom Probanden selbst gesehen wird.

1. Zur Kritik dieses Konzeptes vgl. Bd. 1, Kap. 3.9.1.

Um dies leisten zu können, stützt sich die interpretative Sozialforschung auf eine Reihe von methodologischen Prinzipien (vgl. Bd. 1, Kap. 3.9.1), die nicht erst bei der Auswertung von Textmaterial, sondern bereits bei der Konstruktion der Erhebungssituation zu beachten sind. Dazu gehören insbesondere das *Prinzip der Offenheit*, wonach die Erhebungssituation für Selbstthematisierungen des Probanden offen zu halten ist, sowie das *Prinzip der Kommunikation*, wonach dafür Sorge zu tragen ist, dass der Proband in der ihm vertrauten Sprache und nach den ihm vertrauten Regeln der kommunikativen Verständigung sprechen kann.

Überblick über qualitative Erhebungsinstrumente

Das Spektrum der qualitativen Erhebungsinstrumente, welche dies gewährleisten sollen, reicht von den verschiedenen Formen der teilnehmenden Beobachtung (z. B. Dechmann, 1978; Girtler, 1984)[1] über das Rollenspiel als einer gleichsam experimentellen Variante der teilnehmenden Beobachtung (z. B. Volmerg et al., 1983), und verschiedene Formen des qualitativen Interviews (zu einem Überblick vgl. Lamnek, 1989, 35ff; Spöhring, 1989, 147ff.; Flick, 1995, 94ff.; Friebertshäuser, 1997; Hopf, 2000) bis hin zur Gruppendiskussion (Nießen, 1977; Bohnsack, 2000; Lamnek, 2005b), die als sozialwissenschaftliche Forschungsmethode lange Zeit eher stiefmütterlich behandelt wurde, unter dem Label Focus Group aber neuerdings zunehmende Beachtung findet.

Genetische Rekonstruktion von Handlungen

Auch wenn die Prinzipien der Offenheit und der Kommunikation durch qualitative Erhebungsinstrumente weitgehend gewährleistet sind, können wir nach Kraus (2001, 242) jedoch nicht davon ausgehen, dass die Sinnkonstitution etwas ist, das gleichsam in einem Individuum passiert, welches dem Wissenschaftler sodann darüber Auskunft gibt. Vielmehr wird vor unseren Ohren und Augen in der Interaktion gemeinsam Sinn konstituiert (vgl. Bd. 1, Kap. 2.9.2). Welche Konsequenzen dies für die Gestaltung der Erhebungssituation nach sich zieht, wollen wir im Folgenden anhand der genetischen Rekonstruktion von Handlungen (vgl. Bd. 1, Kap. 3.2.3) diskutieren, indem wir einen Blick auf die Interviewtechniken und Techniken der Gesprächsführung werfen, mittels derer wir jene Art von Texten (Interviewprotokollen) erhalten können, die für die interpretative Analyse von Handlungsverläufen benötigt werden.

---

1. Zu einer vergleichenden Diskussion siehe Lamnek, 1989, 297ff.

Arten der Handlungsbeschreibung

Grundsätzlich stehen uns zwei Arten der Handlungsbeschreibung offen: die Darstellung der Handlung als Verwirklichung eines Tätigkeitsmusters (= schematische Handlungsbeschreibung) oder ihre Darstellung als Teil einer Geschichte. Eine Handlungsbeschreibung, welche die Bedeutung des Handelns nicht übergehen will, ist jedoch niemals die Beschreibung einer isolierten Tätigkeit, sondern immer die Gesamtbeschreibung der Handlung in ihrem Kontext. Auch die "schlichten" schematischen Handlungsbeschreibungen sind nach Schwemmer (1987) *keine kontextfreien Beschreibungen*, sondern nur solche, deren Kontext nicht reflektiert ist, und die die Handlungen meist in einen anderen Kontext rücken als den, in dem sie vom Handelnden ausgeführt worden sind. Die Rekonstruktion der Orientierungen, welche den Gesamtzusammenhang seines Handelns und Erlebens herstellen, kann daher nur auf der Grundlage des Verständnisses jener Geschichte erfolgen, die sich für ihn tatsächlich abgespielt hat und ein Zugang zu dem dafür benötigten empirischen Material kann daher am ehesten über die Ausnutzung der alltagspraktischen Kompetenz des Erzählens gewonnen werden (Kempf, 1987).

Das narrative Interview

Dieser Grundgedanke liegt auch dem narrativen Interview nach Schütze (1977, 1982, 1983) zugrunde, das auf eine vorhergehende Strukturierung der Interviewinhalte durch den Interviewer weitgehend verzichtet. Ein Leitfaden ist zwar vorhanden, "seine Bedeutung schrumpft jedoch durch die Eigenleistung der Forschungssubjekte" (Hoffmann-Riem, 1980, 360). Stattdessen soll durch das "Hervorlocken von Geschichten" ein Maximum an Kontext geschaffen werden, um die subjektive Realität interpretativ erfassen zu können. Gleichzeitig soll dadurch erreicht werden, dass die Erzählung durch die Bedeutungshierarchie des Erzählers selbst strukturiert wird.

Entsprechend empfiehlt Schütze eine Zweiteilung des Interviews in eine Phase der Haupterzählung und eine des narrativen Nachfragens.

"Narrative Interviews setzen in ihrer Anwendung voraus, dass hinsichtlich des interessierenden Gegenstandsbereiches eine zentrale Eingangsfrage mit narrativer Generierungskraft formulierbar ist. Nur dort, wo ein Informant als Akteur oder Betroffener auf die Verstrickung in einen lebensgeschichtlichen oder historischen Ereigniszusammenhang (eine 'Geschichte') zurückblicken kann, ist die Formulierung eines narrativen Themas möglich" (Schütze, 1982, 570).

Die Phase der Haupterzählung

Die Anfangsthemenstellung hat die Funktion, eine längere Haupterzählung zu initiieren. Die Eingangsfrage (mitunter als narrativer Stimulus bezeichnet) soll so gestaltet werden, dass der Forscher sein Interesse am gesamten Hintergrundzusammenhang bekundet, ohne jedoch bereits eine allgemeine Erzählfolie in allen relevanten Aspekten vorzugeben. Gleichzeitig soll die Erzählung durch sie thematisch begrenzt werden, da ansonsten das Problem auftauchen kann, dass der Erzähler den narrativen Zusammenhang (die Gesamtgestalt und den 'roten Faden') der berichteten Ereignisse nicht mehr überblickt.

"Zudem muss der in der Erzählung darzustellende Ereigniszusammenhang zumindest in einem Hauptereignishöhepunkt, einem Ereignis zu Beginn und einem Ereignis gegen Ende des Ereignisablaufes noch im Gedächtnis des Erzählers haften, um überhaupt die Motivation fürs Erzählen über die vorgeschlagene Thematik zu schaffen" (Schütze, 1982, 574).

Während der Haupterzählung beschränkt sich der Interviewer strikt auf die Rolle des Zuhörers mit den entsprechenden erzählunterstützenden Signalen (Kopfnicken, 'hm, hm' usw.), bis der Interviewte durch eine eindeutige Koda zu erkennen gibt, dass seine Erzählung abgeschlossen ist.

Die Phase des narrativen Nachfragens

Die danach einsetzenden Nachfragen des Interviewers dürfen nach Schütze (1982, 570) nicht als Theorie- oder Meinungsfragen formuliert sein, wie z. B. Fragen nach Gründen, Aufforderungen zu Meinungskundgaben usw., sondern sollen *neue Erzählsequenzen* zu Darstellungsbereichen initiieren, die bisher nicht genügend oder überhaupt nicht ausgeführt wurden.

Um dem Interviewer die Möglichkeit zu geben, seine eigenen thematischen Interessen in die Interviewsituation einzubringen, den Erzählfluss des Interviewten zugleich aber so wenig wie möglich zu stören, sollen durch *immanente Fragen* zunächst solche Themenbereiche aufgegriffen werden, die der Interviewte bereits selbst angesprochen hat. Nur wenn er bestimmte Punkte der Vorab-Erzählfolie gänzlich auslässt, sollen am Ende des Interviews auch *exmanente Fragen* eingebracht werden.

Das partnerzentrierte Gespräch

Hoffmann-Riem (1980, 347) hat darauf hingewiesen, dass der Aufbau der Kommunikationsbeziehung darüber entscheidet, "welche Tiefenschicht der Datengewinnung erreicht wird". Damit gewinnt die Art der Gesprächsbeziehung zwischen dem Interviewer und dem Interviewten besondere Bedeutung, weshalb

Kempf (1987) mit dem partnerzentrierten Gespräch eine psychologische Variante des narrativen Interviews vorgeschlagen hat, die Elemente der klientenzentrierten Gesprächsführung nach Rogers (1972) in die Interviewsituation einbringt.

Insbesondere für die *psychologische* Interviewführung stellt sich nämlich die Frage, wie die Gesprächssituation so gestaltet werden kann, dass der Interviewte tatsächlich bereit ist, seine subjektive Realität der Dinge ein Stück weit offen zu legen; selbst dort, wo ihm manche Aspekte vielleicht peinlich sind oder wo er Widerstände dagegen hat, diese Aspekte öffentlich zu machen. Damit unterscheidet sich das psychologische Interview in wesentlichen Punkten vom soziologischen, wie Schütze es im Sinn hat.

Verhaltensregeln im narrativen Interview

Schütze sieht den Interviewten vorwiegend als *Informanten*, den es in bestimmte *Zugzwänge des Erzählens* zu verwickeln gilt. Dem Interviewer weist er die Rolle des produktiv zuhörenden Erzählpartners zu, der sein Interesse an der Erzählung signalisiert und seinen Einfluss auf den Interaktionsprozess und die Entwicklung der Erzählung durch Einhaltung bestimmter *Regeln* auf ein Minimum beschränkt.

Solche Regeln sind u. a.: der Verzicht auf Theorie-, Meinungs- und Warum-Fragen; die Vorschrift, nicht zu erkennen zu geben, wenn dem Interviewer Teile der Erzählung bereits bekannt sind; der Verzicht darauf, Auslassungen explizit nachzufragen, usw. Insbesondere legt Schütze Wert darauf, dass die Fragen des Interviewers nicht als Eindringen in die private Sphäre erlebt werden und daher nur Sachverhalte von *öffentlichem* Interesse betreffen dürfen.

Die Überwindung der Distanz zwischen Forscher und Forschungssubjekt

Im *psychologischen Interview* kann die Beschränkung auf Sachverhalte von öffentlichem Interesse dagegen nicht aufrechterhalten werden und zumindest im psychologischen Tiefeninterview, aber auch z. B. im diagnostischen Anamnesegespräch, bedarf es auf Seiten des Interviewten sehr wohl auch einer *Bereitschaft zur Preisgabe privater Realität*.

Daraus ergibt sich die Notwendigkeit der Überwindung der Distanz zwischen Forscher und Forschungssubjekt, wie sie Smedslund (zit. n. Kempf, 1987) sehr treffend formuliert hat:

"If you want to study a thing, you have to move closer. Ceate a situation of trust, show care, respect and understanding".

Was dem Forscher damit abverlangt wird, ist jedoch mehr als nur die Entwicklung bestimmter Gesprächstechniken, damit sein Gegenüber sich nicht "ausgefragt" fühlt. Er muss eine bestimmte Haltung gegenüber seinem Gesprächspartner einnehmen, die nicht nur einen solchen Eindruck vermeiden hilft, sondern von vorneherein ausschließt, dass der Interviewte bloß zu Forschungszwecken funktionalisiert wird.

Die alltagsweltliche Kommunikationsform, an die wir hierfür anknüpfen können, ist die eines vertraulichen Gesprächs zwischen einem Erzähler und einem Zuhörer, die ein gemeinsames Interesse *aneinander* und *am Thema* des Gesprächs miteinander verbindet.

Wenn unser Gesprächspartner uns Vertrauen und Offenheit entgegenbringen soll, müssen auch wir dasselbe tun. Dies impliziert insbesondere auch die Notwendigkeit der Offenlegung des Interesses, das der Forscher an dem Gespräch hat (denn nur so kann sich ein *gemeinsames* Interesse herausbilden), und der Rollenverteilung während des Gesprächs, wonach das Interesse des Forschers ein Hörinteresse und das Interesse seines Gesprächpartners ein Erzählinteresse ist.

Die Fiktion einer *egalitären Kommunikation* (Hoffmann-Riem, 1980, 380) bewirkt allzu leicht eine Rollenverwischung, die am Ende Zweifel an der Offenheit des Forschers entstehen lässt und die Gesprächssituation belastet.

Kontaktaufnahme und Small-talk-Phase

Gleichzeitig gewinnen auch die Kontaktaufnahme und die dem eigentlichen Gespräch vorangehende Small-talk-Phase eine besondere Bedeutung für den Gesprächsverlauf und die damit erreichbare Tiefenschicht der Datengewinnung. Sie sind nämlich erstens entscheidend für die Herstellung des Rapports zwischen dem Forscher und seinem Gesprächspartner und müssen zweitens für die Konstitution eines gemeinsamen Interesses an dem Gespräch genutzt werden.

Damit wird die Trennung zwischen einer Small-talk-Phase und der das eigentliche Gespräch eröffnenden Eingangsfrage eher künstlich. Denn die Bedeutung der Eingangsfrage wird vom Kontext der vorangegangenen Kommunikation zwischen dem Forscher und seinem Gesprächspartner (mit-)bestimmt und kann davon nicht getrennt werden. Die Eingangsfrage hat daher hauptsächlich die Funktion, das Gesprächsthema noch einmal auf den Punkt zu bringen.

Elemente der klientenzentrierten Gesprächsführung

Um es dem Gesprächspartner zu ermöglichen, seine subjektive Realität der Dinge offenzulegen, ist es erforderlich, eine *personenzentrierte Grundhaltung* einzunehmen und auch zu kommunizieren. Diese ist nach Rogers (1972) definiert durch:

- einfühlendes Verstehen,
- Wertschätzung und emotionale Wärme sowie
- Echtheit und Selbstkongruenz.

Wenn unser Gesprächspartner erzählen soll, was er tatsächlich erlebt hat, und nicht nur seine – bereits durch Rationalisierungen, Abstraktionen und Generalisierungen etc. bereinigten – Alltagstheorien darüber, dann hat dies Konsequenzen für die Art der Fragen bzw. ihrer Formulierung, die Smedslund (zit. n. Kempf, 1987) in die Regel zusammengefasst hat:

"Don't ask what goes on inside people, but ask what people go on inside of."

Diese Regel ist parallel zu Schützes Forderung nach der Vermeidung von Theorie-, Meinungs- und Warum-Fragen. Sie geht aber darüber hinaus, indem sie aufzeigt, wie wir stattdessen fragen können:

- Statt nach einer Erklärung zu fragen (Warum-Fragen),
- fragen wir besser, *wie* es sich *so* ergeben hat (Wieso-Fragen),

oder wir richten eine *implizite* Aufforderung an den Sprecher, das Gesagte mit mehr Kontext zu versehen, indem wir ihm das Gesagte mit einem leicht fragenden Anheben der Stimme am Satzende *spiegeln*.

Um die Bereitschaft zur Preisgabe subjektiver Realität zu schaffen bzw. nicht zu zerstören, ist es erforderlich, die subjektive Realität des Gesprächspartners zu *akzeptieren*, ihn nicht etwa in Diskussionen zu verwickeln oder unter Rechtfertigungsdruck zu setzen und sich mit eigenen Stellungnahmen zurückzuhalten. Das bedeutet aber *nicht* – wie im narrativen Interview gefordert –, dem Gesprächspartner zu signalisieren, dass man seine Typisierungen in einer Weise nachvollzieht, "als ob beim Austausch der Standorte, der Zuhörer als Erzähler die Dinge in der selben Typikalität sähe" (Hoffmann-Riem, 1980, 359); sondern es bedeutet, dass man ihm signalisieren muss, dass es in dem Gespräch nicht um irgendeine "objektive" Realität geht, über die sich Forscher und Gesprächspartner zu einigen hätten, sondern einzig und allein um die Sicht des Gesprächspartners, und dass man bereit ist, sich auf diese seine subjektive Realität voll und ganz einzulassen.

Dazu gehört es, den Erzählfluss des Gesprächpartners nicht durch voreilige Fragen oder ad-hoc-Interpretationen zu stören. Indem wir dies tun, auch (vermeintlichen) Themenabweichungen folgen und Pausen aushalten, vermitteln wir dem Gesprächspartner bezüglich der *Form* seiner Erzählung jene "nicht an Bedingungen geknüpfte Wertschätzung" (Tausch & Tausch, 1981, 70), die wir ihm auch bezüglich der *Inhalte* seiner Erzählung entgegenbringen müssen, wenn wir die Bereitschaft zur ungeschminkten Darstellung seiner subjektiven Realität nicht im Keim ersticken wollen.

Kunstfehler der Gesprächsführung

Im Gegensatz zu der weit verbreiteten Einstellung, welche die Fähigkeit zur Durchführung qualitativer Interviews als einen selbstverständlichen und relativ unproblematischen Bestandteil der Qualifikation von Sozialwissenschaftlern ansieht, sind die Techniken der Gesprächsführung allerdings nichts, das einem gleichsam in die Wiege gelegt ist, sondern sie bedürfen eines intensiven Trainings[1].

Zwar gibt es nach Hopf (2000, 358) in den Sozialwissenschaften einen relativ breiten Konsens darüber, dass die Interviews von Interviewern durchgeführt werden sollten, die mit dem theoretischen Ansatz, den Fragestellungen und den Vorarbeiten des jeweiligen Forschungsprojektes vertraut sind, so dass sie zu einer autonomen Gesprächsführung in der Lage sind und abschätzen können, wann es inhaltlich angemessen ist, vom Leitfaden abzuweichen, an welchen Stellen es erforderlich ist, intensiver nachzufragen oder auch an welchen Stellen es sich empfiehlt, nur sehr unspezifisch zu fragen und den Befragten einen breiten Artikulationsspielraum einzuräumen.

Über die Forderung nach inhaltlich-theoretischer Kompetenz hinaus herrschen in der Literatur jedoch mitunter recht naive Vorstellungen, die den Eindruck wecken, dass sich die Kompetenzprobleme mit der Narration gleichsam von selbst erledigen (z. B. Hoffmann-Riem, 1980, 357ff.). Die daraus resultierenden Kunstfehler der Gesprächsführung sind nach Hopf (2000, 358) zu einem Teil Planungsfehler, die auf einer Fehleinschätzung des Verhältnisses von Informationsinteressen und zur Verfügung stehender Zeit beruhen und überlange Interviewleitfäden hervorbringen, die eine Tendenz zum oberflächlichen Abhaken der

1. So beträgt die Ausbildung in klientenzentrierter Gesprächsführung (= Grundstufe der Ausbildung in klientenzentrierter Beratung) nach den Richtlinien der GwG volle 300 Ausbildungsstunden über einen Zeitraum von zwei Jahren.

Fragen nach sich ziehen. Zu einem anderen Teil sind sie Ausbildungsdefiziten und/oder einer mangelnden Supervision der Interviewer geschuldet.

Als typische ausbildungsbedingte und durch Mangel an Erfahrung verursachte Kunstfehler bei der Durchführung qualitativer Interviews, die unter anderem mit der Angst vor der ungewohnten Situation einer intensiven Kommunikation mit Fremden zu tun haben, nennt Hopf (ebd., 359):

- die Tendenz zu einem dominierenden Kommunikationsstil, der z. B. in der Häufung suggestiver Fragen, suggestiver Vorgaben und Interpretationen (vgl. Richardson et al., 1979) und/oder in der Häufung von bewertenden und kommentierenden Aussagen, die zwar meist unterstützend gemeint, im Endeffekt jedoch lenkend und ggf. störend sind;
- Probleme mit den passiv-reziptiven Anteilen der Gesprächsführung, wie sie etwa in fehlender Geduld beim Zuhören und/oder in Defiziten beim Aufgreifen von Anhaltspunkten für Nachfragen zum Ausdruck kommen; sowie
- mangelnde Souveränität im Umgang mit dem Leitfaden, was sich z. B. darin niederschlägt, dass dieser immer wieder penetrant ins Gedächtnis gerufen oder als Disziplinierungsinstrument verwendet wird und/oder den Blick für relevante aber nicht antizipierte Aspekte des jeweiligen Themas verstellt (vgl. Merton & Kendall, 1979, 185).

Angesichts der Komplexität der inhaltlich-theoretischen und sozial-psychologischen Aspekte der Interviewpraxis ist es nach Hopf daher unerlässlich, sich mit den Interviewtechniken nicht nur theoretisch vertraut zu machen, sondern sie in Form von Interviewertrainings praktisch einzuüben, mögliche Kunstfehler zu diskutieren und textbezogen an einzelnen Interviewpassagen zu erörtern.

1.1.3 Transkriptionsverfahren

War man in den Anfangstagen der qualitativen Sozialforschung noch darauf angewiesen, während der Interviews, Gruppendiskussionen und dergleichen Gesprächsnotizen anzufertigen, so stellte die Erfindung des Tonbandes und der Videokamera und die damit ermöglichte Praxis von Audio- und Videoaufnahmen geradezu einen Meilenstein auf dem Weg zu einer objektivierbaren qualitativen Forschungsmethodik dar. Erst dadurch wurde die

Möglichkeit eröffnet, die gesprochene Rede (einschließlich paraverbaler – z. B. Lachen oder Hüsteln – und nonverbaler Äußerungen – insbes. Mimik und Gestik) – wortgetreu zu verschriftlichen.

Gleichwohl darf nicht übersehen werden, dass auch die daraus resultierenden Transkripte keine Eins-zu-Eins-Wiedergabe der transkribierten Gespräche sind, sondern eine erhebliche Reduktion der fast unbegrenzt reichhaltigen Gesprächsmerkmale (= Primärdaten) und ihrer Aufzeichung (= Sekundärdaten) darstellen (vgl. Cook, 1990). Hinzu kommt, dass schon in den Audio- und Videoaufnahmen einzelne Textpassagen unverständlich sein können. Nonverbale Äußerungen sind nur über Videoaufnahmen zugänglich und können u.U. außerhalb des Kamerablickwinkels stattfinden, etc. Ein weiteres, grundsätzliches Problem besteht darin, dass der Transkribierende z. T. solche Phänomene zu Papier bringen muss, die er als natürlicher Sprecher zu überhören gelernt hat – z. B. Versprecher, Wortabbrüche, Wiederholungen oder Füllaute (vgl. Kowal & O'Connell, 2000, 444). Zugleich muss bei der späteren Analyse der Transkripte berücksichtigt werden, dass Transkribierende nicht selten unzuverlässig transkribieren (vgl. Lindsay & O'Connel, 1995; O'Connell & Kowal, 1994), weshalb man die Textanalyse nicht allein auf das Transkript stützen kann, sondern zu ihrer Validierung auch immer wieder auf die Audio- bzw. Videoaufnahmen zurückgreifen muss.

Damit die Transkripte ein brauchbares Handwerkzeug darstellen, muss zudem auf ihre Lesbarkeit geachtet werden, was gewisse Abstriche bezüglich ihrer Präzision nach sich zieht. So sind zwar verschiedene Formen der Verschriftlichung wie (1) Standardsprache, (2) literarische Umschrift, (3) Dialektumschrift und (4) phonetische Umschrift nach dem internationalen phonetischen Alphabet möglich. Je präziser die Umschrift, desto geringer ist jedoch die Lesbarkeit, weshalb die Dialektumschrift und insbesondere die phonetische Umschrift nur für linguistische Fragestellungen geeignet sind. Die standardsprachliche Umschrift hingegen vernachlässigt die Besonderheiten der gesprochenen Sprache, wie z. B. die Auslassung einzelner Laute ("gehn" statt "gehen"). Diese Abweichungen werden in der literarischen Umschrift berücksichtigt, die sich am besten eignet, weil ja auch Satzbau und Grammatik der gesprochenen Sprache oft von der Standardsprache abweichen.

Die in Tab. 1.1.3.1 wiedergegebenen Transkriptionsregeln versuchen dies zu berücksichtigen und stützen sich auf Vorschläge von Ehlich & Switalla (1976) und Mergenthaler (1989).

Tab. 1.1.3.1: Transkriptionsregeln

Zeilenschreibweise	Aufeinanderfolgende Sprechakte werden jeweils mit einer neuen Transkriptzeile begonnen, so dass das Untereinander der Zeilen das Nacheinander der Sprechakte abbildet.
Identifikation von Sprechakt und Sprecher	Die Sprechakte werden laufend durchnummeriert (001, 002, 003, usw.) und die Sprecher durch ein geeignetes Kürzel identifiziert; z. B. "P" für Peter, "I" für Interviewer.
Transkriptionsmodus der verbalen Äußerungen	Mischform aus Standardsprache und literarischer Umschrift. D. h. die in der gesprochenen Sprache üblichen Verschleifungen ("s" für "es", "nen" für "einen", usw.) werden in das Transkript übernommen, jedoch nur so weit, dass sie die Lesbarkeit nicht beeinträchtigen. Deshalb sind Verschleifungen wie "ham" für "haben", "ma" für "man", "wa"/"ma" für "wir" von der literarischen Umschrift ausgenommen.
Interpunktion	Normale Interpunktion. Am Satzende steigende Intonation wird durch "?", fallende Intonation durch "." dargestellt.
Unvollendete Sätze oder Nebensätze	… enden mit "…".
Sprechpausen	Kürzere Pausen werden durch "…", Pausen ab 3 Sekunden durch Angabe der Pausenlänge in doppelten Klammern dargestellt; z. B. ((5 sec)).
Hervorhebungen, Akzentuierungen	Betont Gesprochenes wird durch einfache Unterstreichung markiert. Beispiel: P: 009 Ja, und dann, ist das Blödsinnigste <u>überhaupt</u>, das überhaupt dort abzustellen, also einfach <u>dort</u>. Also, hättest Dir ja <u>denken</u> können, dass da vielleicht was passiert.
Akustisch Unverständliches	… wird durch Leerzeichen in einfachen Klammern dargestellt; hat der Transkribierende eine Vermutung, was die unverständlichen Worte sein könnten, durch die erratenen Worte in einfachen Klammern. Beispiele: R: 032 Ich denk, gut, vielleicht bin ich's auch, vielleicht müssen wir uns beide ( ). P: 125 Jaja, (das wär's). Das wär's tatsächlich gewesen.

Fortsetzung
Tab. 1.1.3.1: Transkriptionsregeln

Parasprachliche (z. B. Lachen, Hüsteln) und nonverbale Äußerungen (z. B. Mimik, Gestik) sowie Modulation der Stimme	… werden unsystematisch erfasst und in Doppelklammern beschrieben. Beispiel: P: 069 ((lachend)) Ja, eh.. also aber dann zahl ich's ja auch. R 070 ((blickt zum Himmel)) Ja eben! Das musst Du mir doch nicht sagen! P: 071 ((schreit)) Und Du sei mal ganz ruhig! Ausnahme: "äh"s und "mhm"s werden in jedem Fall und ohne Klammern transkribiert
Abbrüche und Unterbrechungen	Abbrüche werden durch das Zeichen "-" am Ende des Sprechaktes markiert. Bei Unterbrechungen wird das Zeichen am Anfang des nächsten Sprechaktes wiederholt. Beispiele: R: 062 Wenn ich z. B. an dem Tag das Auto für mich gehabt hätte, und ich hätt'n Ticket bekommen, ich würd's zahlen. Weil das wär ganz klar meine Sache- P: 063 -Ja- R: 064 -Aber in dem Fall isses nicht ganz klar meine Sache, weil wir ne andere Vereinbarung hatten.
Simultanes Sprechen mehrerer Sprecher	Simultan Gesprochenes wird in eckige Klammern gesetzt. Beispiel: P: 019 Ja ich konnt ja nicht raus, ich war ja beim Professor die ganze Zeit. [Also ich mein, ich konnt ja nicht.] R: 020 [Ja und?] Ich mein, das kann man doch vorher absehen.
In den Sprechakt eines anderen Sprechers eingeschobene parasprachliche Äußerungen	… werden nicht als seperater Sprechakt verschriftet, sondern in die Äußerung des anderen Sprechers in einfachen Klammern eingefügt. Beispiele: R: 096 Du kennst das ja selbst (P: Mhm) und da ist man einfach aufgeschmissen. R: 102 ((lachend)) Ja, und dann hab ich das halt so gemacht (P: lacht mit) und das ist dabei herausgekommen.
Kommentare des Transkribierenden	… werden in spitze Klammern gesetzt. Beispiel: R 054 ((lachend <ironisch>)) Das ist ja fein. So habe ich mir das schon immer erträumt.

1.1.4 Qualitative Analysemethoden

Methodologische Prinzipien

Zwei weitere der methodologischen Prinzipien, die wir bereits in Bd. 1 (Kap. 3.9.1) diskutiert haben, sind zwar ebenfalls bereits bei der Konstruktion der Erhebungssituation (z. B. bei der Gesprächsführung im Interview) zu berücksichtigen, kommen aber insbesondere bei der Auswertung von Textmaterial zum Tragen. Es sind dies das *Prinzip der Fremdheit*, wonach die (soziale) Wirklichkeit, welche in einem Text zum Ausdruck kommt, in ihrer jeweiligen Eigenart, potentiellen Andersheit und Fremdheit zu betrachten ist, weshalb ihr nicht unmittelbar die eigenen Begriffe, Kategorien und Interpretationsmuster übergestülpt werden dürfen, sowie das *Prinzip der Reflexion*, wonach der Forscher seine eigenen (oft impliziten) Vorannahmen ständig reflektieren, offen legen und in Frage stellen muss.

Überblick über qualitative Analysemethoden

Für die qualitative Analyse von Textmaterial wurde in den letzten Jahrzehnten eine Vielzahl an Methoden entwickelt, die – um nur einige der prominentesten zu nennen – von der Grounded Theory (Glaser & Strauss, 1967, 1979; Glaser, 1978; Strauss, 1991; Strauss & Corbin, 1996), die allerdings nicht nur eine Auswertungsmethode, sondern eine umfassende Strategie qualitativer Forschung darstellt (vgl. Bd. 3, Kap. 3), über die psychoanalytische Textinterpretation (Leithäuser & Volmerg, 1979, 1988; Lorenzer, 1981, 1986; einführend: König, 2000), die objektive Hermeneutik (Oevermann, 1983, 1993; einführend: Reichertz, 2000), die Konversationsanalyse (Sudnow, 1972; Turner, 1974; Schenkein, 1978; Psathas, 1979; einführend: Bergmann, 2000), die Diskursanalyse (Potter & Wetherell, 1987, 1995; van Dijk 1985; Jäger, 1994; einführend: Parker, 2000) und die qualitative Inhaltsanalyse (Mayring, 1983, Mayring & Gläser-Zikuda, 2005; einführend: Mayring, 2000) bis hin zur quantitativen Inhaltsanalyse (Berelson, 1952; Bos & Tarnai, 1989; Früh, 2001) reicht, die einen Grenzfall zwischen qualitativen und quantitativen Forschungsmethoden darstellt.

Flexibilität der methodischen Vorgehensweise

Obwohl diese Methoden aus verschiedenen sozialwissenschaftlichen Disziplinen heraus entstanden sind, zum Teil unterschiedliche Textgenres im Blick haben und von divergierenden Analyseintentionen geleitet sind, lassen sich doch einige Gemeinsamkeiten identifizieren, auf die wir im Folgenden kurz eingehen wollen. Gleichwohl ist vorauszuschicken, dass es nicht *die* qualitative Analysemethode gibt, sondern ein methodisches Spektrum

unterschiedlicher Ansätze, die je nach Fragestellung und Forschungstradition ausgewählt und je nach Forschungsgegenstand modifiziert, variiert und abgewandelt werden können.

Diese Flexibilität der methodischen Vorgehensweise gibt es auch innerhalb der verschiedenen Methoden, die sich daher auch nicht auf ein starres Schema aus vorgegebenen methodischen Anleitungen und Regeln reduzieren lassen, sondern an die jeweilige Fragestellung, die Qualität des Textmaterials und/oder an bisher erarbeitete Teilergebnisse und die dadurch neu aufgeworfenen Fragen anzupassen sind.

Objektivitätssicherung und Transsubjektivität

Die darin zum Ausdruck kommende Abhängigkeit der methodischen Vorgehensweise von der Art des zu analysierenden Textes und der Zielsetzung der Untersuchung ist zwar nichts, was die qualitativen Methoden grundsätzlich von den quantitativen unterscheidet und auch dort kann z. B. die experimentelle Versuchsplanung nicht auf einfache Schemata reduziert werden. Diese Abhängigkeit ist in der qualitativen Forschung jedoch weit augenfälliger als in der quantitativen, wo die Objektivitätssicherung durch Distanzierung vom Forschungsgegenstand erreicht werden kann.

In der quantitativen Forschung können wir auf standardisierte Erhebungsinstrumente zurückgreifen und die so gewonnenen Daten mittels statistischer Verfahren auswerten, deren Anwendung zwar fehlerhaft sein kann, die aber über die Wahl des Auswertungsverfahrens und die Festsetzung des Signifikanzniveaus hinaus kaum einen Spielraum für eine Variation des methodischen Vorgehens lassen. In der qualitativen Forschung ist dies nicht der Fall. Die Sicherung der Objektivität kann hier letztlich nur durch die systematische Reflexion der Subjektivität des Forschers, seines Forschungsinteresses und seines Vorwissens (vgl. Bd. 1, Kap. 3.9.1) gewährleistet und der transsubjektive Geltungsanspruch der Forschungsergebnisse nur über die Transparenz des Analyseprozesses und die Explikation der Regeln sichergestellt werden, denen die Analyse folgt.

Nur in wenigen Fällen, nämlich dort, wo es um das Verstehen des vom Urheber des Textes *gemeinten* Sinns geht, kann man darüber hinaus auch auf eine kommunikative Validierung der Analyseergebnisse setzen, indem man die Forschungssubjekte mit den Interpretationen konfrontiert, wie dies u. a. von Sommer (1987), Scheele & Groeben (1988) und Wilke (1992) vorgeschlagen wur-

de. Geht es dagegen um die dahinter liegenden sozialen, kulturellen oder psychologischen *Regeln der Sinnkonstitution*, so ist auch dies nicht möglich, da diese Regeln implizit und den Betroffenen nicht bewusst sind. Auch in allen anderen Fällen, wo kommunikative Validierungsverfahren zum Einsatz gebracht werden können, erweist sich der Konsens zwischen Interpret und Interpretiertem als ein äußerst zweifelhaftes Validitätskriterium. Denn zum einen können Interpretationen auch dann Gültigkeit haben, wenn die Betroffenen sie ablehnen (vgl. Moser, 1995, 67), und zum anderen können sie auch trotz Zustimmung ungültig sein, wenn diese z. B. aufgrund sozialer Erwünschtheit zustande gekommen ist (vgl. Honer, 1994, 90)

Leitgedanken der Textanalyse

Ungeachtet der Vielfalt textanalytischer Verfahren, die in den verschiedenen sozialwissenschaftlichen Disziplinen entwickelt wurden, und ungeachtet der Variationen, welche diese Verfahren in der Forschungspraxis erfahren, gibt es eine Reihe von Leitgedanken, zu denen sich jede textanalytische Methode so oder so verhalten muss und die wir bereits in Bd. 1 (Kap. 3.9.2) diskutiert haben:

1. Der Analysevorgang verläuft nach den Regeln einer auf Verständigung zielenden Kommunikation und enthält zwei aufeinander bezogene Operationen: Einerseits werden aus der Perspektive des Forschers *Fragen* an den Text gerichtet, andererseits steckt in der *Antwort* immer auch mehr oder anderes, als gefragt wurde. Folglich ergeben sich aus dem Text gleichsam *Rückfragen* an die Angemessenheit, Ergiebigkeit etc. der vom Interpreten an den Text herangetragenen Gesichtspunkte.
2. Ein angemessenes Textverständnis ist daher nicht auf Anhieb möglich, sondern erst das Ergebnis einer Interpretationsarbeit, "in der in vielen einzelnen Schritten immer wieder der Versuch unternommen werden muss, Sinn und Bedeutung des 'Ganzen' aus seinen einzelnen Teilen und die Bedeutungsstellung dieser im Rahmen des Ganzen zu ermitteln" (Mollenhauer & Rittelmeyer, 1977, 59).
3. Die methodische Vorgehensweise erfordert daher einen ständigen Wechsel vom Detail zum Ganzen und zurück (= hermeneutischer Zirkel). Die einzelne Textstelle erschließt sich in ihrer Bedeutung erst, wenn der Gesamtkontext bekannt ist, und die Bedeutung des Gesamtkontextes erschließt sich erst aus der Bedeutung der einzelnen Textstellen.

4. Die Bedeutung einer Textstelle/eines Textes ergibt sich aus ihrem/seinem Kontext. Die möglichen Interpretationen werden daher durch den Text selbst (interner Kontext), den (verfügbaren) externen Kontext, auf den der Text verweist, und durch die interpretationsleitenden Fragen begrenzt, welche der Forscher an den Text richtet.
5. Die Annahme des Vorhandenseins einer gewissen Konsistenz der Bedeutungen ist keine empirische Frage, sondern "ein fundamentales Theorem des erkennenden Umgangs mit Dingen überhaupt" (Heinze & Klusemann, 1980, 103), d. h. ein methodisches Prinzip, unter dem die gültigen Interpretationen eines Textes innerhalb der möglichen ausgezeichnet werden können.
6. Während alltagspraktische Verständnisbildung sich auf das Maß konventionell eingespielter Deutungen verlässt und dazu tendiert, mögliche Missverständnisse unkenntlich zu machen, indem sie das gegenseitige Verstehen als unhinterfragtes Axiom voraussetzt, muss diese internalisierte Intersubjektivität (Leithäuser & Volmerg, 1977) in der Textanalyse durch systematisches Misstrauen gegenüber den sich anbietenden Interpretationen, gegenüber den eigenen Voreinstellungen und Erwartungen etc., sowie gegenüber der Vollständigkeit des erfassten Kontextwissens systematisch durchbrochen werden.
7. Unter keinen Umständen kann mit zweifelsfreier Gewissheit und hinreichenden Gründen behauptet werden, dass eine Interpretation richtig ist (Mollenhauer & Rittelmeyer, 1977, 59). Denn da wir nie sicher sein können, *die* gültige Interpretation gefunden zu haben, und da wir nicht ausschließen können, dass es weitere Kontext-Informationen geben könnte, die uns zu einer Änderung unserer Deutung veranlassen müssten (Heinze & Klusemann, 1980, 104), hat das Ergebnis der Textanalyse immer nur den Status einer Hypothese.
8. Damit diese Hypothese überprüfbar ist, ist ein durchsichtiges Analyseverfahren erforderlich, so dass die einzelnen Analyseschritte nachvollziehbar und die verwendeten Termini in ihrer Bedeutung bestimmt und eindeutig auf die einzelnen Textbestandteile beziehbar sind.

Der nicht hintergehbare Vorbegriff

Entgegen der verbreiteten Auffassung, wonach die qualitative Textanalyse "den Inhalt selbst sprechen" lässt und aus ihm heraus die Analyse entfaltet (Lamnek, 1989, 195), ist in Rechnung zu stellen, dass die Analyse immer bereits einen Vorbegriff des

Analysegegenstandes voraussetzt, und zwar in dreierlei Hinsicht (vgl. Mollenhauer & Rittelmeyer, 1977, 59):

- als *Verstehbarkeit des semantischen Gehalts* (der Bedeutung) der verbalen (und ggf. nonverbalen) Zeichen (Worte, Sätze, Gesten) der (Umgangs-)Sprache, die wir mit dem Text teilen;
- als *Klassifikation des Textes* (autobiographische Erzählung, Rollenspielsituation, Gruppendiskussion etc.), der analysiert werden soll und damit als ein Gegenstand mit einer bestimmten, alltagspraktisch eingespielten gesellschaftlichen Bedeutung eingeordnet wird;
- als *theoretisch reflektierter Vorbegriff*, der eine aus der wissenschaftlichen Arbeit gewonnene Annahme darüber enthält, welche Art von Hypothesen am Text entwickelt und/oder überprüft werden kann.

Die Methode der sozialpsychologischen Rekonstruktion

Wie die o.g. Prinzipien und Leitgedanken praktisch umgesetzt werden können, welche Einschränkungen und Variationen sich dabei aufgrund der Art des analysierten Textes und der gewählten Analysemethode – Textinterpretation, qualitative Inhaltsanalyse oder quantitative Inhaltsanalyse – ergeben, wollen wir im Folgenden daher anhand einer Familie von textanalytischen Verfahren herausarbeiten und exemplifizieren, die mit der Kommunikationstheorie von Watzlawick et al. (1967; vgl. Bd. 1, Kap. 3.4.5) und der Konflikttheorie von Deutsch (1973) über einen gemeinsamen theoretischen Vorbegriff verfügen und in der Literatur als sozialpsychologische Rekonstruktion bezeichnet werden.

Kontrollfragen

- Was versteht man unter dem Begriff der Triangulation und welche Formen der Triangulation lassen sich unterscheiden?
- Wie unterscheiden sich quantitative und qualitative Sozialforschung hinsichtlich ihres Gegenstandsverständnisses, ihres Wissensideals und ihres Empirieverständnisses?
- Was versteht man unter dem Primat der Empirie vor der Theorie und welche Konsequenzen ergeben sich daraus für die interpretative Forschung?
- Was versteht man in der qualitativen Sozialforschung unter den Prinzipien der Offenheit und der Kommunikation?
- Nennen Sie einige qualitative Erhebungsmethoden!
- Welche Arten von Handlungsbeschreibungen lassen sich unterscheiden?
- Welcher Grundgedanke liegt dem narrativen Interview zugrunde?
- Wie unterscheidet sich das narrative Interview von Leitfadeninter-

views und welche Phasen des narrativen Interviews lassen sich unterscheiden?

- Welches sind die Gemeinsamkeiten und Unterschiede von narrativem Interview und partnerzentriertem Gespräch und wie sind diese begründet?
- Durch welche Elemente ist die personenzentrierte Grundhaltung nach Rogers definiert und wie werden diese im partnerzentrierten Gespräch umgesetzt?
- Diskutieren Sie einige der verbreiteten Kunstfehler der Gesprächsführung!
- Begründen Sie, warum Transkripte keine Eins-zu-Eins-Wiedergabe der transkrbierten Gespräche sind und welche Konsequenzen sich daraus für den Umgang mit Transkripten ergeben!
- Welche Formen der Verschriftlichung von Gesprächen lassen sich unterscheiden und welches sind ihre Vor- und Nachteile?
- Was versteht man in der qualitativen Sozialforschung unter den Prinzipien der Fremdheit und der Reflexion?
- Nennen Sie einige qualitative Analysemethoden!
- Wie unterscheiden sich quantitative und qualitative Forschungsmethoden hinsichtlich Objektivitätssicherung und Gewährleistung von Transsubjektivität der Forschungsergebnisse?
- Diskutieren Sie Möglichkeiten und Grenzen der kommunikativen Validierung der Analyseergebnisse!
- Welches sind die Leitgedanken der Textanalyse?
- Diskutieren Sie die Auffassung, wonach die qualitative Textanalyse "den Text selbst sprechen" lässt!

1.2
Grundlagen der sozialpsychologischen Rekonstruktion

Gegenstandsverständnis und Wissensideal

Seit den richtungsweisenden Arbeiten von Deutsch (1973) und Mitchell (1981) wurden in der Psychologie und in anderen Sozialwissenschaften eine Reihe von Theorien und Modellen des Konfliktmanagements (Fisher & Ury, 1984; Glasl, 1992) entwickelt, denen allen gemeinsam ist, dass sie zur Lösung von Konflikten nicht (nur) bei der Veränderung der objektseitig definierten Konfliktursachen (z. B. Ressourcenknappheit) ansetzen, sondern auf die Transformation der subjektseitig definierten Wahrnehmung und Interpretation des Konfliktes seitens der Konfliktparteien abzielen.

In methodologischer Hinsicht hat dies Konsequenzen, sowohl für das Gegenstandsverständnis als auch für das Wissensideal der Konfliktforschung, welche die Vermittlung objektseitig definierter Randbedingungen menschlichen Handelns in subjektseitig definierte Handlungsprämissen (Intentionen, Situationsbeurteilungen etc.) zu einem zentralen Forschungsgegenstand machen, dessen empirische Untersuchung einer textanalytischen Methodik bedarf, die auf unterschiedlichste Textsorten anwendbar ist: auf die Kommunikation *zwischen* den Konfliktparteien ebenso wie auf die Berichterstattung *über* Konflikte und auf Gruppendiskussionen, welche den *sozialen* Prozess des Aushandelns von Bedeutungen modellieren, ebenso wie auf Interviews, welche die je *individuelle* Bewältigung intrapersonaler Konflikte abbilden (Kempf et al., 2000).

Der Primat des Gegenstands vor der Methode

Ausgehend vom *Primat des Gegenstands vor der Methode* (Kempf, 1992a) wird dabei die Dichotomie zwischen qualitativen und quantitativen Methoden aufgehoben und deren Verhältnis zueinander in Hinblick auf die jeweils verschiedenen Forschungsziele, denen sie dienen, neu definiert. Entsprechend versteht sich die sozialpsychologische Rekonstruktion als eine Familie kommunikationsanalytischer Methoden, die durch ihre theoretische Kon-

zeption miteinander verbunden sind und ein breites Methodenspektrum abdecken, welches von textinterpretativen bis zu inhaltsanalytischen Verfahren reicht (Kempf & Baumgärtner, 1996) und eine gegenstandsangemessene Integration quantitativer und qualitativer Methoden (Kempf & Palmbach, 1994; Kempf, 2002) anstrebt.

Charakterisierung verschiedener textanalytischer Verfahrensweisen

Damit soll erreicht werden, dass die spezifischen Vorteile der verschiedenen textanalytischen Verfahrensweisen in optimaler Weise ausgenutzt und ihre Nachteile so weit wie möglich ausgeglichen werden (vgl. Tab. 1.2.1).

- Während inhaltsanalytischen Methoden deskriptiv und statisch sind, sind textinterpretative Verfahren prozessorientiert und dynamisch.
- Während Textinterpretation und qualitative Inhaltsanalyse (schon aus Gründen des Arbeitsaufwandes) auf die Untersuchung einiger weniger typischer Fälle abzielen, erlaubt die quantitative Inhaltsanalyse die Auswertung großer Datenmengen und repräsentativer Stichproben von Texten.

Dies wirkt sich auch auf die Art der Ergebnisdarstellung aus, die im Falle der quantitativen Inhaltsanalyse gleichsam *summarisch* ist, während sie im Falle der qualitativen Inhaltsanalyse *exemplarisch* und im Falle der Textinterpretation *narrativ* ist.

Tab. 1.2.1: Gegenüberstellung verschiedener textanalytischer Verfahrensweisen

	Textinterpretation	Inhaltsanalyse	
		Qualitativ	Quantitativ
Text-material	Einige wenige typische Texte		Repräsentative Textstichprobe
Art der Analyse	Prozessorientiert, dynamisch	Deskriptiv, statisch	
Fokussierte Perspektive	Gegenstands- (report aspect), sender- (command aspect) und empfängerbezogen (parade aspect)	Sender- (command aspect) und empfängerbezogen (parade aspect)	Gegenstandsbezogen (report aspect), nur teilweise auch senderbezogen (command aspect)
Ergebnis-darstellung	Narrativ	Exemplarisch	Summarisch

Ebenen der Verständnisbildung

Unabhängig davon, welcher methodische Zugang gewählt wird oder wie die verschiedenen Verfahrensweisen miteinander kom-

biniert werden, erfolgt die Textanalyse stets auf drei verschiedenen, systematisch aufeinander aufbauenden Ebenen der Verständnisbildung (vgl. Tab. 1.2.2), die aber je nach Art des Textes, der Zielsetzung der Analyse und der textanalytischen Verfahrensweise unterschiedlich stark fokussiert werden.

Tab. 1.2.2: Ebenen der Verständnisbildung

Ebene der Verständnisbildung	Analysierter Aspekt der Kommunikation und ihres Gegenstandes	Charakteristische Fragen	Fokussierte Sinnebene
Logisches Verstehen	Report aspect der Kommunikation	"Was wird mitgeteilt?" - "Worüber wird gesprochen?" - "Was wird darüber ausgesagt?"	Manifester Gehalt der Kommunikation
	Die Kommunikation als *Informationsfigur*	"Was will der Sender mitteilen?"	Interaktive Dynamik der Kommunikation
Psycho-logisches Verstehen	Command & parade aspect der Kommunikation	"Wie wird das mitgeteilt?" - "Wie wird miteinander gesprochen?" - "Welche Antwort wird dadurch herausgefordert?"	
	Die Kommunikation als *Interaktionsfigur*	"Welchen (impliziten) Regeln folgt die durch den Text repräsentierte Konstruktion subjektiver Wirklichkeit?"	Latenter Gehalt der Kommunikation
Sozio-logisches Verstehen	Aus der Kommunikation ausgeschlossene oder verzerrte Aspekte des Gegenstandes der Kommunikation	"Was wird dadurch nicht oder nur verzerrt angesprochen?"	Exterritorialisierter Gehalt der Kommunikation
	Die Kommunikation als *Abwehrfigur*	"Warum wird das nicht oder nur verzerrt thematisiert?"	Sekundärer Gewinn aus der Thematisierungsabwehr

Die Ebene des logischen Verstehens

Auf der Ebene des *logischen Verstehens* konzentriert sich die Analyse auf den propositionalen Aspekt (report aspect) der Kommunikation. Dabei interessiert primär der rein informative Charakter der Äußerungen im Text (der Text als *Informationsfigur*). Bei der Frage "Was will der Sender tatsächlich mitteilen?" handelt es sich um einen hermeneutischen Analyseschritt, der nach den Intentionen des Textproduzenten fragt. Diese Intentionen können letztlich nur aus der interaktiven Dynamik der Kommunikation erschlossen werden.

Die Ebene des psycho-logischen Verstehens

Der metakommunikative und pragmatische Aspekt sowie die interaktive Dynamik der Kommunikation (der Text als *Interaktionsfigur*) werden auf der Ebene des *psycho-logischen Verstehens* analysiert. Im Anschluß an Watzlawick et al. (1967) und Schulz von Thun (1981) werden die Selbstdarstellung der Produzenten (Selbstmanifestationsaspekt nach Schulz von Thun) und die Darstellung ihrer Beziehung zur Welt (Beziehungsaspekt, der sowohl bei Watzlawick et al. als auch bei Schulz von Thun allerdings auf die Beziehung zum Rezipienten eingeschränkt ist) als metakommunikativer Aspekt (command aspect) gefasst, der darüber entscheidet, wie die Sprachäußerung zu verstehen ist (report aspect) und welche Antwort auf Seiten des Rezipienten dadurch provoziert wird (parade aspect). In einem tiefenhermeneutischen Interpretationsschritt werden Aspekte des latenten Gehaltes der Kommunikation analysiert. Dabei geht es darum, herauszufinden, welchen (impliziten) Regeln die durch den Text repräsentierte Konstruktion subjektiver Wirklichkeit folgt.

Entfremdung

Den Kerngedanken des sozial-materialistischen Konstruktivismus (Keupp, 1991, 1992) folgend, wird davon ausgegangen, dass sich die Regeln zur Konstruktion der subjektiven Wirklichkeit seitens der Individuen keineswegs beliebig konstituieren, sondern in den sozialen Gegebenheiten selbst verankert sind. Diese impliziten Regeln stellen die psychodramatische Form des Textes dar, in der die Art und Weise zum Ausdruck kommt, wie die gesellschaftliche Entfremdungsproblematik durch diese Wirklichkeitskonstruktion bewältigt wird.

Der Terminus *Entfremdung* beschreibt den dialektischen Gegensatz zwischen Individuum und Gesellschaft, demzufolge die subjektive Logik des Handelns mit ihrer gesellschaftlichen Logik nicht deckungsgleich ist (Sève, 1978). Menschliches Handeln hat nicht nur die damit intendierten Wirkungen, sondern oft eine Vielzahl

an Nebenfolgen, die so weder gewollt noch vorhergesehen werden und vom Standpunkt des Subjektes aus oft nicht einmal erkannt werden (können). Schon die einfachsten zwischenmenschlichen Interaktionen haben daher nicht nur aufgrund der bestehenden Perspektivendivergenz voneinander abweichende (subjektseitig bestimmte) Bedeutungen für die Interaktionspartner, sondern zugleich eine (objektseitig bestimmte) Bedeutung für das System.

Der latente Gehalt der Kommunikation bestimmt somit die Perspektive, unter welcher die Subjekte manifeste Kommunikationsinhalte interpretieren, d. h. bestimmte Aspekte aufnehmen, während sie andere Elemente (Fakten, Erfahrungen, Bedeutungen etc.) thematisch abwehren.

Die Ebene des soziologischen Verstehens

Auf der Ebene des *sozio-logischen Verstehens* wird der exterritorialisierte Gehalt der Kommunikation untersucht (der Text als *Abwehrfigur*). Hier steht die Frage nach den Inhalten im Vordergrund, welche in der Kommunikation nicht oder nur verzerrt thematisiert werden. Dieses Exterritorialisieren von Bedeutungen zeigt sich z. B. deutlich in konfliktären Interaktionen, in welchen die Konfliktpartner den destruktiven Charakter der nicht-intendierten Wirkungen ihres Konfliktlöseverhaltens nicht erkennen (können) bzw. aus ihrem Bewusstsein ausblenden.

In der sozialpsychologischen Rekonstruktion wird diese Perspektivenverengung (sowie Thematisierungsabwehr generell) aber weder auf individuelle Eigenschaften der Konfliktparteien noch auf Verdrängung oder Abwehr im psychoanalytischen Sinne zurückgeführt, sondern vielmehr im Zusammenhang mit dem sozialen Charakter der Kommunikation selbst betrachtet und problematisiert. Über die Erfassung des exterritorialisierten Gehaltes der Kommunikation hinaus stellt sich damit die Frage nach den Gründen für das Weglassen oder für die verzerrte Darstellung bestimmter Inhalte der Kommunikation. Um diese Gründe aufschlüsseln zu können, wird der sekundäre Gewinn aufgezeigt, welchen die Kommunikationspartner aus der Thematisierungsabwehr beziehen, d. h. es wird untersucht, unter welchen subjektiven Prämissen die Subjekte das Ausblenden bestimmter Bedeutungen für sich selbst als funktional wahrnehmen.

Die Berücksichtigung externer Kontexte

Im Gegensatz zu psychoanalytisch orientierten textinterpretativen Verfahren, die über den Text selbst nicht hinausgehen (z. B. Volmerg, 1988), erweitert die sozialpsychologische Rekonstrukti-

on – ähnlich wie die kritische Diskursanalyse (van Dijk, 1985; Jäger, 1994) – ihre Interpretationsbasis auf externe Kontexte: Ziel ist die umfassende Rekonstruktion möglichst aller Kontexte, welche für das Verständnis eines Textes relevant sind. Dazu gehören u. a. die Sachverhalte, über die kommuniziert wird (report aspect), die Situation, in der kommuniziert wird (command aspect) und die (ggf. auch disparaten) Intentionen, mit welchen Sender und Empfänger miteinander kommunizieren (parade aspect). Dabei ist zu beachten, dass auch die externen Kontexte in Textform vorliegen und folglich interpretationsbedürftig sind.

Die Rekonstruktion externer Kontexte erfolgt daher ebenfalls auf allen drei Ebenen der Verständnisbildung, was für die Interpretation des Textmaterials auf psycho-logischer und sozio-logischer Ebene umso mehr erforderlich ist, als es hier ja um jene Bedeutungsaspekte der Kommunikation geht, die von den Subjekten thematisch abgewehrt werden und die daher (da aus der Sprache ausgeschlossen) für textinterne Analysen unzugänglich sind.

Die Erweiterung des hermeneutischen Zirkels

Unter den Kontexten, welche die sozialpsychologische Rekonstruktion berücksichtigt, kann somit bezüglich des internen Kontextes zwischen der einzelnen Textstelle (interner Mikrokontext) und dem Text als Ganzem (interner Makrokontext) sowie bezüglich des externen Kontextes zwischen der Situation, in der kommuniziert wird, und dem Gegenstand, über den kommuniziert wird, unterschieden werden, wobei jeder dieser Kontexte sowohl unter dem Aspekt seiner Darstellungsfunktion (report aspect) als auch seiner Ausdrucksfunktion (command aspect) und seiner Appellfunktion (parade aspect) analysiert wird.

Wie wir bereits in Bd. 1 (Kap. 3.4.5) ausgeführt haben, interpretieren sich diese Aspekte jedoch wechselseitig, so dass der o.g. hermeneutische Zirkel nicht nur einen ständigen Wechsel zwischen der einzelnen Textstelle und dem Text als Ganzem, sondern auch einen ständigen Wechsel zwischen der Fokussierung der Darstellungs-, Ausdrucks- und Appellfunktion erfordert – und zwar: sowohl des Textes als auch seiner externen Kontexte und sowohl der einzelnen Textstelle als auch des Textes als Ganzem.

Einschränkungen der Analysemethode

Wie weit dieser Idealfall einer sozialpsychologischen Rekonstruktion in der Praxis der Textanalyse auch tatsächlich umgesetzt werden kann, hängt allerdings entscheidend von der analysierten Textsorte, von der empirischen Zugänglichkeit externer Kontexte sowie von der Art der textanalytischen Verfahrensweise ab.

Während die quantitative Inhaltsanalyse (schon allein aufgrund des Erfordernisses, große Textmengen mit hinreichender Intercoder-Reliabilität zu kodieren) relativ an der Textoberfläche bleibt und den Analyseschwerpunkt auf die gegenstandsbezogene (report aspect) und (nur teilweise) auch auf die senderbezogene Perspektive (command aspect) legt, untersucht die qualitative Inhaltsanalyse die Texte schwerpunktmäßig unter senderbezogener (command aspect) und empfängerbezogener Perspektive (parade aspect), und nur in der textinterpretativen Methodik kommen alle drei Perspektiven – zumindest potentiell – etwa gleichgewichtig zum Tragen (vgl. Tab. 1.2.1).

Kontrollfragen

- Welche Arten von textanalytischen Verfahren lassen sich unterscheiden und wodurch sind diese charakterisiert?
- Welche Ebenen der Verständnisbildung unterscheidet man in der sozialpsychologischen Rekonstruktion und wie sind diese definiert?
- Was versteht man in der sozialpsychologischen Rekonstruktion unter dem latenten Gehalt eines Textes?
- Was versteht die sozialpsychologische Rekonstruktion unter Entfremdung?
- Inwiefern nimmt die sozialpsychologische Rekonstruktion eine Erweiterung des hermeneutischen Zirkels vor?
- Welcher Zusammenhang besteht zwischen der Art der analysierten Textsorte, der empirischen Zugänglichkeit externer Kontexte und der Art der textanalytischen Verfahrensweise?

1.3
Textinterpretation

1.3.1 Varianten der Textinterpretation

Handlungsanalyse, Interaktionsanalyse und Textanalyse i.e.S.

Welche der drei Perspektiven bei der Interpretation eines Textes fokussiert wird, hängt nicht zuletzt von der Art der analysierten Texte und von der Zielsetzung ab, die mit der Textinterpretation verfolgt wird. Bezüglich letzterer kann grob unterschieden werden zwischen:

- der *Handlungsanalyse*, welche die genetische Rekonstruktion der subjektiven Handlungsprämissen des Sprechers zum Ziel hat und den Text unter einer gegenstandsbezogenen Perspektive interpretiert;
- der *Interaktionsanalyse*, welche darauf gerichtet ist, wie das jeweilige Thema durch die Interaktion zwischen Sprecher und Angesprochenem modifiziert wird und die das Zusammenspiel senderbezogener und empfängerbezogener Perspektive fokussiert, und
- der *Textanalyse i.e.S.*, die auf das Bedeutungspotential abzielt, welches der Text für einen hypothetischen Empfänger entfaltet, und den Text unter einer empfängerbezogenen Perspektive interpretiert.

Nicht jede Textart ist gleichermaßen geeignet, unter den o.g. Zielsetzungen interpretiert zu werden, so dass eine wechselseitige Abhängigkeit zwischen den Erhebungsinstrumenten einerseits und den Zielsetzungen der Textinterpretation andererseits besteht (vgl. Tab. 1.3.1.1).

Das für die Handlungsanalyse benötigte Textmaterial kann am besten mittels narrativer Interviews oder partnerzentrierter Gespräche erhoben werden, die den Einfluss des Interviewers minimieren und dem Interviewten einen maximalen Spielraum bieten, die berichteten Ereignisse gemäß seiner eigenen Bedeutungshierarchie darzustellen.

Bei konfrontativen Interviewtechniken, in der Gruppendiskussion oder auch im Rollenspiel wird der Gesprächsverlauf dagegen viel stärker durch die Interaktion zwischen den Gesprächspartnern bestimmt, weshalb die daraus hervorgegangenen Texte einer Interaktionsanalyse bedürfen.

Öffentliche Texte, wie z. B. Politikerreden, Medienberichterstattung, Filme oder literarische Produkte etc., werden dagegen unter der Zielperspektive einer Textanalyse i.e.S. interpretiert.

	Handlungs-analyse	Interaktions-analyse	Textanalyse i.e.S.
Zielsetzung der Interpretation	Rekonstruktion der subjektiven Handlungsprämissen des Sprechers	Analyse der Modifikation des Themas durch die Interaktion zwischen Sprecher und Angesprochenem	Rekonstruktion des Bedeutungspotentials, welches der Text für einen hypothetischen Empfänger entfaltet
Perspektive der Interpretation	Gegenstands-bezogen	Sender- und empfängerbezogen	Empfänger-bezogen
Geeignete Erhebungs-instrumente und Textarten	Narratives Interview und partnerzentriertes Gespräch	Konfrontative Interviewtechniken, Gruppendiskussion und Rollenspiel	Öffentliche Texte (z. B. Politikerreden, Medienberichte, Filme oder literarische Texte)

Tab. 1.3.1.1: Abhängigkeit zwischen Textart und Ziel der Textinterpretation

Das Ergebnis der Textinterpretation

Das Ergebnis der Textinterpretation ist in jedem Fall nicht (nur) die Nacherzählung des Textes aus der Perspektive der 1. Person (= des Sprechers) und ggf. der 2. Person (= des Angesprochenen), sondern vielmehr eine Rekonstruktion des Textes und seiner Kontexte aus der Perspektive der 3. Person (= des Interpreten) vor dem Hintergrund seines Vorbegriffs des zu interpretierenden Textes und der darin dargestellten Sachverhalte (= externer Kontext) und der während des Interpretationsprozesses systematisch an den Text herangetragenen Fragerichtung, Theorien etc. als Interpretationsfolie.

Exemplifikation anhand eines Textbeispiels

Im Folgenden wollen wir die Vorgehensweise der Textinterpretation anhand einer Textanalyse i.e.S. erläutern, wofür wir auf ein Textbeispiel zurückgreifen, das weiter unten auch der Darstellung der qualitativen Inhaltsanalyse dienen wird. Anschließend werden wir auf einige Besonderheiten eingehen, welche Hand-

lungsanalyse und Interaktionsanalyse von der Textanalyse i.e.S. unterscheiden.

Das Textbeispiel

Das Textbeispiel, dessen wir uns bedienen, ist ein kurzer Kommentar, der am 6. November 1997 auf Seite 1 der südwestdeutschen Regionalzeitung *Südkurier* erschienen ist und die Benennung von Claude Trichet als Kandidat für die Präsidentschaft der Europäischen Zentralbank (EZB) zum Gegenstand hat:

Widersinnig

Der Vorstoß des französischen Präsidenten und seines Regierungschefs, einen eigenen Kandidaten für das Amt des Präsidenten der europäischen Zentralbank zu benennen, ist ebenso ungeschickt wie illusorisch. Für die Ernennung bedarf es der Einstimmigkeit. Das Veto der Holländer, die einen eigenen Kandidaten stellen, gilt aber als sicher.
Tatsächlich befürchtet Paris wieder einmal ein angeblich erdrückendes deutsches Übergewicht im währungspolitischen Bereich. Dem holländischen Kandidaten Wim Duisenberg wird unterstellt, nur der Schatten des deutschen Bundesbankpräsidenten Hans Tietmeyer zu sein – und so ein Werkzeug der Bundesbank.
Frankreich besitzt für seine Offensive gegen das imaginäre Übergewicht der Bundesbank nicht die geringsten Erfolgsaussichten. Es verhindert zwar die Wahl Duisenbergs, hat aber nicht die Gewißheit, damit einem gefügigeren Kandidaten den Weg zu ebnen. Außerdem dürfte Paris auf Schwierigkeiten stoßen, wenn es versucht, Landsleute an die Spitze anderer europäischer und internationaler Einrichtungen zu stellen. Nicht zuletzt gefährdet Frankreich seinen internationalen Ruf. Denn nationalegozentrischer Ehrgeiz und Eitelkeiten sind weniger denn je beliebt.

ALFRED FRISCH, PARIS

1.3.2 Exemplifikation der Vorgehensweise

Das Ablaufschema der Textinterpretation

Das in Abb. 1.3.2.1 dargestellte Ablaufschema der Textinterpretation dient in erster Linie dem Überblick und sollte nicht als starres Schema missverstanden werden, dem die Textinterpretation rigide zu folgen hat. In der Praxis seiner Anwendung kann es vorkommen, dass gelegentlich ein Interpretationsschritt übersprungen wird, oder dass man im Verlaufe der Interpretation zu einem früheren Schritt zurückkehren muss.

Explikation des Vorbegriffs

Am Anfang der Textinterpretation steht die Explikation des Vorbegriffs des Interpreten: Diese umfasst u. a. die Begründung für die Textauswahl, die Charakterisierung der Textart, das Vorwissen bezüglich des externen Kontexts sowie die Theorie, welche als Interpretationsfolie dient, und wird in einem kurzen Memo festgehalten. Mitunter ist die Interpretationsfolie zunächst noch

eher vage und – in Abweichung von dem o.g. Ablaufschema – an späterer Stelle genauer zu präzisieren. In unserem Fall könnte die Explikation des Vorbegriffs etwa so aussehen wie unter Abb. 1.3.2.1 beschrieben.

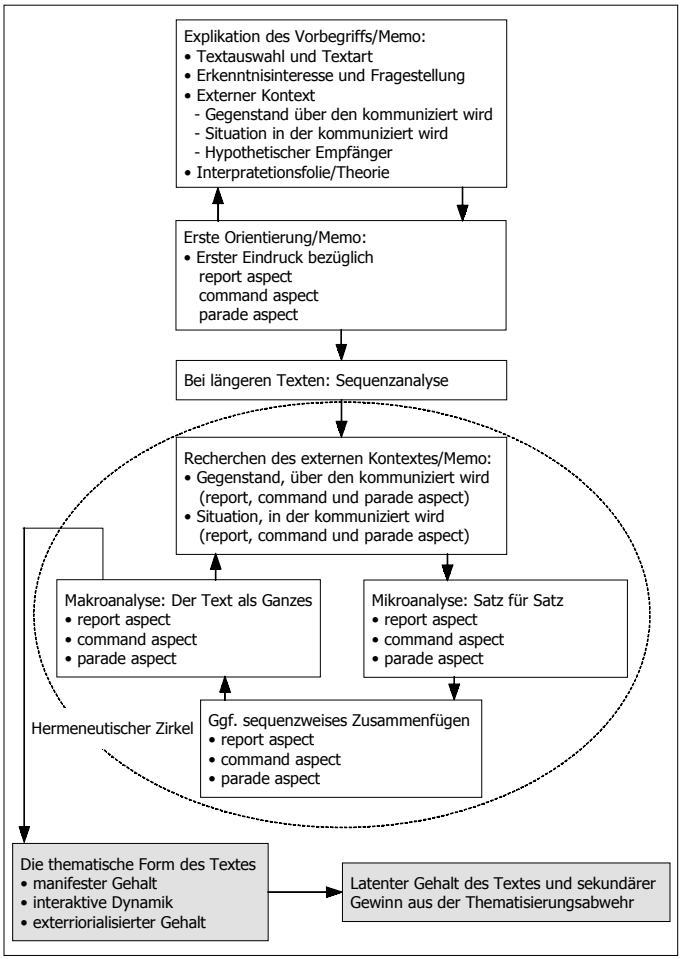


Abb. 1.3.2.1: Das Ablaufschema der Textinterpretation

Memo zur Explikation des Vorbegriffs

Textauswahl: Es handelt sich um einen Text, der einen politischen Konflikt zum Gegenstand hat und aufgrund seiner relativen Kürze für Übungszwecke geeignet ist.

Textart: Kommentar, Beispiel für Meinungsjournalismus.

Erkenntnisinteresse: Exemplarische Analyse der Dramatisierung von Konflikten in den Nachrichtenmedien.

Fragestellung: Wie wird der thematische Horizont der Konfliktwahrnehmung durch den Text gesteuert?

Gegenstand, über den kommuniziert wird: Bevorstehende Wahl des ersten Präsidenten der EZB. Hier: Die Benennung eines französischen Kandidaten.

Situation, in der kommuniziert wird: Deutsch-französische Freundschaft; bevorstehende Einführung des Euro.

Hypothetischer Empfänger: Leserschaft einer Regionalzeitung, die in der ehemals französisch besetzten Zone erscheint, in der viele deutsch-französische Städtepartnerschaften bestehen.

Interpretationsfolie: Durch ihren command und parade aspect entwickelt die Kommunikation eine interaktive Dynamik, welche die Kommunikationspartner selbst in den Gegenstand der Kommunikation verwickelt, der so zu einem Medium wird, in dem sie ihr eigenes Selbstverständnis und ihre Beziehung zueinander inszenieren. Dieser latente Gehalt der Kommunikation bestimmt damit zugleich die Perspektive, unter welcher sie den manifesten Gehalt der Kommunikation interpretieren, d.h. bestimmte Bedeutungen aufnehmen, während andere Bedeutungen (und sogar Tatsachen) ausgeblendet werden und aus dem Gesichtsfeld verschwinden.

Erste Orientierung

Als Einstieg in die eigentliche Interpretationsarbeit empfiehlt es sich, den Text aufmerksam laut vorzulesen – bei interaktiven Texten (z. B. Interviews, Gruppendiskussionen etc.) am besten mit verteilten Rollen. Anschließend wird der erste Eindruck, der beim Lesen des Textes hinsichtlich des report, command und parade aspects des Textes gewonnen wurde, in einem Memo festgehalten, das z. B. wie folgt aussehen könnte.

Memo zum ersten Eindruck

Report aspect: Benennung eines französischen Kandidaten für das Amt des EZB-Präsidenten.

Command aspect: Auffallend negativ konnotierter Wortgebrauch für die französche Seite, ihre Handlungen und Motive: u. a. Worte wie widersinnig, ungeschickt, illusorisch, wieder einmal, angeblich, unterstellen, imaginär, gefügiger Kandidat, nationalegozentrischer Ehrgeiz, Eitelkeiten.

Parade aspect: Stimulierung von Empörung über Frankreich, vielleicht sogar Einschwören des Lesers auf einen bevorstehenden Machtkampf.

Sequenzanalyse

Im Anschluss daran empfiehlt sich (insbesondere bei längeren Texten) eine Sequenzanalyse, die den Text in thematisch geschlossene Abschnitte einteilt. Zu diesem Zweck wird Satz für Satz notiert, wer die darin auftretenden Akteure (Personen, Institutionen etc.) sind, die (a) als Subjekt (Täter) und/oder (b) als Objekt (Opfer) der beschriebenen Ereignisse auftreten und durch welches Thema sie miteinander verbunden sind. Um den Text in Sequenzen einzuteilen, wird darauf geachtet, wie mehrere aufei-

nanderfolgende Sätze thematisch zusammenhängen bzw. wo das Thema wechselt (vgl. Tab. 1.3.2.1).

Der Themenwechsel kann, aber muss nicht am Ende eines Absatzes stattfinden. Mitunter finden sich Überleitungssätze, die sowohl dem vorangegangenen als auch dem Folgethema zugerechnet werden können, so dass die Sequenzeinteilung nicht ganz scharf, sondern fließend ist. Bezüglich der Frage, wie viele Sequenzen man bilden soll und wo genau die Grenze zwischen einer Sequenz und der nächsten zu ziehen ist, lassen sich keine verbindlichen Richtlinien angeben. Die Sequenzeinteilung ist nur heuristischer Natur und soll bei längeren Texten dazu dienen, gleichsam eine Zwischenstufe zwischen der Satz für Satz vorgehenden Mikroanalyse des Textes und seiner Makroanalyse zu bilden, welche das Bedeutungspotential des Textes als Ganzes betrifft. Bei kürzeren Texten kann man diesen Analyseschritt überspringen.

Gleichwohl vermitteltt die Sequenzanalyse auch bei kürzeren Texten bereits einen ersten Eindruck davon, wie der Text aufgebaut ist. So fällt z. B. bereits bei der Sequenzanalyse unseres Beispieltextes (vgl. Tab. 1.3.2.1) auf, dass die französische Seite durchgehend nur in der Rolle der Subjekt-Akteure auftritt, die deutsche Seite dagegen nur in Absatz 2, Satz 2, und dort auf der Objektseite. Die Entwicklung des Themas beginnt mit dem Herausstellen der Widersinnigkeit, weil Aussichtslosigkeit des französischen Unterfangens (1. Sequenz: Überschrift und Absatz 1), wendet sich sodann den haltlosen Motiven zu, die Frankreich dazu bewogen haben (2. Sequenz: Absatz 2), und verbreitet angesichts des zu erwartenden Bumerang-Effektes schließlich Siegeszuversicht (3. Sequenz: Absatz 3), um dann mit einem Ausweis der Urteilskompetenz des Autors zu enden (4. Sequenz: Unterschrift).

Recherche des externen Kontextes

Als nächster Analyseschritt erfolgt eine Recherche des externen Kontextes, die wieder in einem Memo festgehalten wird und die Hintergrundwissen über den Gegenstand zusammenträgt, über den kommuniziert wird, sowie über die Situation, in der kommuniziert wird. Beides wird sowohl hinsichtlich des report als auch des command und des parade aspects beleuchtet.

Tab. 1.3.2.1: Das Ergebnis der Sequenzanalyse des Beispieltextes

Abs.	Satz	Satzbeginn	Akteure	
			Als Subjekt	Als Objekt
Überschrift		Widersinnig		
1	1	Der Vorstoß ...	Französischer Präsident [französischer] Regierungschef	Eigener [französischer] Kandidat
	2	Für die Ernennung ...		
	3	Das Veto der Holländer ...	Die Holländer	Eigener [holländischer] Kandidat
2	1	Tatsächlich ...	Paris	
	2	Dem holländischen Kandidaten ...	[Paris]	Wim Duisenberg Hans Tietmeyer Bundesbank
3	1	Frankreich besitzt ...	Frankreich	Bundesbank
	2	Es verhindert zwar ...	[Frankreich]	Duisenberg Ein [für Frankreich] gefügigerer Kandidat
	3	Außerdem dürfte ...	Paris	Landsleute [Frankreichs] Europäische und internationale Einrichtungen
	4	Nicht zuletzt gefährdet ...	Frankreich	
	5	Denn nationalegozentrischer ...		
Unterschrift		Alfred Frisch, Paris	Alfred Frisch (Autor)	

Tab. 1.3.2.1: Das Ergebnis der Sequenzanalyse des Beispieltextes

Thema	Sequenz
Widersinn [von etwas]	Widersinnigkeit, weil Aussichtslosigkeit des französischen Unterfangens
Frankreichs Verhalten als Gegenstand der Widersinnigkeit	
Aussichtslosigkeit als Begründung für Widersinnigkeit	
Aussichtslosigkeit als Begründung für Widersinnigkeit	
Motive Frankreichs, die zu dem widersinnigen Verhalten geführt haben	Frankreichs haltlose Motive
Haltlosigkeit der Gründe für diese Motive	
Aussichtslosigkeit als Begründung für Widersinnigkeit	Zu erwartender Bumerangeffekt des französischen Unterfangens legt Siegeszuversicht nahe
Aussichtslosigkeit als Begründung für Widersinnigkeit	
Androhung negativer Konsequenzen	
Androhung negativer Konsequenzen	
Rechtfertigung dieser Konsequenzen durch Unterstellung verwerflicher Motive	
Begründung für die Kompetenz des Urteils	Ausweis für die Kompetenz des Urteils

Hierfür können wir zunächst von jenen Informationen ausgehen, auf welche der analysierte Text explizit verweist und/oder die in seiner unmittelbaren Nähe zu finden sind (z. B. andere Artikel in derselben Zeitungsausgabe, welche vom selben Thema handeln oder einen bestimmten Aspekt des Themas aufgreifen). Des Weiteren können wir auch auf andere Quellen zurückgreifen, wie z. B. auf die Berichterstattung in den überregionalen Qualitätszeitungen *Frankfurter Allgemeine* (FAZ), *Frankfurter Rundschau* (FR) und *Süddeutsche Zeitung* (SZ).

Im Falle unseres Beispieltextes könnte das Ergebnis dieser ersten Recherchen in etwa so aussehen:

Memo zum externen Kontext

Thematischer Bezug innerhalb derselben Zeitungsausgabe: Der Kommentar verweist auf einen Nachrichtenartikel (7 Absätze, 33 Spaltenzentimeter) auf der Wirtschaftsseite: "Streit um den EZB-Präsidenten. Frankreich bekräftigt Anspruch". Dort findet sich auch eine kurze Meldung (1 Absatz, 3 Spaltenzentimeter): "Euro wird Stabil".

Gegenstand, über den kommuniziert wird: Frankreich hat den Chef der Bank von Frankreich, Claude Trichet, als Gegenkandidaten zu dem von Deutschland favorisierten Kandidaten Wim Duisenberg nominiert und beruft sich dabei auf eine implizite Einigung, die im Zusammenhang mit der Einigung für Frankfurt als Sitz der EZB getroffen wurde. Die währungspolitische Linie, welche von Claude Trichet vetreten wird, unterscheidet sich nicht von jener der deutschen Bundesbank und des holländischen Kandidaten Wim Duisenberg. Frankreichs Präsident ist Jacques Chirac; Regierungschef ist Lionel Jospin. Deutscher Bundeskanzler ist Helmut Kohl. Deutschland hat seine Interessen bereits in einer Reihe von früheren Entscheidungen durchgesetzt, die im Zusammenhang mit der Euro-Einführung stehen (u. a. Maastricht-Vertrag, Sitz der EZB in Frankfurt/M.) (report aspect).
Für die deutsche Öffentlichkeit bedeutete die DM über Jahrzehnte hinweg ein starkes Identifikationsmoment. Die bevorstehende Euro-Einführung ist insofern nicht nur ein ökonomisches Thema. Die Schwäche anderer europäischer Währungen (einschließlich des französischen Franc) relativ zur starken DM hat in der Öffentlichkeit jedoch Befürchtungen geweckt, was passieren könnte, wenn die gemeinsame Währung nicht unter ähnlich strikter Kontrolle steht wie bisher die DM (command aspect).
Insofern besteht Handlungsbedarf in zweierlei Hinsicht: (1) Fortschreibung der monetären und ökonomischen Verhältnisse (die durch die Kosten der Wiedervereinigung ohnedies belastet sind) und (2) Verteidigung der darauf gegründeten Identitätsaspekte (parade aspect).

Situation, in der kommuniziert wird: Die deutsch-französischen Beziehungen sind vorgeblich gut. Helmut Kohl stellt sich immer wieder als Erbe Adenauers dar, des Vaters der deutsch-französischen Aussöhnung. Seit der Wiedervereinigung ist Deutschland aber nicht nur das wirtschaftsstärkste sondern nun auch das bevölkerungsreichste und größte Land Europas (report aspect).
Dies verschärft die Konkurrenz zwischen Deutschland und Frankreich

(command aspect) und führt dazu, dass Deutschland (unter dem Schlagwort der gestiegenen Verantwortung) zunehmend mehr Mitspracherechte auf dem internationalen Parkett beansprucht (parade aspect).

Nachdem wir erste Mikro- und Makroanalysen des internen Kontextes durchgeführt haben, werden sich sowohl weitere Fragen bezüglich des Gegenstandes ergeben, über den kommuniziert wird, als auch bezüglich der Situation, in der kommuniziert wird. Wir werden dann also zu Recherchen von externem Kontext zurückkehren, bevor wir uns erneut der Analyse des internen Kontextes zuwenden, usw.

Mikroanalyse

Die Analyse des internen Mikro-Kontextes zerlegt den Text Satz für Satz in die einzelnen Aussagen, welche darin enthalten sind (report aspect). Die erkenntnisleitenden Fragen, die an den Text gerichtet werden, lauten:

- "Was wird mitgeteilt?"
- "Worüber wird gesprochen?"
- "Was wird darüber ausgesagt?"

Gleichzeitig notieren wir ihren command aspect:

- "Wie wird das mitgeteilt?"
- "Wie wird miteinander gesprochen?"

Drittens notieren wir den parade aspect:

- "Welche Reaktion wird dadurch herausgefordert?"

Darüber hinaus halten wir fest, ob der Satz auf frühere Textstellen Bezug nimmt und dort gegebene Informationen (report aspect), Interpretationen (command aspect), oder Herausforderungen (parade aspect) wieder aufgreift. Gleichzeitig gehen wir an die frühere Stelle zurück und notieren dort, in welchem Satz die Information, Interpretation oder Herausforderung später wieder aufgegriffen wird.

Weiters notieren wir eventuelle Bezugnahmen auf externen Kontext sowie Fragen an diesen, die es erst noch zu recherchieren gilt, und schließlich halten wir fest, falls relevante Informationen durch die Form der Darstellung ausgeblendet oder verzerrt werden (vgl. Tab. 1.3.2.2).

Entscheidend für die Mikroanalyse ist es, dass wirklich jede mögliche Bedeutung einer Textstelle aufgelistet wird. Welche von ihnen dann tatsächlich vom gesamten Text eingelöst wird, zeigt sich später im Verlaufe der Makroanalyse, welche den Text als Ganzes betrachtet.

Tab. 1.3.2.2: Die Mikroanalyse des Textbeispiels
S = Sequenz-Nr.
A = Absatz-Nr.
S = Satz-Nr.

S	A	S	Text	
Report aspect			Command aspect	Parade aspect
1	Überschrift		Widersinnig	
[Etwas ist] widersinnig			Abwertende Kritik	Aufforderung zur Ablehnung
1	1	1	Der Vorstoß des französischen Präsidenten und Präsidenten der europäischen Zentralbank zu	
Der französische Präsident und der französische Regierungschef haben einen Vorstoß unternommen			"Vorstoß" klingt abrupt und nach Alleingang oder Konfrontation / auch positive Interpretation möglich: initiativ, Tatkraft ---------- "sein Regierungschef" klingt abwertend	Aufforderung zur Ablehnung oder zur Anerkennung
Inhalt des Vorstoßes ist die Benennung eines französischen Kandidaten für die Präsidentschaft der europäischen Zentralbank			Sachinformation, aber: der französische Kandidat wird im Gegensatz zu dem holländischen namentlich nicht genannt / ist dehumanisierend	Ablehnung oder zumindest Nicht-zur-Kenntnis-Nehmen des französischen Kandidaten
Der Vorstoß ist ungeschickt und illusorisch			Abwertende Kritik, "nicht ernst zu nehmen", klingt arrogant	Aufforderung zur Ablehnung bestätigt sich. Zugleich Aufforderung zur Beruhigung: "die werden schon sehen, wo sie damit hinkommen!"
1	1	2	Für die Ernennung bedarf es der Einstimmigkeit.	
Ernennung des EZB-Präsidenten bedarf der Einstimmigkeit			Sachinformation Wortspiel "*Be*nennung (1/1/1) vs. *Er*nennung (1/1/2)	Beruhigung durch "sachliche Begründung" verstärkt
1	1	3	Das Veto der Holländer, die einen eigenen Kan-	
Holland stellt einen Kandidaten Holland wird ein Veto einlegen			Das Eigeninteresse Hollands wird als legitimes Motiv dargestellt, aber nicht explizit als solches benannt	Die Beruhigung wird durch "sachliche Begründung" verstärkt

Tab. 1.3.2.2: Die Mikroanalyse des Textbeispiels

<=	=>	Externer Kontext	Ausgeschlossen oder verzerrt
	1/1/1		
seines Regierungschefs, einen eigenen Kandidaten für das Amt des benennen ist ebenso ungeschickt wie illusorisch.			
1/Ü	1/1/2		Legitimität der Benennung eines französischen Kandidaten und Vereinbarung von 1993, die Frankreich zumindest die Hoffnung machte, es könnte den ersten EZB-Präsidenten stellen (vgl. FAZ, FR, SZ vom selben Tag)
	2/2/2	Demokratischer EU-Prozess? / Ist das Verhalten Frankreichs eine Regelverletzung oder nicht? / Was ist die Bedeutung des Amtes?	
	3/3/2		
1/1/1	3/3/2		
didaten stellen, gilt aber als sicher.			
	3/3/3 3/3/5		Holländisches Interesse wird vorgeschoben, deutsches Interesse wird dagegen *nicht* thematisiert

Fortsetzung
Tab. 1.3.2.2: Die Mikroanalyse des Textbeispiels
S = Sequenz-Nr.
A = Absatz-Nr.
S = Satz-Nr.

S	A	S	Text	
Report aspect			Command aspect	Parade aspect
2	2	1	Tatsächlich befürchtet Paris wieder einmal ein Bereich.	
Paris befürchtet deutsches Übergewicht im währungspolitischen Bereich			Klingt abwertend, delegitimierend: "wieder einmal" "angeblich erdrückendes" Übergewicht	Zurückweisung der Befürchtungen Frankreichs als "paranoid" (wieder einmal), "irreal" (angeblich) und "bösartig" (wir Deutschen erdrücken doch niemanden)
2	2	2	Dem holländischen Kandidaten Wim Duisenberg Hans Tietmeyer zu sein – und so ein Werkzeug	
[Paris] unterstellt dem holländischen Kandidaten etwas			Negative Konnotation: Paris "unterstellt" etwas / verweist auf Ungerechtigkeit	Zurückweisung solcher Unterstellungen
			Der, dem Paris etwas unterstellt, erscheint in der Passivform / versetzt ihn in die Rolle des Opfers, dem etwas angetan wird	Solidarisierung mit dem Opfer, das im Unterschied zu dem französischen Kandidaten namentlich genannt und damit humanisiert wird
Inhalt der Unterstellung ist, dass der holländische Kandidat nur ein Schatten des Bundesbankpräsidenten und damit ein Werkzeug der Bundesbank sei			Negative Konnotation: "nur Schatten", "Werkzeug"	Die Unterstellung ist auch noch frech und daher erst recht zurückzuweisen
3	3	1	Frankreich besitzt für seine Offensive gegen das sichten.	
Frankreich führt eine Offensive durch			Negative Konnotation "Offensive" / Frankreich erscheint als Aggressor	Verurteilung Frankreichs, weil dessen Verhalten aggressiv ist, auf haltlosen Unterstellungen beruht und gegen Deutschland gerichtet ist
Die Offensive richtet sich gegen das imaginäre Übergewicht der Bundesbank			Negative Konnotation des Wortes "imaginär": Frankreichs Motive werden als haltlos und ungerecht gegen Deutschland bzw. die Bundesbank disqualifiziert	

Fortsetzung
Tab. 1.3.2.2: Die Mikroanalyse des Textbeispiels

<=	=>	Externer Kontext	Ausgeschlossen oder verzerrt
angeblich erdrückendes deutsches Übergewicht im währungspolitischen			
	3/3/1	Bisheriger Verlauf der Entscheidungen im Zusammenhang mit der Euro-Einführung	Ausgeschlossen: all die Entscheidungen im Zusammenhang mit der Euro-Einführung, bei denen sich die deutschen Interessen bereits durchgesetzt haben (vgl. FAZ, FR, SZ vom selben Tag)
wird unterstellt, nur der Schatten des deutschen Bundesbankpräsidenten der Bundesbank.			
1/1/1		Beide Kandidaten verfolgen eine ähnliche währungspolitische Linie wie deutsche Bundesbank auch (vgl. FAZ, FR und SZ vom selben Tag)	Die Kompetenz des französischen Gegenkandidaten, (vgl. FAZ, FR und SZ vom selben Tag) schwindet aus dem Gesichtsfeld; es geht darum, das Opfer zu verteidigen
imaginäre Übergewicht der Bundesbank nicht die geringsten Erfolgsaus-			
2/2/1			Deutsche Interessen und tatsächlicher Einfluss Deutschlands in währungspolitischen Fragen

Fortsetzung
Tab. 1.3.2.2: Die Mikroanalyse des Textbeispiels
S = Sequenz-Nr.
A = Absatz-Nr.
S = Satz-Nr.

S	A	S	Text	
Report aspect			Command aspect	Parade aspect
Die Offensive hat nicht die geringsten Erfolgsaussichten				Beruhigung: "die werden damit nicht durchkommen"
3	3	2	Es verhindert zwar die Wahl Duisenbergs, hat zu ebnen.	
Frankreich wird die Wahl Duisenbergs verhindern aber am Ende leer ausgehen			Lose-Lose-Modell	Genugtuung darüber, dass Frankreich auch nicht bekommen wird, was es will
			Gewissheit: "Es verhindert" vs. Ungewissheit, ob die eigentlichen Ziele damit erreicht werden können	Konfrontative Haltung: "die werden schon sehen, wo sie damit hinkommen"
Was Frankreich will, ist ein "gefügigerer Kandidat"			Negative Konnotation "gefügig"	Verurteilung von Frankreichs Motiven
3	3	3	Außerdem dürfte Paris auf Schwierigkeiten stoßen, internationaler Einrichtungen zu stellen.	
Es ist damit zu rechnen, dass Paris auf Schwierigkeiten stoßen wird			"dürfte" = Drohung & zugleich Ausdruck von Gewissheit	Zuversicht: "denen werden wir's zeigen"
Französische Kandidaten für andere europäische Spitzenpositionen werden nicht ohne weiteres akzeptiert werden			Konkrete Drohung	Zuversicht: "wir wissen schon, wie wir es ihnen heimzahlen können"
Paris will seine Landsleute in europäische Spitzenpositionen bringen			Abwertende Kritik: bloßes Eigeninteresse Frankreichs	Dem französischen Egoismus Paroli zu bieten

Fortsetzung
Tab. 1.3.2.2: Die Mikroanalyse des Textbeispiels

<=	=>	Externer Kontext	Ausgeschlossen oder verzerrt
aber nicht die Gewissheit, damit einem gefügigeren Kandidaten den Weg			
1/1/1			Die Möglichkeit, einer gütlichen Einigung im beiderseitigen Interesse
1/1/2		Frankreich kann die Wahl verhindern (wegen erforderlicher Einstimmigkeit)	Frankreich könnte aber auch darauf verzichten, von seinem Vetorecht Gebrauch zu machen
wenn es versucht, Landsleute an die Spitze anderer europäischer und			
1/1/3			Legitimität von Eigeninteressen (die den Holländern durchaus zugestanden wird, während die deutschen unerwähnt bleiben) wird Frankreich abgesprochen

Fortsetzung
Tab. 1.3.2.2: Die Mikroanalyse des Textbeispiels
S = Sequenz-Nr.
A = Absatz-Nr.
S = Satz-Nr.

S	A	S	Text	
Report aspect			Command aspect	Parade aspect
3	3	4	Nicht zuletzt gefährdet Frankreich seinen inter-	
Frankreich gefährdet seinen internationalen Ruf			"Nicht zuletzt": Frankreich setzt sehr viel aufs Spiel	Zuversicht: Frankreich steht allein und der Rest Europas hinter uns / so viel wird Frankreich wohl doch nicht riskieren und vielleicht noch einlenken
3	3	5	Denn nationalegozentrischer Ehrgeiz und Eitel-	
Nationalegozentrischer Ehrgeiz und Eitelkeiten sind weniger denn je beliebt			Unterstellung unlauterer Motive	Selbstgerechtigkeit
4	Unterschrift		Alfred Frisch, Paris	
Autor schreibt aus Paris				Kompetenzzuschreibung verschafft dem Kommentar Glaubwürdigkeit und Autorität

Makroanalyse

Bei der Makroanalyse gehen wir sequenzweise vor und achten insbesondere auf jene Textstellen, die durch Querverweise mit anderen Sequenzen verbunden sind. Manche Details werden sich als wichtiger, andere als weniger wichtig erweisen, weil sie im Gesamtkontext des Textes keine Rolle spielen, und von wieder anderen Interpretationen mag sich zeigen, dass sie vorschnell waren und im Lichte des Gesamtkontextes nicht aufrechterhalten werden können.

So kann z. B. das Wort Vorstoß in Absatz 1, Satz 1 unseres Textbeispiels entweder auf Initiative und Tatkraft verweisen (positive Konnotation) oder auf Alleingang und Konfrontation (negative Konnotation). Während letztgenannte Interpretation durch eine Vielzahl von anderen Textstellen gestützt wird, welche die Benennung eines französischen Kandidaten als widersinnig, ungeschickt, illusorisch und nationalegozentrisch abqualifizieren, ist dies für die positive Interpretation des Wortes nicht der Fall. Diese Bedeutung wird durch den Rest des Textes nicht gestützt und kann daher ausgeschlossen werden.

Fortsetzung
Tab. 1.3.2.2: Die Mikroanalyse des Textbeispiels

<=	=>	Externer Kontext	Ausgeschlossen oder verzerrt
nationalen Ruf.			
keiten sind weniger denn je beliebt.			
1/1/3			Eigene Interessen Deutschlands

Der hermeneutische Zirkel

Während die Mikroanalyse Hypothesen darüber generiert, wie die einzelnen Textstellen (möglicherweise) zu interpretieren sind, liegt der Makroanalyse zunächst eine hypothesentestende Vorgehensweise zugrunde, aufgrund derer manche der möglichen Bedeutungen gestützt und andere falsifiziert werden können. Am Ende der Makroanalyse steht dann eine Hypothese darüber, wie der Text als Ganzer zu verstehen ist.

Diese Hypothese kann erneut Fragen an den externen Kontext aufwerfen (z. B. ob es denn tatsächlich eine Vereinbarung gegeben hat, aufgrund derer sich Frankreich Hoffnungen darauf machen konnte, den ersten Präsidenten der EZB zu stellen) und die Antwort darauf kann wiederum die Bedeutung einzelner Textstellen tangieren und dazu führen, dass bestimmte Interpretationen zurückgewiesen oder modifiziert werden müssen usw.

Dieser hermeneutische Zirkel, der von der Recherche des externen Kontextes über die Mikroanalyse des internen Kontextes zu seiner Makroanalyse und wieder zurück zur Recherche externen Kontextes führt, wird so lange fortgesetzt, bis ein in sich konsistentes Textverständnis erreicht ist.

Die thematische Form des Textes

Als Ergebnis der Analyse erhalten wir schließlich eine detaillierte Beschreibung der *thematischen Form* des Textes mit all seinen Nebenargumentationslinien sowie jener Tatbestände, die durch sie *verzerrt oder unkenntlich gemacht* werden. Das heißt: Wir erhalten eine Darstellung der (subjektiven) Wirklichkeit, welche durch den Text konstruiert wird, einschließlich des externen Kontextes, in dem der Text zu verstehen ist, und der Überredungstechniken, derer sich der Autor bedient, um dieser Konstruktion der Wirklichkeit zu Glaubwürdigkeit zu verhelfen. Im Falle unseres Textbeispiels könnte diese Darstellung etwa so aussehen:

Die thematische Form des Beispieltextes

Bei dem analysierten Text handelt es sich um einen namentlich gezeichneten Kommentar (5 Absätze, 12 Spaltenzentimeter) auf Seite 1 des *Südkurier* vom 6. November 1997. Dieser kommentiert die Nominierung eines französischen Kandidaten (Claude Trichet, Präsident der Bank von Frankreich) für das Amt des Präsidenten der EZB durch Frankreichs Präsident Jacques Chirac und Ministerpräsident Lionel Jospin, die damit einen Gegenkandidaten gegen den von der deutschen Regierung favorisierten holländischen Kandidaten Wim Duisenberg ins Spiel bringen.
Thematischer Bezug besteht zu einem Nachrichtenartikel auf der Wirtschaftsseite "Streit um den EZB-Präsidenten: Frankreich bekräftigt Anspruch" (7 Absätze, 33 Spaltenzentimeter) sowie zu einer kleinen Meldung "Euro wird stabil" (1 Absatz, 3 Spaltenzentimeter).
Schon beim ersten Lesen fällt der feindselige Unterton des Kommentars ins Auge. Auffallend ist dieser Ton nicht zuletzt auch vor dem Hintergrund der intakten deutsch-französischen Beziehungen, zumal mit Helmut Kohl ein deutscher Bundeskanzler regiert, der sich gern als Erbe Adenauers darstellt und die Achse Paris-Bonn auf seine Fahnen geschrieben hat.
Mit der deutschen Wiedervereinigung (1990) hat sich das Kräfteverhältnis innerhalb Europas allerdings verschoben. War die frühere Bundesrepublik bereits die wirtschaftsstärkste Nation Europas gewesen, so ist Deutschland durch die Wiedervereinigung auch zum bevölkerungsreichsten und größten Land Europas aufgestiegen, das daraus eine gewachsene internationale Verantwortung ableitet, die im Anspruch auf einen ständigen Sitz im UN-Sicherheitsrat ebenso zum Ausdruck kommt wie im Schulterschluss mit den USA, den US-Präsident Bush mit seiner Definition der amerikanisch-deutschen Beziehungen als "partners in leadership" besiegelt hat.
Das gewachsene Gewicht Deutschlands ist in Verbindung mit dem von Deutschland forcierten europäischen Einigungsprozess durchaus geeignet, bei den europäischen Partnern Befürchtungen hinsichtlich einer zukünftigen deutschen Vormachtstellung auszulösen.
Dem gewachsenen nationalen Selbstbewusstsein Deutschlands stehen im ökonomischen Bereich Irritationen und eine gewisse Verunsicherung der Bevölkerung gegenüber: Die Kosten der Wiedervereinigung, die sich eben nicht aus der Portokasse finanzieren ließen, die explosionsartige Steigerung der Arbeitslosenrate seit 1990, die neo-liberale Entwicklung nach dem Wegfall der Systemkonkurrenz mit Ende des kalten Krieges, die

daraus resultierende Infragestellung sowohl des Systems der sozialen Sicherung als auch des Wirtschaftsstandortes Deutschland rührt an grundlegende Aspekte (bundes-)deutscher Nachkriegsidentität, deren Symbol die Stabilität der DM war.
Mit der bevorstehenden Einführung der gemeinsamen europäischen Währung zum 1. Januar 2002 sind daher in der deutschen Bevölkerung nicht nur ökonomische Ängste verbunden. Der Wunsch nach einer strikten Kontrolle der europäischen Währung nach dem Vorbild der deutschen Bundesbank ist nicht nur Ausdruck des Bedürfnisses nach ökonomischer Sicherheit, sondern rührt an die Wurzeln (bundes-)deutschen Selbstverständnisses.
Diesem doppelten Bedürfnis nach ökonomischer Sicherheit und Wahrung idenstitätsstiftender Momente wurde in den vorangegangenen Jahren entgegengekommen, indem eine Vielzahl an währungspolitischen (Maastricht-Vertrag, Stabilitätskriterien für den Euro-Beitritt) und symbolischen (Prestige-)Entscheidungen (Namensgebung für die europäische Währung, Sitz der europäischen Zentralbank in Frankfurt/Main) im Sinne Deutschlands getroffen wurden.
Auch die bevorstehende Ernennung des Präsidenten der EZB hat sowohl eine sachliche als auch eine Prestige-Dimension: In sachlicher Hinsicht kann davon ausgegangen werden, dass Claude Trichet ein kompetenter Bewerber ist, der zudem dieselbe währungspolitische Linie vertritt wie Wim Duisenberg und der Präsident der deutschen Bundesbank Hans Tietmeyer. Bezüglich der Prestige-Fragen kann der von Frankreich erhobene Anspruch durchaus als legitim angesehen werden: nicht nur wegen der vielen Entscheidungen, die im Vorfeld bereits im Sinne der deutschen Interessen getroffen wurden, so dass jetzt auch einmal Frankreich zum Zuge kommen könnte, sondern auch in Hinblick auf die Abmachungen zwischen Kohl und Chirac im Zusammenhang mit der Verlegung des Sitzes der EZB nach Frankfurt/Main.
Diese Abmachungen waren aber offensichtlich nicht sehr eindeutig: Während sie von deutscher Seite lediglich als Verzicht darauf interpretiert wurden, einen Deutschen Bewerber an die Spitze der EZB zu setzen, konnte sich Frankreich Hoffnung darauf machen, dass das prestigeträchtige Amt des EZB-Präsidenten mit einem französischen Kandidaten besetzt würde.
In dieser Situation benennen nun Chirac und Jospin den Präsidenten der Bank von Frankreich als Gegenkandidaten des von der deutschen Bundesbank und der deutschen Bundesregierung favorisierten Holländers Wim Duisenberg. Dieses Ereignis ist Gegenstand des Kommentars.
Sein Autor, Alfred Frisch, berichtet direkt aus Paris, sozusagen "vor Ort" (1/U/1). Dies verleiht ihm den Nimbus besonderer Kompetenz und verspricht, dass der (überraschende) Schritt Frankreichs dem (südwest-) deutschen Leser aus dem französischen Kontext heraus verständlich gemacht wird.
Eben dies findet jedoch nicht statt. Der Kommentar entbehrt jeglicher Sachinformationen, welche die Hintergründe des deutsch-französischen Konfliktes um den Posten des EZB-Präsidenten aufhellen könnten. Solche Informationen werden auch in dem Bericht auf der Wirtschaftsseite des *Südkurier* nicht gegeben, so dass der unerwartete Schritt Frankreichs

schon von daher als widersinnig erscheinen muss: Die französische Regierung beschwört ohne nachvollziehbaren Grund einen Konflikt herauf. Dieses Urteil wird dem Kommentar bereits als Überschrift (1/Ü/1) vorangestellt; gleichsam als Appell an den Leser, sich dem kompetenten Urteil des Pariser Korrespondenten anzuschließen. Der Kommentar selbst dient dann nur noch der Begründung dieses Urteils, wofür im Wesentlichen von den folgenden Stilelementen Gebrauch gemacht wird:

- Bis auf eine einzige Ausnahme (1/1/3: Holland) treten ausschließlich französische Subjekt-Akteure auf. Alle anderen Akteure sind (fast) ausschließlich auf der Objektseite zu finden. Lediglich Holland (1/1/3) tritt einmal (defensiv) in Aktion. Schon dies erzeugt das Bild, dass der Konflikt einseitig von Frankreich ausgeht.
- Die französischen Akteure werden tendenziell dehumanisiert. Im Unterschied zu holländischen und deutschen Akteuren (2/2/2, 3/3/2) wird keiner der französischen je namentlich genannt, und die Formulierung "der französische Präsident und sein Regierungschef" (1/1/1) degradiert Lionel Jospin zum bloßen Handlanger von Jacques Chirac.
- Frankreichs Handlungen und Motive werden zum Gegenstand abwertender Kritik gemacht, die zur Ablehnung herausfordert und von den Sachfragen ablenkt.
- Schon im ersten Satz des Artikels erscheint die Benennung eines eigenen Kandidaten als konfrontativer "Vorstoß", der im Alleingang unternommen (1/1/1) und weiter unten zu einer "Offensive" gegen das Übergewicht der Bundesbank gesteigert wird (3/3/1). Die Frage nach der Legitimität des Schrittes Frankreichs kann so gar nicht erst aufkommen. Frankreichs Verhalten ist aggressiv und daher zu verurteilen, was auch immer seine Motive sein mögen.
- Diese Motive sind "Eitelkeit" (3/3/5) und "nationalegozentrischer Ehrgeiz" (3/3/5). Frankreich wird vorgeworfen, dass es "wieder einmal" ein "angeblich erdrückendes" deutsches Übergewicht (2/2/1) befürchte, und die Auseinandersetzung mit (tatsächlichen oder vermeintlichen) Empfindlichkeiten Frankreichs wird abgewehrt, indem diese als paranoid ("wieder einmal"), realitätsfremd ("angeblich") und bösartig (Wir Deutschen erdrücken doch niemanden; W. K.) denunziert werden. Weder die Frage nach dem tatsächlichen Gewicht Deutschlands noch die Frage nach Frankreichs Vorleistungen (s. o.) kann sich aus dieser Sichtweise heraus ergeben.
- Stattdessen wird Frankreich vorgeworfen, Wim Duisenberg zu unterstellen "nur der Schatten des deutschen Bundesbankpräsidenten zu sein – und so ein Werkzeug der deutschen Bundesbank" (2/2/2). Dieser Vorwurf fordert zur Solidarisierung mit Wim Duisenberg auf, der namentlich genannt und damit als greifbare Person humanisiert wird, während sein französischer Gegenkandidat anonym bleibt (1/1/1) und schließlich als Kandidat, der Paris "gefügig" ist (3/3/2), in genau der Weise abgewertet wird, in der Frankreich (angeblich) Wim Duisenberg disqualifiziert. Seine Kompetenz wird ebensowenig zum Thema wie die währungspolitische Linie, die er verfolgt, und die durchaus mit der Wim Duisenbergs und der Bundesbank vergleichbar ist. Rein sachlich gesehen stellt Claude Trichet eine vertretbare Alternative zu Wim Duisenberg dar, doch davon wird abgelenkt.

- Die Motive Frankreichs werden als haltlos und als Angriff gegen die Bundesbank denunziert: Frankreich will "einem gefügigeren Kandidaten den Weg ebnen" (3/3/2). Es führt eine "Offensive gegen das imaginäre Übergewicht der Bundesbank" (3/3/1), womit die Konfliktlinie durchaus zwischen Frankreich und Deutschland gezeichnet wird, wenngleich Deutschland als Akteur nicht in Erscheinung tritt, die deutschen Interessen nicht thematisiert werden und der Einfluss Deutschlands im währungspolitischen Bereich verleugnet wird.
- Die deutsche Politik und die deutschen Interessen, welche durch den Konflikt berührt werden, finden mit keinem Wort Erwähnung, lediglich die Bundesbank (als unabhängige und überpolitische Institution) wird als Gegenstand (ungerechtfertigter) Angriffe Frankreichs angesprochen. Dies legt die Grundlage, auf welcher der Artikel schließlich in einen Appell an Selbstgerechtigkeit mündet, die Frankreich unlautere Motive unterstellt und schon den bloßen Gedanken an eine Involvierung deutschen Eigeninteresses zurückweist: "Nationalegozentrischer Ehrgeiz und Eitelkeiten sind weniger denn je beliebt" (3/3/5).
- Während Frankreich ein Eigeninteresse vorgeworfen und das deutsche Eigeninteresse verleugnet wird, wird das holländische Eigeninteresse als ein legitimer Grund dafür ins Spiel gebracht, dass die französische Initiative scheitern wird: "Das Veto der Holländer, die einen eigenen Kandidaten stellen, gilt aber als sicher" (1/1/3).

Die Argumentationslinie des Artikels ist darauf ausgerichtet, gegen die Nominierung eines französischen Kandidaten Stimmung zu machen (1/Ü, 1/1/1, 2/2/1, 2/2/2; 3/3/1; 3/3/2) und zugleich die Öffentlichkeit zu beruhigen, dass dem Schritt Frankreichs kein Erfolg beschieden sein wird.

- Schon in der Überschrift und im ersten Satz wird die Nominierung Claude Trichets als "illusorisch" bezeichnet (1/Ü; 1/1/1) und
- gleich darauf die (scheinbare) Aussichtslosigkeit des französischen Unterfangens unter Verweis auf die Einstimmigkeit von EU-Entscheidungen (1/1/2) und das zu erwartende Veto der Holländer (1/1/3) vor Augen geführt: "Benennung" eines Kandidaten (1/1/1) bedeutet noch lange nicht dessen "Ernennung" (1/1/2) und dieser steht das Veto entgegen (1/1/3).
- Frankreich besitzt "nicht die geringsten Erfolgsaussichten" (3/3/1), seinen Kandidaten durchzubringen und kann allenfalls die Wahl Duisenbergs verhindern (3/3/2).

Indem der Autor des Artikels keinen Zweifel daran lässt, dass Frankreich dies auch tun wird, konstruiert er den Konflikt nach dem Muster eines Lose-Lose-Modells, in dem es nicht mehr darum geht, selbst zu gewinnen, sondern nur noch darum, dass der Gegner nicht gewinnen darf: "Es (Frankreich, W.K.) verhindert zwar die Wahl Duisenbergs, hat aber nicht die Gewissheit, damit einem gefügigeren Kandidaten den Weg zu ebnen" (3/3/2).

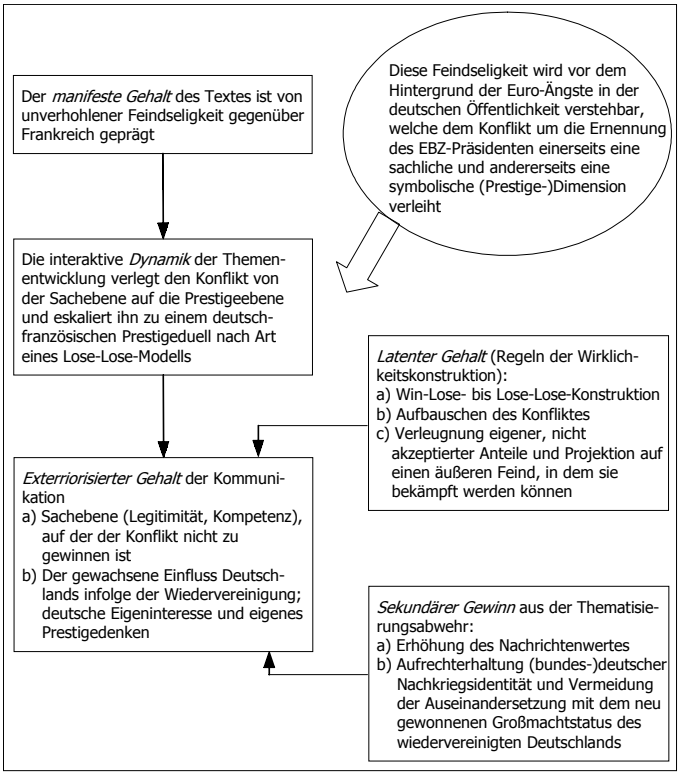
Im Rahmen dieses Lose-Lose-Modells entwickelt der Autor schließlich, was er gemeint hat, als er bereits einleitend den "Vorstoß" Frankreichs als "ungeschickt" bezeichnet hat (1/1/1): Der ungewissen Erfolgsaussicht Frankreichs stehen härteste Konsequenzen entgegen, die Frankreich nun unverhohlen angedroht werden:

- "Paris dürfte auf Schwierigkeiten stoßen, wenn es (künftig; W. K.) versucht, Landsleute an die Spitze anderer europäischer und internatio-

naler Einrichtungen zu stellen" (3/3/3) (Im Klartext: "Wir wissen schon, wie wir uns revanchieren können").

- Außerdem: "Nicht zuletzt gefährdet Frankreich seinen internationalen Ruf" (3/3/4) (Soll heißen: "Frankreich steht allein und der Rest Europas hinter uns").
- Und das schafft Zuversicht, denn so viel wird Frankreich wohl doch nicht riskieren und vielleicht noch einlenken, bevor es soweit kommt. "Denn nationalegoistischer Ehrgeiz und Eitelkeiten (die es scheinbar nur auf französischer Seite gibt; W.K.) sind weniger denn je beliebt" (3/3/5).

Abb. 1.3.2.2: Das Ergebnis der Textinterpretation



Der latente Gehalt des Textes

Auf dem Wege der Abstraktion wird von der thematischen Form des Textes schließlich auf seine latenten Gehalt zurückgeschlossen, indem die Regeln der Wirklichkeitskonstruktion rekonstruiert werden, welche der Thematisierungsabwehr zugrunde liegen (vgl. Abb. 1.3.2.2). Dazu halten wir fest:

1. Der Text handelt von einem deutsch-französischen Konflikt um die Präsidentschaft der EZB.

2. Der manifeste Gehalt des Textes ist von unverhohlener Feindseligkeit gegenüber Frankreich geprägt.
3. Diese Feindseligkeit wird vor dem Hintergrund der Euro-Ängste in der deutschen Öffentlichkeit verstehbar, welche dem Konflikt einerseits eine sachliche und andererseits eine symbolische (Prestige-)Dimension verleihen.
   Auf der *Sachebene* könnte dem Konflikt leicht seine Bedrohlichkeit genommen werden (vgl. Kompetenz von Trichet), jedoch spricht alles für Frankreich (Legitimität, Kompetenz), so dass der Konflikt auf dieser Ebene nicht zu gewinnen ist.
   Auf der *Prestigeebene* können die Sicherheitsbedürfnisse der Öffentlichkeit nur symbolisch befriedigt werden ("Durchsetzung unseres Kandidaten"), jedoch trägt dies die Gefahr in sich, dass dies seitens der europäischen Partner als deutscher Vormachtsanspruch interpretiert werden könnte.
4. Die Dynamik der Themenentwicklung verlegt den Konflikt von der Sachebene auf die Prestigeebene und eskaliert ihn schließlich zu einem deutsch-französischen Prestigeduell nach Art eines Lose-Lose-Modells.
5. Die Regeln der Wirklichkeitskonstruktion, denen der Text dabei folgt (= der latente Gehalt des Textes), konzeptualisieren den Konflikt als Konkurrenzsituation (Win-Lose bis Lose-Lose), sie bauschen den Konflikt über die anstehenden Sachfragen (= Wahl eines EZB-Präsidenten) hinaus auf und gehen mit der Verleugnung eigener, nicht akzeptierter Anteile einher, die auf einen äußeren Feind (= Frankreich) projiziert werden, in dem sie bekämpft werden können.
6. Dadurch verschwindet die Sachebene, auf der der Konflikt nicht zu gewinnen ist, ebenso aus dem thematischen Horizont des Textes wie das deutsche Eigeninteresse und das eigene Prestigedenken, welche das europäische Kräfteverhältnis zuungunsten von Deutschland verschieben könnten (= der exterritorialisierte Gehalt des Textes).

Der sekundäre Gewinn aus der Thematisierungsabwehr

In einem letzten Interpretationsschritt können schließlich Hypothesen bezüglich des sekundären Gewinns aus der Thematisierungsabwehr formuliert werden, zu deren Überprüfung jedoch eine Kontexterweiterung (Baros & Reetz, 2002) erforderlich wäre.

So könnte man z. B. aus der von Walter Lippmann (1922) begründeten Nachrichtenwerttheorie (Schulz, 1976; Staab, 1990) die Hypothese ableiten, dass die Thematisierungsabwehr eine Art Kollateralschaden darstellt, welcher daraus resultiert, dass der

Kommentator den Konflikt zwecks Steigerung seines Nachrichtenwertes polarisiert. Oder man könnte die Hypothese formulieren, dass die Thematisierungsabwehr der Aufrechterhaltung der (bundes-)deutschen Nachkriegsidentität und der Vermeidung einer Auseinandersetzung mit dem neu gewonnenen Großmachtstatus des wiedervereinigten Deutschlands dient.

Diese Hypothesen müssen sich nicht unbedingt ausschließen und der sekundäre Gewinn aus der Thematisierungsabwehr kann auch multipler Natur sein. Die Überprüfung der Hypothesen geht über die Textinterpretation jedoch hinaus.

1.3.3 Gegenstandsbedingte Differenzierungen

Themenverschiebungen und thematische Brüche

In der Praxis der Textinterpetation haben wir es nicht selten mit Texten zu tun, deren Themenentwicklung nicht so linear ist wie in unserem Beispieltext, sondern Brüche enthält. Dies ist z. B. dann der Fall, wenn im Verlaufe eines Textes ein neuer Gedanke auftaucht, der das bisher Gesagte in neuem Licht erscheinen lässt und das Thema in eine andere Richtung weiterentwickelt, was so weit gehen kann, dass die am Ende gezogenen Schlussfolgerungen dem zuvor Gesagten diametral entgegenstehen.

Das methodische Prinzip der Konsistenz, dessen wir uns bedienen, um die gültigen Interpretationen einer Textstelle gegenüber den anderen möglichen Interpretationen auszuzeichnen (s. o.), ist daher nur relativ zu verstehen und darf nicht dazu führen, dass Widersprüche harmonisiert, thematische Brüche geglättet oder Themenverschiebungen verschleiert werden. Während die Übereinstimmung zwischen Textstellen einerseits als Indiz dafür genommen werden kann, dass sie richtig verstanden wurden, bilden andererseits gerade die Widersprüchlichkeiten und Ungereimtheiten den Anknüpfungspunkt für eine weitergehende Analyse des Textes (Leithäuser & Volmerg, 1979).

Interaktionsanalyse

Bei der Analyse von interaktivem Textmaterial, also z. B. bei der Interpretation von konfrontativen Interviews, Gruppendiskussionen oder Rollenspielen, aber auch bei der Interpretation von narrativen Interviews oder partnerzentrierten Gesprächen – kurz: bei der Interpretation jeglicher Textart, an deren Produktion mehrere Sprecher beteiligt sind und die Rolle zwischen Sprecher und Angesprochenem wechselt – sind Themenverschiebungen und thematische Brüche nahezu der Regelfall.

Bei der Interaktionsanalyse – der es ja um die Modifikationen des Themas geht, welche aus der Interaktion zwischen Sprecher und Angesprochenem resultieren – stehen sie sogar im Fokus der Aufmerksamkeit des Interpreten. Dies hat zur Folge, dass die Rekonstruktion des mit dem Rollenwechsel zwischen Sprecher und Angesprochenem einhergehenden Spannungsverhältnisses von gemeintem und aufgenommenem Sinn (vgl. Abb. 1.3.3.1) den Kern der Interpretationsarbeit bildet: "Was sagt der Sprecher?, "Was will er damit mitteilen?", "Wie wird das mitgeteilt?", "Welche Antwort wird dadurch herausgefordert", "Wie wird das vom Angesprochenen tatsächlich aufgenommen?", "Welche Antwort gibt er?" usw.

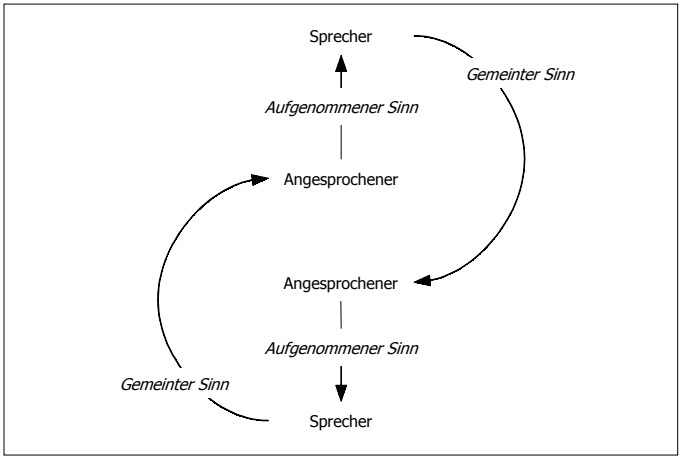


Abb. 1.3.3.1: Das Spannungsverhältnis zwischen gemeintem und aufgenommenem Sinn

Als empirische Grundlage für die Rekonstruktion möglicher Diskrepanzen zwischen gemeintem und aufgenommenem Sinn können wir uns nur auf den externen Kontext (z. B. "Was wusste der Sprecher, das dem Angesprochenen unbekannt war?" oder "Welche Empfindlichkeiten des Angesprochenen werden durch das Gesagte tangiert?") und auf den je internen Kontext der Sprechbeiträge der jeweiligen Sprecher stützen. Interpretationsverfahren, welche den externen Kontext ausklammern und/oder (schon bei der Transkription) zwar Sprecherwechsel notieren, die Sprecher aber unidentifiziert lassen (z. B. Volmerg, 1980), sind für die Interaktionsanalyse daher ungeeignet.

Narratives Interview und partnerzentriertes Gespräch

Indem wir davon ausgehen müssen, dass die Sinnkonstitution – zumindest teilweise – erst in der Kommunikation zwischen Interviewer und Interviewten stattfindet (s. o.), ist die Interaktionsanalyse auch bei der Interpretation von narrativen Interviews und partnerzentrierten Gesprächen unverzichtbar, obwohl der Einfluss, welchen der Interviewer auf den Gesprächsverlauf nimmt, dort auf ein Minimum beschränkt wird.

Anders als z. B. bei der Interpretation von Gruppendiskussionen, wo die Interaktionsanalyse im Zentrum des Forschungsinteresses steht, hat sie hier jedoch in erster Linie eine methodologische Hilfsfunktion zwecks Problematisierung und Absicherung der Validität der Selbstauskünfte des Probanden: "Hätte der Interviewte eine bestimmte Aussage auch von sich aus gemacht, oder wurde sie ihm vom Interviewer in den Mund gelegt?", "Vermeidet der Interviewte ein bestimmtes Thema anzusprechen, oder hat der Interviewer davon abgelenkt?", usw.

Handlungsanalyse

Ziel der Handlungsanalyse ist es, die subjektiven Handlungsprämissen der Befragten zu rekonstruieren. Erkenntnisgegenstand der Handlungsanalyse ist daher nicht nur "die Beschreibung isolierter Tätigkeiten, sondern die Analyse der Handlungszusammenhänge und Kontexte, in denen diese Tätigkeiten stehen, und aus welchen die Personen ihre jeweiligen Handlungsvorsätze subjektiv als 'logische' bzw. 'vernünftige' Konsequenz ableiten" (Baros & Reetz, 2002, 2).

Namentlich in der Migrationsforschung, wo die Autoren die sozialpsychologische Rekonstruktion auf die Interpretation von qualitativen Interviews anwenden (vgl. Baros, 2001; Reetz, 2007), bedeutet dies, über eine einfache Beschreibung individueller Subjektivität hinauszugehen und zu analysieren, wie individuelle Subjektivität gesellschaftlich hervorgebracht wird. Davon ausgehend, dass "das Bewusstsein kein Abbild des Realen ist, und dass das Reale nicht die willkürliche Konstruktion des Bewusstseins ist" (Freire, 1981, 51f), wird dies dahingehend zugespitzt, dass die Diskrepanz zwischen gesellschaftlichem Sein und Bewußt-Sein nicht einfach auf einem Irrtum der Subjekte beruht, sondern gesellschaftlich bedingte Abwehrprozesse beinhaltet.

"Das heißt, konsequent zu Ende gedacht, dass die Rekonstruktion gesellschaftlich hervorgebrachter Subjektivität auf keinen Fall die Vernachlässigung tiefenpsychologischer Aspekte, also die Vernachlässigung von Momenten der komplexen Struktur des Psychischen bedeutet. Im Gegenteil:

Gerade die Problematisierung subjektiver Formen der Realitätsbewältigung (wie Rationalisierungen, Verdrängungen, Abwehrmechanismen etc.) und ihre Funktion für den einzelnen steht im Vordergrund der Analyse" (Baros & Reetz, 2002, 2).

Der Fokus der sozialpsychologischen Rekonstruktion liegt bei Baros & Reetz damit nicht nur auf der Herausarbeitung der Regeln, nach denen die Migranten ihre Wirklichkeit konstruieren (= latenter Gehalt des Textes), sondern darüber hinausgehend auf der Frage nach der Art und Weise, wie sie mithilfe dieser Konstrukte ihre Migrationsrealität bewältigen.

Strukturelle Verallgemeinerung

Unter Einbeziehung von Handlungs- und Konflikttheorien sowie der Entfremdungstheorie von Sève zielen Baros & Reetz auf die theoretische Erfassung wesentlicher Aspekte der Lebenssituation von Migranten sowie deren Handlungsmöglichkeiten, -einschränkungen und -widersprüche, was auf dem Wege einer strukturellen Verallgemeinerung der empirisch gewonnenen Ergebnisse erzielt wird, d. h. indem 'typische', theoretisch verallgemeinerte Konflikt- und Problemkonstellationen herausgearbeitet werden.

Sekundärer Gewinn aus der Thematisierungsabwehr

Als empirische Grundlage, auf welcher die Bewältigung der Migrationssituation eruiert wird, dient

- die Rekonstruktion jener Bedeutungsaspekte der Kommunikation, die vom Interviewten systematisch abgewehrt werden (= exterritorialisierter Gehalt des Textes) und – weil aus der Sprache ausgeschlossen – durch eine Analyse des Textes allein nicht zugänglich sind, sondern die Einbeziehung von externem Kontext unerlässlich machen, sowie
- die Beantwortung der Frage nach dem sekundären Gewinn aus der Thematisierungsabwehr, die nur im Zusammenhang mit der aktuellen Situationsbewältigung der Interviewten angemessen beantwortet werden kann und daher eine über die Interpretation eines gegebenen Interviews hinausgehende Kontexterweiterung in Richtung auf eine umfassende Analyse der aktuellen Lebenspraxis der Interviewten erfordert.

Um dies zu leisten, wird untersucht, inwieweit sich die als latenter Gehalt der analysierten Kommunikation herausgearbeiteten Regeln subjektiver Wirklichkeitskonstruktion auch in anderen Bereichen der Lebenspraxis der Interviewten als Bestandteile ihrer Realitätsbewältigung wiederfinden, womit sich der Analyseschwerpunkt auf den Aspekt der Handlungsfähigkeit der Interviewten verlagert.

"Dabei soll der subjektive Wert aufgezeigt werden, welchen die Subjekte aus der Fixierung auf Handlungsmuster beziehen, die zwar kurzfristig einer vordergründigen, belastungslindernden Situationsbewältigung dienen mögen, jedoch langfristig betrachtet zu einer permanenten Reproduktion ihrer Problemlage führen (können)" (ebd., 11).

Das argumentative Rückkoppelungsgespräch

Als methodischen Zugang zu dem dafür erforderlichen empirischen Material entwickelten Baros & Reetz daher die Methode des argumentativen Rückkoppelungsgespräches, das die im Verlaufe des gesamten Forschungsprozesses geleisteten Interpretationen im Sinne einer argumentativen Geltungsprüfung von Interpretationen (Terhart, 1981, 769ff.) zum Gegenstand eines Diskurses mit den Betroffenen macht.

Zentrales Charakteristikum des argumentativen Rückkoppelungsgespräches ist die Aufhebung der Trennung zwischen Interpretation und Intervention, wodurch die Möglichkeit geschaffen wird, bereits im Forschungsverlauf eine Prämissenerweiterung für die Beteiligten (Forscher und Betroffene) herbeizuführen. Entsprechend geht der Argumentationsprozess weder als Bevormundung der Betroffenen vonstatten noch verfolgt er das Ziel einer reibungslosen kommunikativen Einigung, sondern ist als ein nach begründbaren Regeln ablaufender Beratungsprozess zwischen Interpreten und Interpretierten und insofern auch als eine innovationserzeugende Praxis (Reetz, 2007, 174) zu verstehen.

Von den gebräuchlichen Modellen einer kommunikativen Validierung, welche die Geltungsfrage auf dem Wege einer Einigung zwischen Interpreten und Interpretierten lösen (z. B. Heinze & Klusemann, 1979, 195f.) und/oder sich der Gültigkeit der Interpretationen vergewissern, indem sie eine dialogische Absicherung aus der Eigenperspektive des Interpretierten vornehmen lassen (z:B. Scheele & Groeben, 1988, 49; Wilke, 1992, 334), unterscheidet sich das argumentative Rückkoppelungsgespräch somit dadurch, dass es nicht die Einigung zum Kriterium für die Gültigkeit der Interpretationen macht, sondern darauf abzielt, mit Hilfe der aus dem Argumentationsprozess gewonnenen neuen Erkenntnisse den hermeneutischen Zirkel weiterzuführen.

Kontrollfragen

- Welche Varianten der Textinterpretation lassen sich unterscheiden und welche Abhängigkeit besteht zwischen der Textart einerseits und den Zielsetzungen der Textinterpretation andererseits?

- Worin besteht das Ergebnis der Textinterpretation?
- Erläutern Sie das Ablaufschema der Textinterpretation!
- Welche Punkte umfasst die Explikation des Vorbegriffs?
- Welche Schritte umfasst die erste Orientierung und was sollte dazu in einem Memo festgehalten werden?
- Erläutern Sie die Vorgehensweise der Sequenzanalyse!
- Welches ist der Ausgangspunkt der Recherche des externen Kontextes, bezüglich welcher Fragen wird dabei Hintergrundwissen zusammengetragen und unter welchen Aspekten sind diese zu betrachten?
- Erläutern Sie die Vorgehensweise bei der Mikroanalyse des Textes und welche erkenntnisleitenden Fragen dabei an den Text gerichtet werden!
- Erläutern Sie die Vorgehensweise bei der Makroanalyse des Textes!
- Erläutern Sie den hermeneutischen Zirkel der Textinterpretation!
- Was versteht man unter der thematischen Form des Textes?
- Diskutieren Sie, warum bezüglich des sekundären Gewinns aus der Thematisierungsabwehr zwar Hypothesen generiert, diese jedoch anhand des Textes allein nicht überprüft werden können!
- Welches Spannungsverhältnis besteht zwischen dem Prinzip der Konsistenz einerseits und Themenverschiebungen und/oder thematischen Brüchen andererseits?
- Mittels welcher interpretationsleitender Fragen kann das Spannungsverhältnis zwischen gemeintem und aufgenommenem Sinn analysiert und auf welche Arten von Kontexten kann dabei zurückgegriffen werden?
- Bei welcher Art von Texten steht die Interaktionsanalyse im Vordergrund, welche Rolle spielt sie bei der Interpretation narrativer Interviews und in welcher Form werden die Selbstauskünfte des Probanden dabei problematisiert?
- Worin besteht das Ziel der Handlungsanalyse?
- Was verstehen Baros & Reetz unter einem argumentativen Rückkoppelungsgespräch, wie unterscheidet es sich von einer kommunikativen Validierung und inwiefern bedeutet es eine Kontexterweiterung?

1.4
Inhaltsanalyse

1.4.1 Varianten der Inhaltsanalyse

Die Leitfrage der Inhaltsanalyse

Die Methode der Inhaltsanalyse entstand während der 20er Jahre des vergangenen Jahrhunderts zunächst aus dem Bedürfnis nach einer systematischen Auswertung großer Textdatenmengen der sich entfaltenden Massenmedien (vgl. Merten, 1983; Lissmann, 1997). Entsprechend zielte sie zunächst auf quantifizierende Analysen ab: Häufigkeitsanalysen, in denen bestimmte manifeste Inhalte ausgezählt werden (z. B. wie oft eine bestimmte politische Partei in einem bestimmten Medium erwähnt wird), Indikatoranalysen, in denen latente Inhalte durch bestimmte Textmerkmale operationalisiert werden (z. B. die Verwendung von Wörtern wie 'muss', 'niemals', 'notwendigerweise' als Indikatoren für Dogmatismus); Valenz- und Intensitätsanalysen, die den Text auf Rating-Skalen einschätzen (z. B. wie stark ein Zeitungskommentar die Meinung der jeweiligen Regierungspartei ausdrückt) und Kontingenzanalysen, die Zusammenhänge zwischen verschiedenen Textmerkmalen untersuchen (z. B. wie oft ein bestimmter Politiker im Kontext von positiven Ereignissen erwähnt wird).

Die Leitfrage dieser frühen inhaltsanalytischen Untersuchungen wurde von Lasswell, Lerner & Pool (1952; zit.n. Friedrichs, 1973) in die viel zitierte Formel

"Wer sagt was zu wem, wie, warum und mit welchem Effekt?"

gefasst, die allerdings darüber hinausgeht, was die Inhaltsanalyse als solche tatsächlich zu leisten vermag.

Inhaltsanalytische Versuchspläne

Dies wird auch anhand der verschiedenen Forschungsprobleme deutlich, zu deren Untersuchung auf inhaltsanalytische Methoden zurückgegriffen wird, die Holsti (1969) in drei Hauptkategorien gruppiert (vgl. Tab. 1.4.1.1) und zu den Grundbegriffen der klassischen Kommunikationstheorie – Nachricht, Sender, Kodierung, Kanal, Dekodierung und Empfänger – in Beziehung setzt.

Absicht	Bezug	Fragerichtung	Forschungsproblem
Rückschlüsse auf Vorgänge vor der Mitteilung	Sender	Wer?	Quelle der Mitteilung
	Kodierung	Warum?	Zweck der Mitteilung
Deskription von Merkmalen der Mitteilung	Kanal	Wie?	Form der Mitteilung
	Nachricht	Was?	Inhalt der Mitteilung
Rückschlüsse auf Vorgänge nach der Mitteilung	Empfänger	Zu wem?	Zielgruppe der Mitteilung
	Dekodierung	Mit welchem Effekt?	Reaktion auf die Mitteilung

Tab. 1.4.1.1: Inhaltsanalytische Forschungsprobleme (nach Holsti, 1969, modifiziert)

Quantitative Inhaltsanalyse

Namentlich die quantitative Inhaltsanalyse, die Berelson (1952) als eine "Forschungstechnik zur objektiven, systematischen und quantitativen Deskription von manifestem Kommunikationsinhalt" definiert, hat lediglich die Deskription von Merkmalen der Mitteilung zum Gegenstand, wozu ggf. auch die Frage nach den vom Text angegebenen Quellen (z. B. Nachrichtenagenturen, Korrespondentenberichten, etc.) gezählt werden können. Schon die Klärung einer umstrittenen Autorschaft sowie jegliche Rückschlüsse auf Vorgänge vor oder nach der Mitteilung gehen über das von Berelson formulierte Programm jedoch hinaus und erfordern eine Interpretation der Analyseergebnisse.

Berelsons Verständnis der Inhaltsanalyse geriet daher schon bald in die Kritik und bereits Kracauer (1952, 636f.) wies in seinem berühmten Plädoyer für eine qualitative Inhaltsanalyse darauf hin, dass auch die Beschreibung der manifesten Textinhalte in vielen Fällen eine Interpretationsleistung voraussetzt, weshalb die Daten, die in der quantitativen Inhaltsanalyse statistisch weiterverrechnet werden, keineswegs so objektiv und reliabel sind, wie Berelson dies sah. Im Gegenteil kann die Fixierung auf manifeste Textinhalte und deren Beschreibung durch an der Textoberfläche bleibende Variablen zu unangemessenen Vereinfachungen führen, die das Bild verzerren (ebd., 632).

Die Forderung nach qualitativen Analysemethoden

Mit der Forderung nach qualitativen Analysemethoden vollzog Kracauer (ebd., 637f.) aber nicht etwa einen radikalen Bruch mit den quantitativen Methoden, sondern er vertritt die Auffassung, dass quantitative und qualitative Analysen keineswegs so verschieden sind, wie es auf den ersten Blick erscheinen mag. Auch die quantitative Inhaltsanalyse beinhaltet qualitative Aspekte und

auch die qualitative Analyse ist für quantifizierende Analyseschritte dort offen, wo sie gebraucht werden.

Wogegen Kracauer sich wandte, ist allerdings die dogmatische Fixierung auf quantifizierende Methoden als Ausweis der Wissenschaftlichkeit schlechthin sowie die teilweise auch heute noch vorzufindende Auffassung, wonach qualitative Studien bloß explorativen Charakter hätten und in den Bereich der Voruntersuchungen zu verweisen seien – etwa zwecks Generierung von Hypothesen, die es dann mit 'harten' quantitativen Methoden zu überprüfen gilt. Quantitative und qualitative Methoden haben vielmehr ihre je spezifischen Vor- und Nachteile und ergänzen sich wechselseitig.

Gegenüberstellung quantitativer und qualitativer Analysemethoden

Zu den unbezweifelbaren Vorzügen der quantitativen Inhaltsanalyse gehört ihre Fähigkeit zur Bewältigung großer Textmengen. Entsprechend ist sie immer dann die Methode der Wahl, wenn man auf repräsentative Aussagen (z. B. über *die* Darstellung des Konfliktes um die Präsidentschaft der EZB in der deutschen Presse schlechthin) abzielt. Andererseits bleiben die Beschreibungsmittel, der quantitativen Inhaltsanalyse mehr an der Textoberfläche und die Beschreibung der Texte erfolgt durch eine Reihe von Variablen, welche die erfassten Textmerkmale aus ihrem Zusammenhang reißt, um sie anschließend auf dem Wege der statistischen Datenaggregation wieder zusammenzufügen. Diese Rekontextualisierung der Textmerkmale stellt sie damit aber in einen anderen Kontext als jenen, in welchem sie ursprünglich standen. Dies kann am Ende zu Fehlinterpretationen führen, denen man vorbeugen kann, indem man zumindest eine kleine Stichprobe von typischen Textbeispielen daraufhin analysiert, ob sie die Interpretationen auch tatsächlich einzulösen vermögen und/oder wie die Interpretationen ggf. zu modifizieren sind. Dies kann erst durch qualitative Methoden geleistet werden, die mehr in die Tiefe der Texte gehen, wegen ihrer Aufwändigkeit aber auf die Analyse einer relativ kleinen Anzahl von Texten beschränkt bleiben, die einander kontrastierend gegenübergestellt werden können.

Qualitative Inhaltsanalyse

Auch die qualitative Inhaltsanalyse beschreibt die Texte durch eine Reihe von Variablen. Diese berücksichtigen aber verstärkt auch den command und parade aspect der Texte und werden – in ihrem ursprünglichen Kontext belassen – auf argumenta-tivem Wege zueinander in Beziehung gesetzt. Letzteres geschieht exemplarisch qua Einzelfall und nicht erst summarisch qua statis-

tisch aggregierter Textstichprobe. Hinsichtlich ihrer Systematik ist sie von der quantitativen Inhaltsanalyse jedoch nicht grundsätzlich verschieden.

Beide Analysemethoden sind sowohl regelgeleitet (d. h. sie folgen vorab formulierten Ablaufmodellen), als auch theoriegeleitet (d. h. sie folgen theoretisch begründeten Fragestellungen und Kodierregeln) und durch eine schrittweise, den Text in einzelne Analyseeinheiten zergliedernde und an Kategoriensystemen orientierte Vorgehensweise (vgl. Krippendorff, 1980) gekennzeichnet. Auch die qualitative Inhaltsanalyse erhebt den Anspruch, sich an Gütekriterien wie der Interkoderreliabilität zu messen. Die Ansprüche sind zwar etwas niedriger, das Ziel aber bleibt, dass mehrere Inhaltsanalytiker nachweislich zu ähnlichen Ergebnissen kommen (vgl. Mayring, 2000).

Inhaltsanalytische Kodierung

Beiden Analysemethoden ist gemeinsam, dass sie – ähnlich wie die textinterpretativen Methoden – bestimmte Fragen an den Text richten. So kann man z. B. bei der Analyse von Konfliktgesprächen den Text daraufhin untersuchen, ob der Sprecher Schuldzuweisungen vornimmt und/oder abfällige Worte verwendet, usw.

Die *quantitative Inhaltsanalyse* antwortet darauf z. B. mit ja/nein, oder mit dem Auszählen der abfälligen Worte, oder mit der Einschätzung der Abfälligkeit des Sprachgebrauchs auf einer Likertskala (z. B. nicht abfällig – leicht abfällig – ziemlich abfällig – total abfällig).

Diese Antworten bilden die *Kategorien* der entsprechenden inhaltsanalytischen *Variablen* und werden in Zahlen übersetzt, die als Grundlage der statistischen Auswertung dienen. Zum Beispiel:

0 = nein
1 = ja

oder

0 = kein abfälliges Wort
1 = ein abfälliges Wort
2 = zwei abfällige Worte
3 = mehr als zwei abfällige Worte

oder

0 = nicht abfällig
1 = leicht abfällig

2 = ziemlich abfällig
3 = total abfällig

Dieser Prozess der Zuordnung von Kategorien bzw. Zahlen zu den analysierten Texten bzw. Textstellen wird als *inhaltsanalytische Kodierung* bezeichnet.

Quantitative Variablen

Um für die statistische Weiterverrechnung geeignet zu sein, müssen die Kategorien so definiert werden, dass

1. alle Kategorien hinreichend oft vorkommen[1] und
2. die Kodierung der Texte hinreichend objektiv ist, um mit verlässlichen Zahlen weiterrechnen zu können.

Dies hat zur Folge, dass die an den Text gerichteten Fragen – z. B. ob der Sprecher abfällige Worte verwendet – vergleichsweise einfach bleiben.

Qualitative Variablen

In der qualitativen Inhaltsanalyse können die Fragen komplexer sein – z. B. ob und ggf. wie der Sprecher einen unversöhnlichen Gegensatz zwischen sich und seinem Gegenüber konstruiert. Entsprechend fallen auch die Antworten komplexer aus, z. B:

"Der Sprecher konstruiert insofern einen unversöhnlichen Gegensatz, als er den Konflikt als Win-Lose-Prozess konzeptualisiert und jegliche Kompromissbereitschaft ablehnt."

Wegen dieser Komplexität und weil es der qualitativen Inhaltsanalyse auch nicht um die statistische Weiterverrechnung von Daten geht, wird auf eine Übersetzung in Zahlen zumeist verzichtet. Gleichwohl bedient sich auch die qualitative Inhaltsanalyse einer normierten und standardisierten Begrifflichkeit, die bestimmte, als relevant erachtete Aspekte des Textinhalts fokussiert. Die dazu verwendeten Beschreibungsprädikatoren (z. B. Konzeptualisierung des Konfliktes als Win-Lose-Prozess, Ablehnung von Kompromissbereitschaft) werden anhand von Beispielen und Gegenbeispielen exemplarisch eingeübt und in ihrer Verwendungsweise mittels Prädikatorenregeln abgesichert. In der Literatur werden sie als qualitative Variablen bezeichnet und entsprechend spricht man auch in der qualitativen Inhaltsanalyse von einer Kodierung des Textmaterials.

Die Offenheit des Kodierverfahrens

Ein wesentlicher Unterschied zwischen der quantitativen und der qualitativen Inhaltsanalyse besteht jedoch in der Offenheit des

1. Was hinreichend ist, kann je nach statistischer Analysemethode variieren.

inhaltsanalytischen Kategoriensystems für Erweiterungen und Modifikationen.

In der quantitativen Inhaltsanalyse besteht sie nur während der Konstruktion des Systems. Danach wird es zur Auswertung des Textmaterials schematisch angewendet, was schon allein dafür notwendig ist, dass die Bedeutung der statistisch weiterverrechneten Zahlen invariant bleibt (vgl. Intercoder-Reliabilität).

In der qualitativen Inhaltsanalyse bleibt das Kodiersystem dagegen während des gesamten Auswertungsprozesses offen für Modifikationen, z. B. für neue Aspekte, deren Relevanz für die Beschreibung des Textmaterials erst aus der Kenntnis des Textmaterials heraus offenkundig wird. Deshalb wird es oft erforderlich, am Ende der qualitativ-inhaltsanalytischen Auswertung noch einmal an ihren Anfang zurückzugehen, weil die relevanten neuen Aspekte vielleicht schon in früheren Textstellen manifest waren, dort aber noch nicht beachtet wurden.

Bei den oben dargestellten textinterpretativen Verfahren gibt es dagegen gar keine Trennung von Auswertung und Interpretation. Die Interpretation geht von der einzelnen Textstelle zum Gesamttext und zum externen Kontext und wieder zurück und bleibt so ständig offen für neue Akzentuierungen, bis die Interpretation am Ende zu einem (für die erkenntnisleitende Fragestellung) angemessenen Verständnis des Textes konvergiert.

Sicherung von Transsubjektivität

Allen drei Arten von textanalytischen Methoden ist der Anspruch auf Objektivität im Sinne einer Transzendierung der Subjektivität des Forschers gemeinsam, so dass jeder sachkundige und unvoreingenommene Analytiker (im Wesentlichen) zu dem selben Ergebnis kommen kann. Während die Sachkundigkeit in allen drei Fällen durch korrekte Methodenanwendung und Regelbefolgung sichergestellt werden kann, lässt sich die Beeinflussung der Forschungsergebnisse durch Voreingenommenheiten des Forschers dagegen nicht so leicht kontrollieren.

Selbst in der quantitativen Inhaltsanalyse, wo die Voreingenommenheit des Forschers qua Umwandlung empirischer Wirklichkeit in anonyme Zahlen partiell ausgeschaltet werden kann[1], ist bereits die Entwicklung des Kodierschemas interpretationshaltig und müssen die Ergebnisse der statistischen Datenaggregation

1. Vorausgesetzt, dass der Kodierprozess hinreichend reliabel ist.

schließlich interpretiert werden. Die Unvoreingenommenheit von Interpretationen kann aber nicht durch Distanzierung vom Forschungsgegenstand erzielt werden, sondern nur durch Distanzierung von der je eigenen Voreingenommenheit.

Bezüglich der Frage, wie diese Distanzierung gelingen kann, können wir davon ausgehen, dass Forschung nicht Selbstzweck ist, sondern einem praktischen Interesse dient. Entsprechend besteht auch ein praktisches Interesse daran, richtige von falschen Interpretationen zu trennen. Nicht indem wir der Fiktion einer wertfreien Wissenschaft anhängen, sondern indem wir uns auf ein expliziertes Erkenntnisinteresse verpflichten, können wir unsere Voreingenommenheiten daher am ehesten neutralisieren. Wenn sich z. B. in der Konfliktforschung unser Interesse darauf richtet, welchen Gesetzmäßigkeiten die Konstruktion sozialer Wirklichkeiten folgt, die einen gewaltförmigen Konfliktaustrag unabweisbar machen, so werden wir eine gültige Antwort darauf nicht finden können, solange wir den Analyseergebnissen unsere eigenen Feindbildkonstruktionen überstülpen, sondern erst dann, wenn wir uns von unserer Parteilichkeit zugunsten der einen oder anderen Konfliktpartei ebenso distanzieren wie von unseren eigenen pazifistischen oder bellizistischen Einstellungen.

1.4.2 Qualitative Inhaltsanalyse

Abgrenzung gegenüber textinterpretativen Verfahren

Auch wenn der Unterschied zwischen Textinterpretation und qualitativer Inhaltsanalyse oft fließend ist, können wir die beiden Analysemethoden doch insofern voneinander abgrenzen, als die Inhaltsanalyse eine stärkere Trennung zwischen der Analyse des Textinhaltes einerseits und ihrer kontextuellen Interpretation andererseits vornimmt. Während die Textinterpretation in einem einzigen großen hermeneutischen Zirkel von der einzelnen Textstelle zum Gesamttext und zum externen Kontext und wieder zurück vorgeht, beinhaltet die Inhaltsanalyse zwei getrennte hermeneutische Zirkel, deren erster in die Deskription des Textinhaltes mündet und deren zweiter die Analyseergebnisse vor dem Hintergrund seines externen Kontextes interpretiert (vgl. Abb. 1.4.2.1).

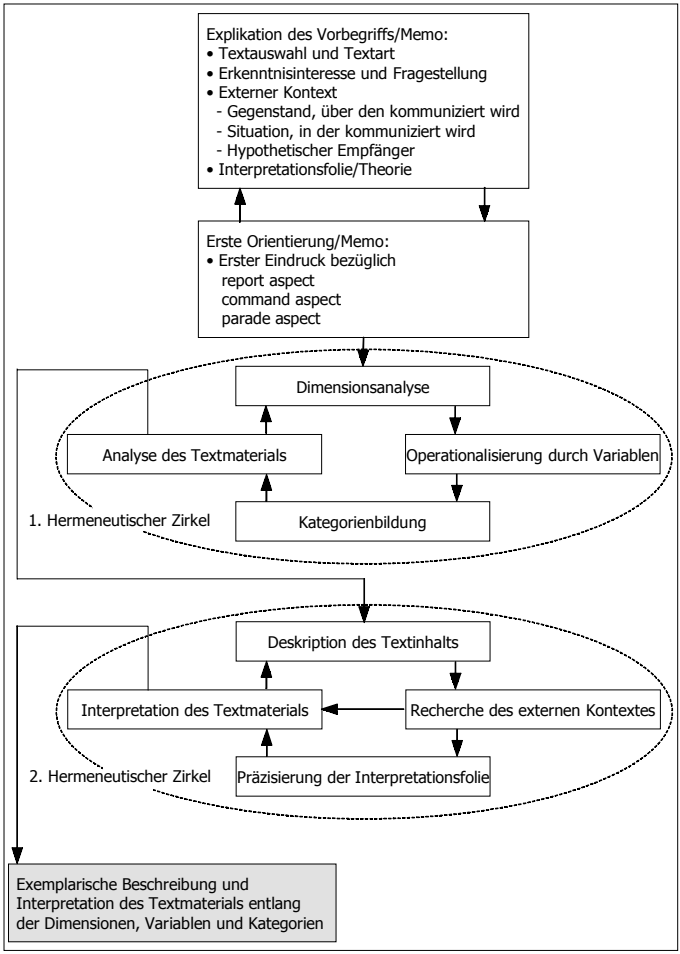


Abb. 1.4.2.1: Ablaufschema der qualitativen Inhaltsanalyse

Explikation des Vorbegriffs

Ähnlich wie bei der Textinterpretation steht auch am Anfang der Inhaltsanalyse eine Explikation des Vorbegriffs des Interpreten, die in einem Memo festgehalten wird und insbesondere eine erste Skizzierung der Interpretationsfolie enthält, die später der Konstruktion der inhaltsanalytischen Variablen und der Interpretation der Analyseergebnisse zugrunde gelegt werden soll. Je nachdem, ob der Interpret bereits über eine ausgearbeitete Theorie verfügt, die als Interpretationsfolie dienen kann, oder nicht, kann zwischen einer inspektiven und einer explorativen Analyse des Textmaterials unterschieden werden. Jedoch ist auch die ex-

plorative Analyse in theoretischer Hinsicht nicht ganz voraussetzungsfrei und beinhaltet zumindest einige Vorannahmen darüber, welche Art von Fragen an den Text gerichtet werden sollen und welche Art von Erkenntnis sich der Interpret davon verspricht.

Erste Orientierung

Als Einstieg in die Analyse des Textmaterials – d. h. noch vor der Konstruktion der Beschreibungsmittel, mit denen es beschrieben werden soll – empfiehlt sich auch hier wieder, den ersten Eindruck, welchen das Textmaterial hinsichtlich seines report, command und parade aspects vermittelt, in einem Memo festzuhalten.

Der erste hermeneutische Zirkel

Die eigentliche Analyse des Textmaterials vollzieht sich dann in einem ersten hermeneutischen Zirkel, der zunächst die (zumindest grobe) Richtung der Fragen festlegt, die an das Textmaterial gestellt werden sollen. Dieser Schritt wird in der Literatur (z. B. Friedrich, 1973, 321) auch als Dimensionsanalyse bezeichnet. Diese Fragerichtungen werden sodann durch Variablen und Kategorien operationalisiert, welche die Beschreibungsmittel zur Analyse des Textinhalts bereitstellen. Zielsetzung ist es, alle relevanten Merkmale des Textmaterials zu beschreiben. Dabei kann sich zeigen, dass die zunächst formulierten Fragerichtungen (Dimensionen) um weitere ergänzt, die Variablen weiter ausdifferenziert und/oder die Definition der Kategorien modifiziert werden muss, usw.

Deskription des Textinhaltes

Am Ende dieses hermeneutischen Zirkels steht ein inhaltsanalytisches Kategoriensystem bzw. Kodierschema (engl. coding schedule), das aus Dimensionen (= Variablengruppen), Variablen und Kategorien besteht, mittels derer der Inhalt des Textmaterials beschrieben wird. Während derselbe Text oder dieselbe Textstelle (je nach Kodiereinheit) dabei durch jede der Variablen zu beschreiben ist, sind die Kategorien so zu bilden, dass jeder Text (bzw. jede Textstelle) bezüglich jeder Variable in genau eine Kategorie gefasst werden kann.

Anforderungen an das Kodierschema

Dazu müssen die Kategorien so konstruiert werden, dass sie ein Klassifikationssystem bilden, d. h. sie müssen erschöpfend (engl. *exhaustive*) und einander ausschließend (engl. *mutually exclusive*) sein.

Exhaustive bedeutet, dass sich für jede Variable stets eine Kategorie finden lässt, unter welche der Text oder die Textstelle fällt.

Notfalls muss man eine Restkategorie bilden (in die dann all die Texte bzw.Textstellen fallen, die sich mit den bisherigen Kategorien nicht beschreiben lassen). Im Verlaufe der Weiterentwicklung des Kodierschemas kann man ggf. weitere Kategorien entwickeln, unter welche sich jeweils ein Teil der Texte fassen lässt, die bisher in die Restkategorie gefallen sind. Ziel ist es, die Texte so umfassend wie möglich zu beschreiben und die Restkategorie so klein wie möglich zu halten.

Mutually exclusive heißt, dass sich die Kategorien derselben Variable wechselseitig ausschließen. Dies gilt selbstverständlich nicht für die Kategorien verschiedener Variablen, da jeder Text bzw. jede Textstelle ja bezüglich jeder der Variablen in eine Kategorie fassbar sein muss.

Festlegung der Kodiereinheiten

Anders als in der quantitativen Inhaltsanalyse, die eine Vorabfestlegung der Kodiereinheiten erfordert, ist dies bei der qualitativen Inhaltsanalyse nicht dar Fall. Zwar wird man in der Regel satzweise vorgehen, doch hängt die Definition der jeweiligen Kodiereinheit von der Art der Variable ab, die zu kodieren ist, und die qualitative Inhaltsanalyse ist dafür offen, dass verschiedene Variablen sich auf verschiedene Kodiereinheiten beziehen. So können manche Variablen den Text als Ganzes betreffen, andere die einzelnen Textstellen und wiederum andere (z. B. Doppelbindungen und zweiseitige Botschaften; vgl. Reimann, 2002) mögen es erfordern, verschiedene Textstellen zueinander in Beziehung zu setzen. In der quantitativen Inhaltsanalyse müssen die Variablen dagegen so formuliert werden, dass sie alle auf dieselben Kodiereinheiten anwendbar sind.

Der zweite hermeneutische Zirkel

In einem zweiten hermeneutischen Zirkel erfolgt die Interpretation der inhaltsanalytischen Ergebnisse, indem diese auf den externen Kontext des Textmaterials bezogen und zu der theoretischen Interpretationsfolie in Beziehung gesetzt werden.

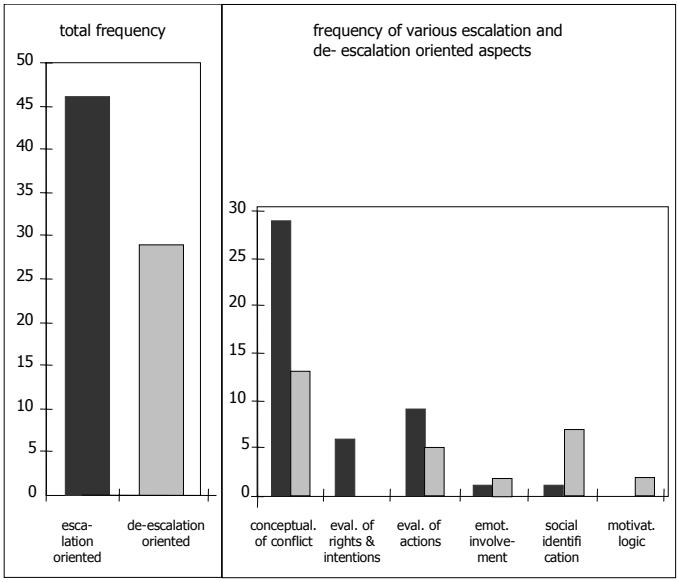
Darstellung der Analyseergebnisse

Die abschließende Darstellung der Analyseergebnisse ist exemplarisch und deskriptiv. Sie beschreibt den Textinhalt entlang der oben entwickelten Dimensionen, Variablen und Kategorien und interpretiert ihn in Hinblick auf die eingangs formulierte Fragestellung der Untersuchung. Dabei kann ggf. auch auf Quantifizierungen und/oder graphische Darstellungen zurückgegriffen werden.

Exkurs: Die Berichterstattung der *Berliner Zeitung* über das nordirische Friedensabkommen

Als Beispiel hierfür gehen wir kurz auf eine Studie von Hamdorf (2001) ein, welche die Berichterstattung über das nordirische Friedensabkommen in der *Berliner Zeitung* vom 11. April 1998 zum Gegenstand hat.
Obwohl der Bericht der Berliner Zeitung voll von Sympathie für das Friedensabkommen zu sein schien und an einer Stelle sogar eine Motivationslogik zugunsten des Friedensprozesses entwickelte, zeigten die Ergebnisse der Studie eine klare Dominanz von eskalationsorientierten Textmerkmalen, sowohl insgesamt als auch bezüglich der Dimensionen (1) der Konzeptualisierung des Konfliktes, (2) der Evaluation von Rechten und Intentionen der Konfliktparteien und (3) der Bewertung ihrer Handlungen. Deeskalationsorientierte Textmerkmale, die – wenngleich weit seltener – auch bezüglich (1) der Konzeptualisierung des Konfliktes und (2) der Bewertung der Handlungen der Konfliktparteien vorhanden sind, dominieren dagegen die – nur gelegentlich kodierten – Dimensionen (1) der emotionalen Verwicklung, (2) der sozialen Identifikation und (3) der Motivationslogik (vgl. Abb. 1.4.2.2).

Abb. 1.4.2.2: Häufigkeit eskalations- und deeskalationsorientierter Textmerkmale in der Berichterstattung über den nordirischen Friedensvertrag (nach ASPR, 2003, 90)



Zur Interpretation des Analyseergebnisses greift Hamdorf (ebd.) auf das Modell der konstruktiven Konfliktberichterstattung nach Kempf (2003) zurück und diskutiert die Analyseergebnisse in Hinblick auf die Fragestellung, ob und wie weit der Friedensprozess durch die Art seiner Darstellung in der *Berliner Zeitung* unterstützt wird.
Dazu stellt Hamdorf in einem ersten Schritt die Verteilung der eskalations- und deeskalationsorientierten Textmerkmale über den Text hinweg dar (vgl. Abb. 1.4.2.3) und zeigt, dass der Bericht der *Berliner Zeitung* zwar deeskalationsorientiert beginnt, diese Art der Berichterstattung jedoch im weiteren Verlauf des Textes kontinuierlich abnimmt, bis sie im

fünften und sechsten Paragraphen völlig fehlt und erst am Ende des Artikels wieder zum bestimmenden Stil wird.
Die deeskalationsorientierten Aspekte des Textes beschreiben den Friedensvertrag als eine Struktur, die geeignet ist, künftige Kooperation zu ermöglichen (10x); eine neue Perspektive, durch die Versöhnung einen Schritt näher rücken könnte (2x) und damit als eine Brücke in eine bessere Zukunft (2x). Win-Win-Orientierung (10x) und die Beschreibung von kooperativem Verhalten (5x) stehen im Vordergrund. Es wird deutlich, dass Zugeständnisse beider Seiten und vereinte Kräfte das Abkommen ermöglicht haben (5x) und der persönliche Einsatz beider Premierminister wird gleichermaßen (3x) honoriert (7x).
Ein geradezu entgegen gesetztes Bild bieten die eskalationsorientierten Textmerkmale, die erstmals im zweiten Paragraphen auftreten und im weiteren Verlauf des Textes zunehmen, bis sie im sechsten Paragraphen ihren Höhepunkt erreichen und zum Ende des Textes hin wieder verschwinden.
Diese eskalationsorientierten Aspekte negieren gemeinsame Interessen (5x) und Kooperationsmöglichkeiten der Konfliktparteien (9x) und präsentieren das Abkommen im Lichte einer Nullsummenorientierung. Es wird als keine wirkliche Lösung beschrieben (7x) und die Möglichkeit einer friedlichen Konfliktlösung wird direkt zurückgewisen oder indirekt in Frage gestellt (12x). Der Antagonismus zwischen den Konfliktparteien steht im Vordergrund und Übereinkünfte werden als Nachgeben interpretiert (10x). Auf der emotionalen Ebene wird dies durch Misstrauen gegen die andere Seite (1x) und die Dämonisierung ihrer Intentionen (1x) unterstrichen und positive (emotionale) Reaktionen auf den Friedensvertrag werden als Blauäugigkeit abgetan.
Damit lässt der Artikel zwar Sympathie für den Friedensvertrag erkennen, die er lesergerecht am Anfang und am Ende des Textes positioniert, die für gewöhnlich am besten erinnert werden. In der Sache jedoch wird der Friedensvertrag entwertet und somit unter dem Strich als eine zwar schöne Illusion, aber eben doch nur als eine Illusion porträtiert.

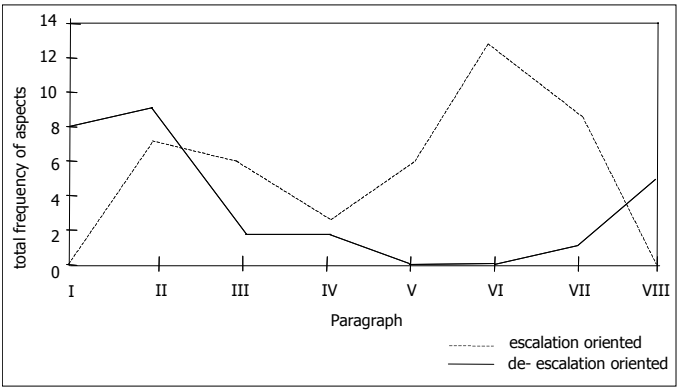


Abb. 1.4.2.3: Verteilung der eskalations- und deeskalationsorientierten Textmerkmale über den Artikel hinweg (nach ASPR, 2003, 91)

1.4.3 Materialbezogene Kategorienbildung

Explorative und inspektive Analyse

Je nachdem, ob zur Entwicklung des Kodierschemas bereits auf eine ausgearbeitete Theorie zurückgegriffen wird oder nicht, kann zwischen einer theoriebezogenen (im Falle der inspektiven Analyse) und einer materialbezogenen Kategorienbildung (im Falle der explorativen Analyse) unterschieden werden. Auch die materialbezogene Kategorienbildung ist jedoch nicht völlig frei von theoretischen Vorannahmen und erfordert zumindest einen groben Vorbegriff, welche Art von Textmerkmalen zur Beschreibung des zu analysierenden Textmaterials geeignet und in Hinblick auf die Fragestellung der Untersuchung relevant ist. Und unabhängig davon, welche Art von Kategorienbildung gewählt wird, erfolgt die Entwicklung des Kodierschemas stets anhand des Textmaterials. Der Unterschied zwischen den beiden Verfahrensweisen ist daher nicht grundsätzlicher, sondern eher gradueller Natur.

Unterschiede zwischen qualitativer und quantitativer Inhaltsanalyse

Während das Kategoriensystem bei der quantitativen Inhaltsanalyse anhand einer Teilstichprobe von Texten entwickelt und dann auf das gesamte Textmaterial angewendet wird, bleibt es bei der qualitativen Inhaltsanalyse bis zuletzt offen für Erweiterungen, Ergänzungen und/oder Modifikationen (s. o.).

Dimensionsanalyse

Zum Beispiel können wir die explorative Analyse unseres Textbeispiels *Widersinnig* (vgl. Kap. 1.3.1) damit beginnen, dass der Text zwei Arten von Aussagen enthält, die für die Analyse der Darstellung des Konfliktes um die Präsidentschaft der EZB relevant sind: (1) Handlungsbeschreibungen (z. B., dass der französische Präsident und sein Regierungschef einen eigenen Kandidaten für das Amt des EZB-Präsidenten benannt haben) und (2) Eigenschaftszuschreibungen (z. B., dass der Vorstoß des französischen Präsidenten und seines Regierungschefs ebenso ungeschickt wie illusorisch ist). Um den Text erschöpfend beschreiben zu können, müssen wir für beide Arten von Aussagen ein geeignetes Beschreibungssystem entwickeln.

Handlungsbeschreibungen: Um die Handlungsbeschreibungen zu identifizieren, die in dem Text enthalten sind, stellen wir an den Text die Frage:

- "Wer tut wem was?"

Unter dieser Frage listen wir alle Handlungsbeschreibungen auf, die der Text enthält.

Eigenschaftszuschreibungen: Um die im Text enthaltenen Eigenschaftszuschreibungen zu identifizieren, stellen wir an den Text die Frage:

- "Wem oder was wird welche Eigenschaft zugeschrieben?"

Unter dieser Frage listen wir alle Eigenschaftszuschreibungen auf, die der Text enthält.

Abgrenzung und Zuordnungsschwierigkeiten: Aussagen, die sich nicht eindeutig als Handlungsbeschreibungen oder Eigenschaftszuschreibungen identifizieren lassen, tragen wir in eine dritte Liste ein.

In der Praxis empfiehlt es sich, die Aussagen auf Karteikarten zu schreiben und anstelle von Listen Stapel von Karteikarten zu bilden, die sich später umsortieren lassen.

Nachdem wir diese Listen erstellt haben (vgl. Tab. 1.4.3.1), klären wir die Zweifelsfälle durch Formulierung von Prädikatorenregeln.

So können wir z. B. die Regeln formulieren, dass (1) substantivierte Handlungsverben als Handlungsbeschreibungen gelten, während (2) Aussagen der Form "x ist y" als Eigenschaftszuschreibungen gelten und (3) Eigenschaftszuschreibungen auch dann kodiert werden, wenn sie implizit formuliert sind (d. h. wenn die Formulierung wie "x ist y" nicht explizit im Text steht, aber sinngemäß erschlossen werden kann).

Daraus ergibt sich die Folgerung, dass die Aussage "Der Vorstoß ... ist ebenso ungeschickt wie illusorisch" sowohl unter den Handlungsbeschreibungen ("Der französische Präsident und sein Regierungschef haben einen Vorstoß unternommen") als auch unter den Eigenschaftsbeschreibungen ("Die Benennung eines eigenen Kandidaten ... ist ein Vorstoß des französischen Präsidenten und seines Regierungschefs") aufzulisten und die Aussage "Paris ist nationalegozentrisch ehrgeizig und eitel" in die Liste der Eigenschaftszuschreibungen aufzunehmen ist.

Weiters formulieren wir eine Regel, wonach (4) Verben wie 'besitzen' oder 'haben' keine Handlungsverben sind und daher als Eigenschaftszuschreibungen gelten, weshalb die Aussagen "Frankreich besitzt ... nicht die geringsten Erfolgsaussichten" und "[Frankreich] hat aber nicht die Gewissheit" unter den Eigenschaftszuschreibungen aufzulisten sind.

Tabelle 1.4.3.1: Liste der Handlungs- und Eigenschaftsbeschreibungen

Handlungsbeschreibungen: "Wer tut wem was?"
• Der französische Präsident und sein Regierungschef haben einen eigenen Kandidaten für das Amt des EZB-Präsidenten benannt. • Die Holländer stellen einen eigenen Kandidaten. • Paris befürchtet ein deutsches Übergewicht im währungspolitischen Bereich. • (Paris) unterstellt etwas: "Dem holländischen Kandidaten wird unterstellt ..." • (Frankreich) verhindert zwar die Wahl Duisenbergs ... • Paris dürfte auf Schwierigkeiten stoßen, wenn ... • Frankreich gefährdet seinen internationalen Ruf.
Eigenschaftszuschreibungen: "Wem oder was wird welche Eigenschaft zugeschrieben?"
• Der Vorstoß des französischen Präsidenten und seines Regierungschefs ist ebenso ungeschickt wie illusorisch. • Für die Ernennung (des EZB-Präsidenten) bedarf es der Einstimmigkeit. • Das Veto der Holländer gilt als sicher. • Implizit: Paris ist ein Wiederholungstäter: "es befürchtet tatsächlich wieder einmal" etwas. • Das von Paris befürchtete deutsche Übergewicht ist "angeblich erdrückend" • Der holländische Kandidat ist (laut Pariser Unterstellung) nur der Schatten des deutschen Bundesbankpräsidenten – und so ein Werkzeug der Bundesbank. • Das Übergewicht der Bundesbank ist imaginär.
Zweifelsfälle:
• *Handlungsbeschreibung*: Der französische Präsident und sein Regierungschef haben einen Vorstoß unternommen, oder *Eigenschaftszuschreibung*: die Benennung eines eigenen Kandidaten ist ein Vorstoß des französischen Präsidenten und seines Regierungschefs? • *Handlungsbeschreibung*: Frankreich *besitzt* ... nicht die geringsten Erfolgsaussichten, oder *Eigenschaftszuschreibung*: Frankreichs Initiative ist chancenlos? • *Handlungsbeschreibung*: [Frankreich] *hat* aber nicht die Gewissheit, oder *Eigenschaftszuschreibung*: Frankreichs Erfolgsaussichten sind ungewiss? • *Handlungsbeschreibung*: Frankreich [will] einem gefügigeren Kandidaten den Weg ebnen, oder eher *Motivzuschreibung*? • *Handlungsbeschreibung*: Paris [wird auch in Zukunft wieder] versuchen, Landsleute an die Spitze anderer europäischer und internationaler Einrichtungen zu stellen oder *Motivzuschreibung*? • *Eigenschaftszuschreibung*: "Nationalegozentrischer Ehrgeiz und Eitelkeiten sind weniger beliebt denn je", und/oder (implizit) "Paris ist nationalegozentrisch ehrgeizig und eitel" oder *Motivzuschreibung*: [Paris will seinen] nationalegoistischem Ehrgeiz und [seine] Eitelkeiten [befriedigen], oder beides, oder alles drei?

Aufgrund einer weiteren Regel, wonach (5) 'wollen' weder ein Handlungsverb noch eine Eigenschaftszuschreibung ist, sondern eine Motivzuschreibung, ergibt sich schließlich, dass wir zur Beschreibung des Textmaterials eine dritte Hauptdimension 'Motivzuschreibungen' einführen müssen:

Motivzuschreibungen: Um die im Text enthaltenen Motivzuschreibungen zu identifizieren, stellen wir an den Text die Frage:

- "Was will wer mit seiner Handlungsweise erreichen?"

und formulieren hierfür die Regel, wonach (6) Motivzuschreibungen auch dann kodiert werden, wenn sie implizit formuliert sind (d. h. wenn Worte wie 'wollen' nicht explizit im Text stehen, aber sinngemäß erschlossen werden können).

Entsprechend sind nun die Aussagen "Frankreich [will] einem gefügigeren Kandidaten den Weg ebnen", "Paris [wird auch in Zukunft wieder] versuchen, Landsleute an die Spitze anderer europäischer und internationaler Einrichtungen zu stellen, und "[Paris will seinen] nationalegoistischen Ehrgeiz und [seine] Eitelkeiten [befriedigen]" in die Liste der Motivzuschreibungen aufzunehmen.

Aufgrund einer weiteren Regel, wonach (7) auch antizipierte Handlungen unter die Handlungsbeschreibungen fallen, wird die Aussage "Paris [wird auch in Zukunft wieder] versuchen, Landsleute an die Spitze europäischer und internationaler Einrichtungen zu stellen" zugleich in die Liste der Handlungsbeschreibungen aufgenommen.

Zwischenergebnis der Dimensionsanalyse

Als Zwischenergebnis der Dimensionsanalyse können wir somit festhalten, dass der Text entlang der *drei Hauptdimensionen* (1) Handlungsbeschreibungen, (2) Eigenschaftszuschreibungen und (3) Motivzuschreibungen beschrieben werden soll. Unter den verschiedenen Hauptdimensionen sind jeweils alle Textstellen zu subsumieren, die Antwort auf die in Tab. 1.4.3.2 dargestellten Leitfragen geben:

Hauptdimension	Leitfrage
Handlungs-beschreibungen	"Wer tut wem was?"
Eigenschafts-beschreibungen	"Wem oder was wird welche Eigenschaft zugeschrieben?"
Motiv-zuschreibungen	"Was will wer mit seiner Handlungsweise erreichen?"

Tab. 1.4.3.2: Hauptdimensionen und Leitfragen der explorativen Analyse des Textbeispiels *Widersinnig*

Was als Antwort auf die Leitfragen zu gelten hat, lässt sich anhand einer Liste von materialnah gewonnenen Ankerbeispielen (vgl. Tab. 1.4.3.3) erläutern und durch Prädikatorenregeln absichern, die in Tab. 1.4.3.4 zusammengefasst sind.

Tab. 1.4.3.3: Liste der Ankerbeispiele

Aussage	Originaltextstelle
Beispiele für Handlungsbeschreibungen	
• Der französische Präsident und sein Regierungschefs haben einen Vorstoß unternommen (Regel 1) • Der französische Präsident und sein Regierungschef haben einen eigenen Kandidaten für das Amt des EZB-Präsidenten benannt.	Der Vorstoß des französischen Präsidenten und seines Regierungschefs, einen eigenen Kandidaten für das Amt des Präsidenten der europäischen Zentralbank zu benennen ist ebenso ungeschickt wie illusorisch (1/1/1)
• Die Holländer stellen einen eigenen Kandidaten.	Das Veto der Holländer, die einen eigenen Kandidaten stellen, gilt aber als sicher (1/1/3)
• Paris befürchtet ein deutsches Übergewicht im währungspolitischen Bereich.	Tatsächlich befürchtet Paris wieder einmal ein angeblich erdrückendes deutsches Übergewicht im währungspolitischen Bereich (1/2/1)
• [Paris] unterstellt etwas	Dem holländischen Kandidaten Wim Duisenberg wird unterstellt, nur der Schatten des deutschen Bundespräsidenten Hans Tietmeyer zu sein – und so ein Werkzeug der Bundesbank (1/2/2)
• [Frankreich] verhindert die Wahl Duisenbergs ...	Es verhindert zwar die Wahl Duisenbergs, hat aber nicht die Gewissheit, damit einem gefügigeren Kandidaten den Weg zu ebnen (1/3/2)
• Paris wird auf Schwierigkeiten stoßen ... • Paris [wird auch in Zukunft wieder] versuchen, Landsleute an die Spitze anderer europäischer und internationaler Einrichtungen zu stellen (Regel 7).	Außerdem dürfte Paris auf Schwierigkeiten stoßen, wenn es versucht, Landsleute an die Spitze anderer europäischer und internationaler Einrichtungen zu stellen (1/3/3)
• Frankreich gefährdet seinen internationalen Ruf.	Nicht zuletzt gefährdet Frankreich seinen internationalen Ruf (1/3/4)

Fortsetzung
Tab. 1.4.3.3: Liste der Ankerbeispiele

Beispiele für Eigenschaftszuschreibungen	
• Die Bennennung eines eigenen Kandidaten ist ein Vorstoß des französischen Präsidenten und seines Regierungschefs (Regeln 2 und 3) • Der Vorstoß des französischen Präsidenten und seines Regierungschefs ist ebenso ungeschickt wie illusorisch.	Der Vorstoß des französischen Präsidenten und seines Regierungschefs, einen eigenen Kandidaten für das Amt des Präsidenten der europäischen Zentralbank zu benennen ist ebenso ungeschickt wie illusorisch (1/1/1)
• Die Ernennung [des EZB-Präsidenten] bedarf der Einstimmigkeit.	Für die Ernennung bedarf es der Einstimmigkeit (1/1/2)
• Das Veto der Holländer gilt als sicher.	Das Veto der Holländer, die einen eigenen Kandidaten stellen, gilt aber als sicher (1/1/3)
• *Implizit*: *Paris ist ein Wiederholungstäter:* "es befürchtet tatsächlich wieder einmal" etwas. • Das von Paris befürchtete deutsche Übergewicht ist "angeblich erdrückend"	Tatsächlich befürchtet Paris wieder einmal ein angeblich erdrückendes deutsches Übergewicht im währungspolitischen Bereich (1/2/1)
• Der holländische Kandidat ist [laut Pariser Unterstellung] nur der Schatten Hans Tietmeyers – und so ein Werkzeug der Bundesbank. • Hans Tietmeyer ist Präsident der deutschen Bundesbank	Dem holländischen Kandidaten Wim Duisenberg wird unterstellt, nur der Schatten des deutschen Bundespräsidenten Hans Tietmeyer zu sein – und so ein Werkzeug der Bundesbank (1/2/2)
• Frankreich besitzt ... nicht die geringsten Erfolgsaussichten (Regel 4) • Das Übergewicht der Bundesbank ist imaginär.	Frankreich besitzt für seine Offensive gegen das imaginäre Übergewicht der Bundesbank nicht die geringsten Erfolgsaussichten (1/3/1)
• [Frankreich] hat keine Gewissheit (einem gefügigeren Kandidaten den Weg zu ebnen) (Regel 4)	Es verhindert zwar die Wahl Duisenbergs, hat aber nicht die Gewissheit, damit einem gefügigeren Kandidaten den Weg zu ebnen (1/3/2)
• Paris ist nationalegozentrisch ehrgeizig und eitel (Regel 3) • Nationalegoistischer Ehrgeiz und Eitelkeiten sind weniger denn je beliebt (Regel 2)	Denn nationalegoistischer Ehrgeiz und Eitelkeiten sind weniger denn je beliebt (1/3/5)

Fortsetzung
Tab. 1.4.3.3: Liste der Ankerbeispiele

Beispiele für Motivzuschreibungen	
• Frankreich [will] einem gefügigeren Kandidaten den Weg ebnen) (Regel 5 und 6)	Es verhindert zwar die Wahl Duisenbergs, hat aber nicht die Gewissheit, damit einem gefügigeren Kandidaten den Weg zu ebnen (1/3/2)
• [Paris will] Landsleute an die Spitze anderer europäischer und internationaler Einrichtungen stellen (Regel 5 und 6)	Außerdem dürfte Paris auf Schwierigkeiten stoßen, wenn es versucht, Landsleute an die Spitze anderer europäischer und internationaler Einrichtungen zu stellen (1/3/3)
• Paris [will seinen] nationalegoistischen Ehrgeiz und [seine Eitelkeiten befriedigen] (Regel 5 und 6)	Denn nationalegoistischer Ehrgeiz und Eitelkeiten sind weniger denn je beliebt (1/3/5)

Tab. 1.4.3.4: Prädikatorenregeln

Regel 0:	Dieselbe Textstelle kann unter mehrere Hauptdimensionen gleichzeitig fallen.
Regel 1:	Substantivierte Handlungsverben gelten als Handlungsbeschreibungen.
Regel 2:	Aussagen der Form "x ist y" gelten als Eigenschaftszuschreibungen.
Regel 3:	Eigenschaftszuschreibungen werden auch dann kodiert, wenn sie implizit formuliert sind (d. h. wenn die Formulierung wie "x ist y" nicht explizit im Text steht, aber sinngemäß erschlossen werden kann).
Regel 4:	Verben wie "besitzen", "haben" sind keine Handlungsverben und gelten daher als Eigenschaftszuschreibungen.
Regel 5:	"Wollen" ist weder ein Handlungsverb noch eine Eigenschaftszuschreibung, sondern eine Motivzuschreibung.
Regel 6:	Motivzuschreibungen werden auch dann kodiert, wenn sie implizit formuliert sind (d. h. wenn Worte wie "wollen" nicht explizit im Text stehen, aber sinngemäß erschlossen werden können).
Regel 7:	Antizipierte Handlungen fallen ebenfalls unter die Handlungsbeschreibungen.

Subdimensionen

In einigen Fällen kann es sinnvoll sein, die Beschreibungsdimensionen vor ihrer Operationalisierung durch Variable in Subdimensionen aufzuschlüsseln.

Z. B. können wir bei den Handlungsbeschreibungen a priori drei Subdimensionen unterscheiden: (1) Die Darstellung der *Subjekte* der Handlungen (*wer* tut etwas?), (2) die Darstellung der *Objekte* der Handlungen (*wem* wird etwas getan?), und (3) die Darstellung des *Inhalts* der Handlung (*was* wird getan?).

Bei den Eigenschaftszuschreibungen können wir zwei Subdimensionen unterscheiden: (1) Das *Objekt* der Eigenschaftszuschreibung (*wem oder was* wird eine Eigenschaft zugeschrieben?) und (2) den *Inhalt* der Eigenschaftszuschreibung (*welche* Eigenschaft wird ihm zugeschrieben?)

Bei den Motivzuschreibungen können wir ebenfalls zwei Subdimensionen unterscheiden: (1) Das *Objekt* der Motivzuschreibung (*wer* will mit seinen Handlungen etwas erreichen?) und (2) den Inhalt des Motivs (*was* will er damit erreichen?).

Operationalisierung der Subdimensionen durch Variablen

Jede dieser Subdimensionen können wir durch Variablen operationalisieren. Z. B. konstruieren wir ein Variablensystem zur Beschreibung der Akteure, die in dem Text vorkommen – sei als Subjekte oder Objekte von Handlungen, als Objekte von Eigenschaftszuschreibungen oder als Objekte von Motivzuschreibungen.

Dabei fertigen wir aufbauend auf die Liste der Ankerbeispiele zunächst wieder eine Liste der Akteure an (vgl. Tab. 1.4.3.5).

Tab. 1.4.3.5: Liste der Akteure, die in dem Text vorkommen

- Der französische Präsident
- Sein [Frankreichs] Regierungschef
- Ein eigener Kandidat Frankreichs
- Die Holländer
- Ein eigener [holländischer] Kandidat
- Paris
- Der holländische Kandidat Wim Duisenberg
- Der deutsche Bundesbankpräsident Hans Tietmeyer
- Die Bundesbank
- Frankreich
- Duisenberg
- Ein [Frankreich] gefügiger Kandidat
- Landsleute [Frankreichs]
- Andere europäische und internationale Einrichtungen

Die Einträge in der Liste können wir nach verschiedenen (relevant erscheinenden) Kriterien klassifizieren und so ein Variablensystem zur Beschreibung der Akteure erstellen, die in dem Text vorkommen. Z. B. klassifizieren wir die Akteure nach ihrer nationalen oder übernationalen Zugehörigkeit (Variable 1, vgl. Tab. 1.4.3.6), nach ihrem Vorkommen als Subjekt oder Objekt einer Handlung (Variable 2, vgl. Tab. 1.4.3.7), nach der Art ihrer Darstellung (Variable 3, vgl. Tab. 1.4.3.8) und nach der emotionalen Konnotation der Darstellung (Variable 4, vgl. Tab. 1.4.3.9).

Was die einzelnen Kategorien bedeuten und wie die Grenze zwischen den verschiedenen Kategorien zu ziehen ist, wird wieder anhand von Ankerbeispielen erläutert und durch die Formulierung von Prädikatorenregeln abgesichert. Aus Platzgründen wird dieser Analyseschritt jedoch hier nicht im Detail ausgeführt, sondern dem Leser als Übungsbeispiel überlassen.

Tab. 1.4.3.6: Klassifikation der Akteure nach ihrer nationalen oder übernationalen Zugehörigkeit

Kategorie	Beispiele
Frankreich	• Der französische Präsident • Sein [Frankreichs] Regierungschef • Ein eigener Kandidat Frankreichs • Paris • Frankreich • Ein [Frankreich] gefügiger Kandidat • Landsleute [Frankreichs]
Deutschland	• Der deutsche Bundesbankpräsident Hans Tietmeyer • Die Bundesbank
Niederlande	• Die Holländer • Ein eigener [holländischer] Kandidat • Der holländische Kandidat Wim Duisenberg • Duisenberg
Europäische Union	• Andere europäische und internationale Einrichtungen
Sonstige	

Tab. 1.4.3.7: Klassifikation der Akteure nach ihrem Vorkommen als Subjekt oder Objekt einer Handlung

Kategorie	Beispiele
Subjekt	• Der französische Präsident • Sein [Frankreichs] Regierungschef • Die Holländer • Paris • Frankreich
Objekt	• Ein eigener Kandidat Frankreichs • Ein eigener [holländischer] Kandidat • Der holländische Kandidat Wim Duisenberg • Duisenberg • Ein [Frankreich] gefügiger Kandidat • Landsleute [Frankreichs]
Weder noch	• Der deutsche Bundesbankpräsident Hans Tietmeyer • Die Bundesbank • Andere europäische und internationale Einrichtungen
Sowohl als auch	

Klasse	Beispiele	Regel
Politische Persönlichkeiten		
a) Namentlich genannt	• Duisenberg	
b) Über Funktion dargestellt	• Der französische Präsident • Sein Regierungschef • Ein eigener Kandidat Frankreichs • Ein [Frankreich] gefügiger Kandidat • Landsleute [Frankreichs]	
c) Sowohl als auch	• Der holländische Kandidat Wim Duisenberg • Der deutsche Bundesbankpräsident Hans Tietmeyer	
Die Nation	• Paris • Frankreich	Eliteorientierte Formulierung
Das Land	• Die Holländer	Menschenorientierte Formulierung
Institutionen	• Die Bundesbank • Andere europäische und internationale Einrichtungen	

Tab. 1.4.3.8: Klassifikation der Akteure nach der Art ihrer Darstellung

Klasse		Beispiele	Regel
Positiv/ Nähe	Affirmativ	• Der holländische Kandidat Wim Duisenberg • Der deutsche Bundesbankpräsident Hans Tietmeyer	Humanisierung
	Negierend		
Neutral		• Duisenberg • Der französische Präsident • Ein eigener Kandidat Frankreichs • Paris • Frankreich • Die Holländer • Die Bundesbank • Andere europäische und internationale Einrichtungen	
Negativ/ Distanz	Affirmativ	• "Sein" Regierungschef • Ein [Frankreich] gefügiger Kandidat • [Bloße] Landsleute [Frankreichs]	
	Negierend		

Tab. 1.4.3.9: Klassifikation der Akteure nach der emotionalen Konnotation ihrer Darstellung

In ähnlicher Weise können wir auch Variablensysteme zur Klassifikation des Inhaltes der im Text dargestellten Handlungen, zugeschriebenen Eigenschaften und unterstellten Handlungsmotive entwickeln. Auch dies wird hier nicht im Detail ausgeführt, sondern bleibt dem Leser als Übungsbeispiel überlassen.

Unterschiede zwischen qualitativer und quantitativer Inhaltsanalyse

In der qualitativen Inhaltsanalyse werden die Variablen und Kategorien des Variablensystems als Beschreibungstermini verwendet, mittels derer der Inhalt des Textmaterials verbal beschrieben wird und die verschiedenen Variablen argumentativ zueinander in Beziehung gesetzt werden.

In der quantitativen Inhaltsanalyse werden die Kategorien der Variablen durch Zuordnung der Zahlen 0,1,2,3… auf Nominalskalen abgebildet und anschließend statistisch weiterverarbeitet.

Tab. 1.4.3.10: Kodierschema zur Klassifikation der Akteure

Nr	Variable	Kategorie		Code
1	Nationale/übernationale Zugehörigkeit	Frankreich		0
		Deutschland		1
		Niederlande		2
		Europäische Union		3
		Sonstige		4
2	Vorkommen als Subjekt oder Objekt einer Handlung	Subjekt		0
		Objekt		1
		Weder noch		2
		Sowohl als auch		3
3	Art der Darstellung	Politische Persönlichkeiten	Namentlich genannt	0
			Über Funktion dargestellt	1
			Sowohl als auch	2
		Die Nation		3
		Das Land		4
		Institutionen		5
4	Emotionale Konnotation	Positiv/ Nähe	Affirmativ	0
			Negierend	1
		Neutral		2
		Negativ/ Distanz	Affirmativ	3
			Negierend	4

Quantitative Kodierschemata

Das Kodierschema muss daher so gestaltet sein, dass jede Kodiereinheit bezüglich jeder Variable in genau eine Kategorie fällt (s.o.) und jede Kodiereinheit am Ende durch ein eindeutiges *Codemuster* beschrieben wird. Um dies zu gewährleisten, wird die endgültige Form des Kodierschemas davon abhängen, wie wir die Kodiereinheiten definiert haben.

Wählen wir als *Kodiereinheiten* z. B. die im Text vorkommenden *Akteure*, dann kann das Kodierschema z. B. so aussehen wie in Tab. 1.4.3.10.

Die Codematrix

Das Codemuster einer Kodiereinheit hat dann die Form eines Zahlenvektors, der in der Reihenfolge der Variablen 1,2,3,… für jede Variable eine Zahl 0,1,2,… enthält (d. h. den Code der Kodiereinheit bezüglich dieser Variablen). Entsprechend wird der analysierte Textkörper (die Gesamtheit der analysierten Kodiereinheiten) durch eine Codematrix beschrieben, wie sie in Tab. 1.4.3.11 dargestellt ist.

Text	Kodiereinheit	Variable Nr.				Wortlaut
Nr.	Nr.	1	2	3	4	
111	001	0	0	1	2	Der französische Präsident
111	002	0	0	1	3	Sein Regierungschef
111	003	0	1	1	2	Ein eigener Kandidat
113	004	2	0	4	2	Die Holländer
113	005	2	1	1	2	Ein eigener Kandidat
121	006	0	0	3	2	Paris
122	007	2	1	2	0	Der holländische Kandidat Wim Duisenberg
usw.	usw.					

Tab. 1.4.3.11: Die Codematrix (doppelt umrandet) Text-Nr. = Text (1=Widersinnig)/Absatz/Satz

Sätze, Absätze oder ganze Texte als Kodiereinheiten

Wenn wir nicht die Akteure, sondern ganze Sätze, Absätze oder ganze Texte als Kodiereinheiten verwenden, ist das oben dargestellte Kodierschema nicht geeignet, da in derselben Kodiereinheit (im selben Satz, Absatz oder Text) mehrere Akteure vorkommen können. Wir können dann aber z. B. ein Kategoriensystem entwickeln, das für jede Nation und/oder übernationale Organisation ihr Vorkommen als Subjekt oder Objekt, die Art der Darstellung der Akteure und die emotionale Konnotation der Variablen durch mehrere binäre Variablen operationalisiert (vgl. Tab. 1.4.3.12).

Tab. 1.4.3.12: Binäres Kodierschema zur Analyse von Sätzen, Absätzen oder ganzen Texten

Variablen-Nr.			Definition	Code	Definition
1			Französische Akteure		
	1.1		Als Subjekt oder Objekt		
		1.1.1	Als Subjekt	0	Kommt nicht vor
				1	Kommt vor
		1.1.2	Als Objekt	0	Kommt nicht vor
				1	Kommt vor
	1.2		Art der Darstellung		
		1.2.1	Namentlich genannte politische Personen	0	Kommt nicht vor
				1	Kommt vor
		1.2.2	Über Funktion genannte politische Personen	0	Kommt nicht vor
				1	Kommt vor
		1.2.3	Die Nation	0	Kommt nicht vor
				1	Kommt vor
		1.2.4	Das Land	0	Kommt nicht vor
				1	Kommt vor
		1.2.5	Institutionen	0	Kommt nicht vor
				1	Kommt vor
	1.3		Emotionale Konnotation der Darstellung		
		1.3.1	Positiv/Nähe affirmativ	0	Kommt nicht vor
				1	Kommt vor
		1.3.2	Positiv/Nähe negierend	0	Kommt nicht vor
				1	Kommt vor
		1.3.3	Negativ/Distanz affirmativ	0	Kommt nicht vor
				1	Kommt vor
		1.3.4	Negativ/Distanz negierend	0	Kommt nicht vor
				1	Kommt vor
2			Deutsche Akteure		
	2.1 bis 2.3	2.1.1 bis 2.3.4	Analog zu oben		
3			Niederländische Akteure		
	3.1 bis 3.3	3.1.1 bis 3.3.4	Analog zu oben		

4			Europäische Union
	4.1 bis 4.3	4.1.1 bis 4.3.4	Analog zu oben
5			Sonstige Akteure
	5.1 bis 5.3	5.1.1 bis 5.3.4	Analog zu oben

Fortsetzung
Tab. 1.4.3.12: Binäres Kodierschema zur Analyse von Sätzen, Absätzen oder ganzen Texten

Die Codematrix sieht dann aus wie in Tab. 1.4.3.13 dargestellt und die Eintragung in der Codematrix sind lauter Nullen und Einsen.

Nr.	Variable Nr.								
	Französische Akteure								usw. Deutschl. bis sonstige
	1.1.1	1.1.2	1.2.1	...	1.2.5	1.3.1	...	1.3.4	usw. 2.1.1 bis 5.3.4
001									usw. Codes
002									usw. Codes
003									usw. Codes
004									usw. Codes
usw.	usw.								

Tab. 1.4.3.13: Die zugehörige Codematrix
Nr. = Nummer der Kodiereinheit

Die Operationalisierung durch binäre Variable ist jedoch nicht zwingend. Wir können auch polytome Variablen bilden. So könnten wir z. B. die emotionale Konnotation statt durch vier binäre Variablen auch durch zwei vierstufige Variablen operationalisieren, wie sie in Tab. 1.4.3.14 dargestellt sind.

Zusammenlegung von Kategorien und/ oder Variablen

Wenn einzelne Kategorien bestimmter Variablen für eine statistische Weiterverarbeitung zu selten vorkommen, kann man Kategorien zusammenlegen, wie wir dies z. B. in Tab. 1.4.3.15 mit den Kategorien der Variablengruppe #.2 (Art der Darstellung) getan haben.

Alternativ dazu man kann auch Variablen zusammenlegen. Ein Beispiel dafür gibt Tab. 1.4.3.16, in der die ursprünglichen Variablen der Variablengruppe #.3 (Emotionale Konnotation der Darstellung) zu einer geringeren Zahl von dichotomen Variablen zusammengelegt wurden.

Tab. 1.4.3.14: Operationalisierung der emotionalen Konnotation durch polytome Variablen

Variablen-Nr		Definition	Code	Definition
#.3		Emotionale Konnotation der Darstellung		
	#.3.1	Positiv/Nähe	0	Kommt nicht vor
			1	Affirmativ
			2	Negativ
			3	Sowohl als auch
	#.3.2	Negativ/Distanz	0	Kommt nicht vor
			1	Affirmativ
			2	Negativ
			3	Sowohl als auch

Tab. 1.4.3.15: Zusammenlegung von Kategorien

Variablen-Nr		Definition	Code	Definition
#.2		Art der Darstellung		
	#.2.1	Namentlich genannte und/oder über ihre Funktion dargestellte politische Persönlichkeiten	0	Kommt nicht vor
			1	Kommt vor
	#.2.2	Nationen, Länder und/oder Institutionen	0	Kommt nicht vor
			1	Kommt vor

Tab. 1.4.3.16: Dichotome Variablenzusammenlegung

Variablen-Nr		Definition	Code	Definition
#.3		Emotionale Konnotation der Darstellung		
	#.3.1	Positive emotionale Konnotation = Positiv/Nähe affirmativ oder Negativ/Distanz negativ	0	Kommt nicht vor
			1	Kommt vor
	#.3.2	Negative emotionale Konnotation = Negativ/Distanz affirmativ oder Positiv/Nähe negativ	0	Kommt nicht vor
			1	Kommt vor

Auch dabei ist man aber nicht daran gebunden, dass unbedingt binäre Variablen zu bilden sind. Wie das Beispiel in Tab. 1.4.3.17 zeigt, können wir bei der Zusammenlegung von Variablen auch polytome Variablen bilden. Arbeitstechnisch ist es sowohl für den Kodiervorgang als auch für die spätere Zusammenlegung von Variablen jedoch am günstigsten, wenn man zunächst binäre Variablen bildet.

Variablen-Nr		Definition	Wert	Definition
#.3		Emotionale Konnotation der Darstellung		
	#.3.1	Emotionale Konnotation der Darstellung	0	Weder noch (= neutral)
			1	Positiv
			2	Negativ
			3	Sowohl als auch (= ambivalent)

Tab. 1.4.3.17: Polytome Variablenzusammenlegung

1.4.4 Theoriebezogene Kategorienbildung

Vorbegriff und Theoriebezug der Kategorienentwicklung

Schon bei der materialbezogenen Kategorienbildung war es nicht wirklich so, dass wir "den Inhalt (des Textes) selbst sprechen lassen" haben, wie Lamnek (1989, 194f) dies sieht. Zwar haben wir "aus ihm [dem Text, W. K.] heraus die Analyse entfaltet", dies ist jedoch aufgrund von wissenschaftlichem und alltagsweltlichem Vorwissen geschehen, d. h. aufgrund des Vorbegriffs, den wir bereits hatten: bezüglich der Textart (Zeitungskommentar), bezüglich des Textgegenstandes (Konflikt), bezüglich des Konfliktbegriffs (Unverträglichkeit der Handlungen oder Ziele zweier oder mehrerer Konfliktparteien), bezüglich einer möglichen Parteilichkeit von Zeitungskommentaren bis hin zur Anheizung des Konfliktes usw.

All dies sind Dinge, die wir bereits im Hinterkopf hatten und aufgrund derer uns erst die Dinge aufgefallen sind, zu deren Beschreibung wir das Kategoriensystem entwickelt haben. Texte sprechen nicht einfach "für sich", sondern sie antworten auf die Fragen, die wir an sie herantragen; und wenn es denn schon so ist, dass man zumindest einen Vorbegriff braucht, damit man Fragen an den Text stellen kann und damit der Text "zu einem spricht", dann spricht auch nichts dagegen, diese Fragen so fundiert wie möglich zu stellen; oder anders ausgedrückt: Wenn wir bereits über eine fundierte Theorie des Gegenstandsbereichs der Analyse verfügen, dann ist es angemessen, das inhaltsanalytische Kategoriensystem aus dieser Theorie heraus zu entwickeln, so dass die Ergebnisse der Analyse dann wieder präzise auf die Theorie zurückbezogen werden können.

Materialbezug der Kategorienentwicklung

Auch dabei können wir das Kodierschema jedoch nicht einfach aus der Theorie ableiten, ohne zuvor das Textmaterial einzusehen (vgl. Friedrich, 1973, 321), sondern auch bei der inspektiven Analyse müssen wir materialnahe Beschreibungskategorien ent-

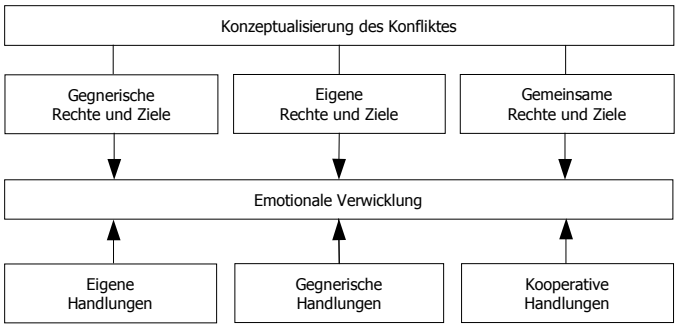
wickeln und untersuchen, wie die theoretischen Begriffe sich in der zu analysierenden Art von Texten abbilden.

Als Beispiel, anhand dessen wir die Vorgangsweise demonstrieren, verwenden wir wieder den in Abschnitt 1.3.1 vorgestellten Text *Widersinnig* und entwickeln ein Kategoriensystem, mittels dessen wir das mentale Modell rekonstruieren wollen, nach welchem der Konflikt konstruiert wird.

Ein kognitives Eskalationsmodell

Den theoretischen Ausgangspunkt hierfür bilden die Konfliktheorie von Morton Deutsch (1973), Friedrich Glasls (1992) Stufenmodell der Konflikteskalation und das von Kempf (2000) daraus abgeleitete kognitive Eskalationsmodell, welches entlang der Dimensionen (1) der Konzeptualisierung des Konfliktes, (2) der Bewertung der Rechte und Ziele der Konfliktparteien, (3) der Bewertung ihrer Handlungen und (4) der daraus resultierenden emotionalen Verwicklung (vgl. Abb. 1.4.4.1) beschreibt, wie sich das mentale Modell eines Konfliktes im Verlaufe der Konflikteskalation verändert.

Abb. 1.4.4.1: Die Struktur des Eskalationsmodells



Ausgangspunkt des Modells ist die Auffassung, dass Konflikte grundsätzlich dafür offen sind, entweder als kooperativer (Win-Win) oder als kompetitiver (Win-Lose) Prozess konzeptualisiert zu werden. Wegen der Perspektivendivergenz der Konfliktparteien besteht jedoch ein gewisser Bias zugunsten einer Win-Lose-Interpretation. Wird der Konflikt kompetitiv ausgetragen, so entwickelt er die Tendenz, sich auszubreiten und hochzuschaukeln. Die Konkurrenz schlägt um in Kampf, der schließlich – wenn ihm nicht Einhalt geboten wird – im Krieg kulminiert. Den Motor dieser Konflikteskalation bilden kompetitive Fehlwahrnehmungen der Konfliktparteien, welche die Konfliktwahrnehmung auf jeder

der genannten Dimensionen polarisieren und radikalisieren. Diese Wahrnehmungsverzerrungen sind in Tab. 1.4.4.1 zusammengefasst und charakterisieren jede der Eskalationsstufen durch ein entsprechendes (typisches) mentales Konfliktmodell.

Die Eckpunkte des Eskalationsprozesses

Zur Konstruktion eines inhaltsanalytischen Kodierschemas, welches geeignet ist, solche Wahrnehmungsverzerrungen abzubilden und die zugrundeliegenden mentalen Modelle zu rekonstruieren, können wir im Anschluss an Kempf, Reimann & Luostarinen (1996) so vorgehen, dass wir zunächst die beiden Eckpunkte des Eskalationsprozesses durch Variablen beschreiben, die aus der Theorie abgeleitet sind (vgl. Tab. 1.4.4.2). Dabei bildet Kooperation den deeskalationsorientierten und Krieg den eskalationsorientierten Pol des Variablenschemas.

Operationalisierung durch Ankerbeispiele und Prädikatorenregeln

Um von der theoretischen Beschreibung der Variablen zu einem Kodierschema zu gelangen, welches auf das zu analysierende Textmaterial anwendbar ist, muss die Bedeutung der Variablen wieder durch Ankerbeispiele exemplifiziert und durch die Formulierung von Prädikatorenregeln abgesichert werden. D. h. wir müssen untersuchen, wie sich die Variablen in den Texten manifestieren.

Dafür können wir zwar in einem ersten Schritt auf den von Bläsi et al. (2004) erstellten Katalog von Ankerbeispielen zurückgreifen, der aber auf den hier zu analysierenden Text nur bedingt anwendbar ist. Während wir es mit der Darstellung eines politischen Konfliktes zu tun haben, der sich weit unterhalb der Schwelle zur Gewalt bewegt, wurde der Katalog von Bläsi et al. auf Grundlage der Kriegsberichterstattung über den Golfkrieg (Reimann, 1997) und die Nachkriegsberichterstattung über Frankreich nach dem Zweiten Weltkrieg (Jaeger, 2005) sowie über Jugoslawien nach Milosević (Bläsi et al., 2005) erstellt.

Die Art und Weise, wie sich die abstrakten theoretischen Kategorien in konkretem Textmaterial manifestieren, hängt sowohl von der Art des Textes als auch von der Art des Konfliktes ab. So manifestieren sich die kompetitiven Wahrnehmungsverzerrungen im interpersonalen Streitgespräch (vgl. Lang, 1998) anders als in der Medienberichterstattung über Konflikte, und in der Berichterstattung über politische Konflikte (vgl. Plontz, 2005) anders als in der Kriegs- und Nachkriegsberichterstattung. Das Kodierschema muss daher stets an die jeweilige Textart *und* an die Art des untersuchten Konfliktes angepasst werden.

Tab. 1.4.4.1: Verzerrung der Konfliktwahrnehmung im Laufe der Eskalation von Konflikten (nach PFK, 2005, 20)

Eskalationsstufe	Kooperation	Perspektivendivergenz
Konzeptualisierung des Konfliktes	Win-Win-Orientierung	Bias in Richtung Win-Lose, aber Win-Win noch möglich
Bewertung von Rechten und Zielen	Gegenseitige Respektierung der Rechte aller Beteiligten und Betonung gemeinsamer Interessen	Fokussierung eigener Rechte und Bedürfnisse (einschließlich gemeinsamer Interessen), die Rechte des anderen verschwinden jedoch aus dem Blickfeld
Bewertung von Handlungen	Berücksichtigung des Nutzens für jede der Parteien	Fokussierung des eigenen Nutzens (auch aus der gemeinsamen Beziehung)
Emotionale Verwicklung	Empathie und gegenseitiges Vertrauen	Konflikt zwischen Bedrohung und Vertrauen
Identifikationsangebot	Allseitig	Selbstzentriert

Konkurrenz	Kampf	Krieg
Win-Lose (ggf. entschärft durch Regeln der Fairness)	Win-Lose (verschärft durch Drohstrategien)	Nullsummenorientierung, Gewalt als angemessenes Mittel der Konfliktlösung, Betonung militärischer Werte, Umschlagen von Win-Lose in Lose-Lose
Fokussierung eigener Rechte und Bedürfnisse; die gemeinsamen Interessen verschwinden jedoch aus dem Blickfeld	Betonung der eigenen Rechte und Bedürfnisse, bei gleichzeitiger Infragestellung der Rechte des Gegners und Verurteilung seiner Intentionen	Idealisierung der eigenen Rechte und Bedürfnisse, bei gleichzeitiger Bestreitung der Rechte des Gegners, Dämonisierung seiner Intentionen und Leugnung gemeinsamer Interessen
Fokussierung des eigenen Nutzens	Rechtfertigung eigener und Verurteilung der gegnerischen Handlungen	Idealisierung eigener Handlungen und Dämonisierung der Handlungen des Gegners
Fokussierung der eigenen Bedrohung, die des Gegners verschwindet jedoch aus dem Blickfeld, gegenseitiges Vertrauen geht verloren	Betonung der eigenen Stärke und der Gefährlichkeit des Gegners schaffen ein labiles Gleichgewicht zwischen Bedrohung und Siegeszuversicht, die Bedrohung des Gegners wird dagegen aktiv geleugnet, es entsteht Misstrauen	Gleichgewicht zwischen Bedrohung und Siegeszuversicht besteht weiter, Misstrauen richtet sich auch gegen neutrale Drittparteien, die in dem Konflikt zu vermitteln versuchen, die Empörung über den Krieg schlägt um in Empörung über den Gegner
Dualistisch	Antagonistisch	Polarisiert

Tab. 1.4.4.1: Verzerrung der Konfliktwahrnehmung im Laufe der Eskalation von Konflikten (nach PFK, 2005, 20)

Tab. 1.4.4.2: Das Variablenschema

	Eskalationsorientierter Pol		Deeskalationsorientierter Pol	
1. Konzeptualisierung des Konfliktes				
	E 1	Polarisierung	D 1	Depolarisierung
	E 1.1	Interpretation des Konfliktes als Win-Lose oder Lose-Lose-Modell und/oder Infragestellen von Win-Win	D 1.1	Interpretation des Konfliktes als Win-Win-Modell und/oder Infragestellen von Win-Lose und/oder Lose-Lose
2. Bewertung der Rechte und Ziele der Konfliktparteien				
	E 2	Antagonismus	D 2	Balance
	E 2.1	Dämonsierung des Gegners, Absprechen von Rechten oder Dämonisierung seiner Intentionen	D 2.1	Respektierung der Rechte des Gegners und/oder unverzerrte Darstellung seiner Intentionen
	E 2.2	Idealisierung eigener Rechte und Intentionen	D 2.2	Realistische und selbstkritische Bewertung der eigenen Rechte und Intentionen
	E 2.3	Bestreiten gemeinsamer Interessen und/oder Betonung der Unvereinbarkeit der Interessen	D 2.3	Betonung gemeinsamer Interessen und/oder des Gewinnes, den beide Seiten aus einem kooperativen Konfliktaustrag ziehen könnten
3. Bewertung der Handlungen der Konfliktparteien				
	E 3	Konfrontation	D 3	Kooperation
	E 3.1	Rechtfertigung eigener Handlungen und/oder Betonung ihrer Korrektheit	D 3.1	Realistische und selbstkritische Bewertung der eigenen Handlungen
	E 3.2	Verurteilung gegnerischer Handlungen	D 3.2	Realistische und faire Bewertung der generischen Handlungen
4. Emotionale Verwicklung				
	E 4	Destruktive Emotionen	D 4	Konstruktive Emotionen
	E 4.1	Siegeszuversicht	D 4.1	Vertrauen in die Möglichkeit kooperativer Konfliktlösung
	E 4.2	Argwohn gegenüber dem Gegner und/oder neutralen Drittparteien	D 4.2	Vertrauen gegenüber dem Gegner und/oder neutralen Drittparteien

Offenheit des Kodierschemas

Bei der Suche nach Ankerbeispielen im Textmaterial können wir zwar wieder Satz für Satz vorgehen, müssen die Sätze aber in ihrem Kontext verstehen. Mitunter werden wir eine Variable auch nicht auf einen einzelnen Satz, sondern auf eine längere Textstelle beziehen müssen. Gleichzeitig halten wir das Kodierschema offen für Erweiterungen, indem wir Textstellen notieren, die uns für die Beschreibung der Eskalations- vs. Deeskalationsorientierung des Textes relevant erscheinen, von unserem bisherigen Variablensystem aber noch nicht erfasst werden. Die so gesammelten Textstellen können wir dann wieder klassifizieren und in unser Variablensystem einbauen und auf diese Weise auch unsere ursprüngliche Theorie um neue Aspekte erweitern.

Eigene, gegnerische und dritte Partei

Sowohl bei der Operationalisierung der Variablen durch Ankerbeispiele als auch bei der späteren Anwendung des Kodierschemas auf das Textmaterial stoßen wir auf die Schwierigkeit, dass die Variablen zwischen der je eigenen, der gegnerischen und ggf. dritten Parteien unterscheiden. Wer diese sind und wie die Trennlinie zwischen den Parteien gezogen wird, stellt sich aber erst am Ende der Inhaltsanalyse des Textes heraus. Insbesondere bei längeren Texten kann es sogar sein, dass die Konfliktlinie sich im Laufe des Textes verändert.

Sowohl beim Notieren der Ankerbeispiele als auch bei der Kodierung des Textmaterials werden wir daher auf diese Begrifflichkeit (eigene, gegnerische und dritte Partei) zunächst verzichten und die jeweilige Partei stattdessen namentlich identifizieren.

Dies hat zur Folge, dass z. B. dieselbe Textstelle entweder als Verurteilung gegnerischer Handlungen (Variable E 3.2) oder als selbstkritische Bewertung der eigenen Handlungen (Variable D 3.1) zu kodieren sein kann. Wir notieren sie daher zunächst unter beiden Variablen und entscheiden später, wenn sich die Parteilichkeit des Textes herausgestellt hat, ob sie auf dem deeskalationsorientierten Pol (Variable D 3.1) oder auf dem eskalationsorientierten Pol (Variable E 3.2) zu kodieren ist.

Dies ist insbesondere dann erforderlich, wenn der Text – wie z. B. bei der Analyse der deutschen Presseberichterstattung über den Krieg in Bosnien – einen Konflikt zum Gegenstand hat, an dem der Berichterstatter und/oder die soziale Gruppe, Institution oder Nation, welcher er angehört, (noch) nicht unmittelbar beteiligt ist, denn auch in diesem Fall muss damit gerechnet werden, dass der Bericht nicht wirklich aus der Perspektive eines neutralen Dritten erfolgt, sondern zugunsten der einen oder anderen Partei verzerrt ist.

Im Falle unseres Beispieltextes *Widersinnig*, der einer deutschen Zeitung entstammt und einen Konflikt zwischen Deutschland und Frankreich zum Gegenstand hat, mag dies auf den ersten Blick als relativ unproblematisch erscheinen, aber auch hier wird sich erst am Ende der Analyse herausstellen, wie die Konfliktlinien gezogen werden und ob z. B. Holland oder die EU als unbeteiligte Dritte oder als Koalitionspartner Deutschlands porträtiert werden.

Die Kodierung des Beispieltextes

Aus Platzgründen verzichten wir im Folgenden auf die Darstellung der Beispiele für die einzelnen Variablen, sondern geben lediglich die zu den einzelnen Textstellen zugehörigen Codes wieder (vgl. Tab. 1.4.4.3). Die Produktion einer nach Variablen geordneten Liste von Textbeispielen sei dem Leser als Übungsaufgabe überlassen.

Tab. 1.4.4.3: Eskalations- und deeskalationsorientierte Aspekte des Beispieltextes

Abs.	Satz	Wortlaut und Codierung
Überschrift		Widersinnig
		• Verurteilung französischer (= gegnerischer) Handlungen (hier: als "widersinnig") (E3.2)
1	1	Der Vorstoß des französischen Präsidenten und seines Regierungschefs, einen eigenen Kandidaten für das Amt des Präsidenten der europäischen Zentralbank zu benennen ist ebenso ungeschickt wie illusorisch.
		• Verurteilung französischer (= gegnerischer) Handlungen (hier: als ungeschickter und illusorischer "Vorstoß") (E3.2) • Deutsche (= eigene) Siegeszuversicht (hier: Vorstoß als "illusorisch") (E4.1)
	2	Für die Ernennung bedarf es der Einstimmigkeit.
		• Keine Kodierung
	3	Das Veto der Holländer, die einen eigenen Kandidaten stellen, gilt aber als sicher.
		• Deutsche (= eigene) Siegeszuversicht (hier: Frankreich hat keine Chance) (E4.1)
2	1	Tatsächlich befürchtet Paris wieder einmal ein angeblich erdrückendes deutsches Übergewicht im währungspolitischen Bereich.
		• Dämonisierung Frankreichs (= des Gegners), Absprechen von Rechten oder Dämonisierung seiner Intentionen (hier: "wieder einmal", implizit: "die sind ja paranoid") (E2.1) • Idealisierung deutscher (= eigener) Rechte und Intentionen (hier: Zurückweisung des Übergewichts als "angeblich erdrückend") (E2.2) • Verurteilung französischer (= gegnerischer Handlungen) (hier: Befürchtungen als gegenstandslos) <implizit> (E3.2)

Fortsetzung
Tab. 1.4.4.3: Eskalations- und deeskalationsorientierte Aspekte des Beispieltextes

2	2	Dem holländischen Kandidaten Wim Duisenberg wird unterstellt, nur der Schatten des deutschen Bundesbankpräsidenten Hans Tietmeyer zu sein – und so ein Werkzeug der Bundesbank.
		• Idealisierung deutscher (= eigener) Rechte und Intentionen (hier: Unterstellungen sind gegenstandslos) <implizit> (E2.2) • Verurteilung französischer (= gegnerischer) Handlungen (hier: ungerechtfertigte Unterstellung) (E3.2)
3	1	Frankreich besitzt für seine Offensive gegen das imaginäre Übergewicht der Bundesbank nicht die geringsten Erfolgsaussichten.
		• Dämonisierung Frankreichs (= des Gegners), Absprechen von Rechten oder Dämonisierung seiner Intentionen (hier: Ziel der Offensive als "imaginär") (E2.1) • Idealisierung deutscher (= eigener) Rechte und Intentionen (hier: Übergewicht als "imaginär") (E2.2) • Verurteilung französischer (= gegnerischer) Handlungen (hier: als "Offensive") (E3.2) • Deutsche (= eigene) Siegeszuversicht (hier: "nicht die geringsten Erfolgsaussichten") (E4.1)
	2	Es verhindert zwar die Wahl Duisenbergs, hat aber nicht die Gewissheit, damit einem gefügigeren Kandidaten den Weg zu ebnen.
		• Konstruktion des Konfliktes nach Lose-Lose-Modell (E1.1) • Dämonisierung Frankreichs (= des Gegners), Absprechen von Rechten oder Dämonisierung seiner Intentionen (hier: die wollen einem gefügigen Kandidaten den Weg ebnen) (E2.1) • Eigene (= deutsche) Siegeszuversicht (hier: die können keine Gewissheit haben, dass sie es schaffen. Impliziert: sie werden es nicht schaffen) (E4.1)
	3	Außerdem dürfte Paris auf Schwierigkeiten stoßen, wenn es versucht, Landsleute an die Spitze anderer europäischer und internationaler Einrichtungen zu stellen.
		• Eigene (= deutsche) Siegeszuversicht (hier: die dürften Schwierigkeiten damit bekommen) (E4.1) • Argwohn gegenüber Frankreich (= dem Gegner) (hier: die wollen [aus national-egozentrischem Ehrgeiz, und Eitelkeit, 1/3/5] womöglich auch noch [ihnen gefügige, 1/3/2] Landsleute an die Spitze anderer Internationaler Einrichtungen stellen (E4.2)
	4	Nicht zuletzt gefährdet Frankreich seinen internationalen Ruf
		• Deutsche (= eigene) Siegeszuversicht (hier: Paris wird damit seinen Ruf verlieren) (E4.1)

Fortsetzung
Tab. 1.4.4.3: Eskalations- und deeskalationsorientierte Aspekte des Beispieltextes

3	5	Denn nationalegozentrischer Ehrgeiz und Eitelkeiten sind weniger denn je beliebt.
		• Dämonisierung Frankreichs (= des Gegners), Absprechen von Rechten oder Dämonisierung seiner Intentionen (hier: als nationalegozentrischer Ehrgeiz und Eitelkeit (E2.1)
Unterschrift		Alfred Frisch, Paris
		• Keine Kodierung

Die Interpretation des Analyseergebnisses

Zur Interpretation des Analyseergebnisses gehen wir wieder von den einzelnen Textstellen, welche wir kodiert haben (interner Mikro-Kontext), zum Text als Ganzem (interner Makro-Kontext), zum externen Kontext und zurück. Dabei zeigt sich zunächst, dass die Konfliktlinie zwischen Deutschland und Frankreich gezogen wird. Neutrale Drittparteien kommen nicht vor. Holland und die EU finden zwar Erwähnung, aber nur als Unterstützer der deutschen Position.

Obwohl keine einzelne Textstelle als Win-Lose kodiert wurde, konzeptualisiert der Text insgesamt den Konflikt im Sinne eines Win-Lose-Modells, das durch Drohstrategien verschärft ist (vgl. Abb. 1.4.4.2). Paris wird auf Schwierigkeiten stoßen, wenn es versucht, andere europäische und internationale Spitzenpositionen mit Landsleuten zu besetzen (3/3) und riskiert, seinen internationalen Ruf zu verlieren (3/4). An einer Stelle (3/2) wird sogar ein Umschlagen des Konfliktes in einen Lose-Lose-Prozess angedroht: "[Frankreich] verhindert zwar die Wahl Duisenbergs, hat aber nicht die Gewissheit, damit einem gefügigeren Kandidaten den Weg zu ebnen". Nur einer kann gewinnen, entweder Deutschland oder Frankreich. Gemeinsame Interessen werden nicht thematisiert und die Möglichkeit eines kooperativen Konfliktausganges (wie sie ja tatsächlich gefunden wurde[1]) wird nicht gesehen.

Der gesamte Kommentar beinhaltet ausschließlich eskalationsorientierte Textmerkmale, deeskalationsorientierte Merkmale kommen nicht vor. Nur eine einzige Textstelle (1/2) ist eine reine Sachinformation. Aber auch diese dient dazu, die im Satz davor (1/1) und im Satz danach (1/3) geschürte Siegeszuversicht von der Sache her zu untermauern.

1. Wim Duisenberg wurde erster Präsident der EZB, trat vereinbarungsgemäß nach der Hälfte seiner Amtszeit zurück und machte das Amt frei für Claude Trichet.

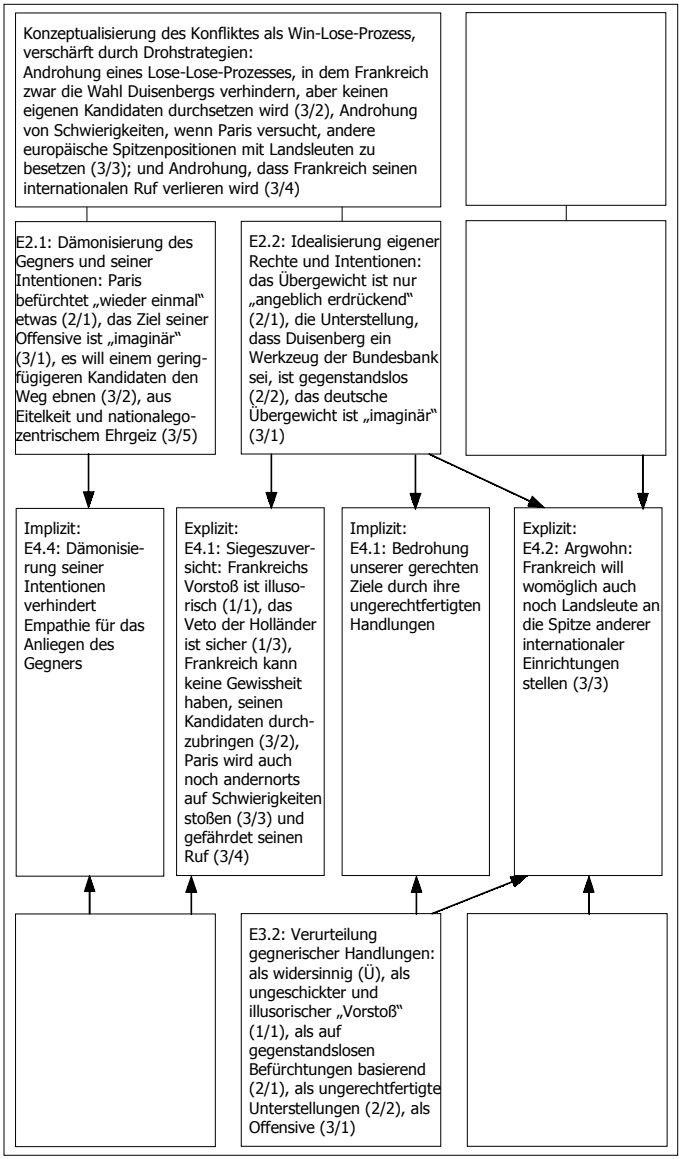


Abb. 1.4.4.2: Das mentale Modell, nach welchem der Konflikt konstruiert wird

Damit entspricht die Konstruktion des Konfliktes dem mentalen Modells eines (Macht-)Kampfes. Die Zeichen stehen auf allen Dimensionen der Konfliktwahrnehmung auf Kampf.

- Frankreich und seine Intentionen werden dämonisiert: Paris befürchtet "wieder einmal" etwas (2/1), das Ziel seiner Offensive ist "imaginär" (3/1), es will einem gefügigeren Kandidaten den Weg ebnen (3/2), und zwar aus Eitelkeit und nationalegozentrischem Ehrgeiz (3/5).
- Frankreichs Handlungen werden als "widersinnig" (Ü) verurteilt, als ein ungeschickter und illusorischer "Vorstoß" (1/1), der auf gegenstandslosen Befürchtungen basiert (2/1); als ungerechtfertigte Unterstellung (2/2) und als Offensive gegen die Bundesbank (3/1).
- Die eigenen Rechte und Intentionen werden idealisiert: das deutsche Übergewicht ist nur "angeblich erdrückend" (2/1), die Unterstellung, dass Duisenberg ein Werkzeug der Bundesbank sei, ist "gegenstandslos" (2/2) und das deutsche Übergewicht ist "imaginär" (3/1).
- Lediglich eine Rechtfertigung der deutschen Handlungen und die Betonung der eigenen Stärke fehlt. Dies liegt aber daran, dass Deutschland in dem Artikel als Subjekt-Akteur gar nicht vorkommt. Das Heft des Handelns wird den Holländern und anderen Europäern überlassen und die Rechtfertigung ihres Handeln wird ebenso stillschweigend unterstellt (1/3: "Das Veto der Holländer gilt als sicher ..."; 3/3: "Paris dürfte auf Schwierigkeiten stoßen ...") wie ihre Intentionen idealisiert werden (3/5: "Denn nationalegoistischer Ehrgeiz und Eitelkeiten sind weniger denn je beliebt").
- Die Dämonisierung Frankreichs und seiner Intentionen verhindert Empathie für das Anliegen des Gegners (implizit), dessen Handlungsweise die selbstlosen eigenen Ziele (= die Ernennung Duisenbergs) bedroht (implizit) und dem (explizit) mit Argwohn (1/3: Frankreich wird auch noch andere Spitzenpositionen mit seinen Landsleuten besetzen wollen) und Siegeszuversicht entgegengetreten wird: Frankreichs Vorstoß ist illusorisch (1/1), das Veto der Holländer ist sicher (1/3), Frankreich kann keine Gewissheit haben, seinen Kandidaten durchzubringen (3/2), und Paris wird auch noch andernorts auf Schwierigkeiten stoßen (3/3) und gefährdet seinen Ruf (3/4).

1.4.5 Quantitative Inhaltsanalyse

Das Ablaufschema der quantitativen Inhaltsanalyse

Das Ablaufschema der quantitativen Inhaltsanalyse ist von jenem der qualitativen nicht grundsätzlich verschieden. Es enthält jedoch einige zusätzliche Schritte, die dem Ziel der statistischen

Auswertung einer repräsentativen Stichprobe von Texten geschuldet sind. Diese betreffen insbesondere: die Festlegung der Grundgesamtheit der zu analysierenden Texte, die Stichprobenziehung, die Definition der Kodiereinheiten, die Kontrolle der Intercoder-Reliabilität und die statistische Datenauswertung (vgl. Abb. 1.4.5.1).

Grundgesamtheit der zu analysierenden Texte

Wollen wir z. B. die Berichterstattung über den deutsch-französischen Konflikt um die Präsidentschaft der EZB analysieren (vgl. Plontz, 2006), so stellt sich zunächst die Frage, was denn überhaupt die Grundgesamtheit der zu analysierenden Texte sein soll:

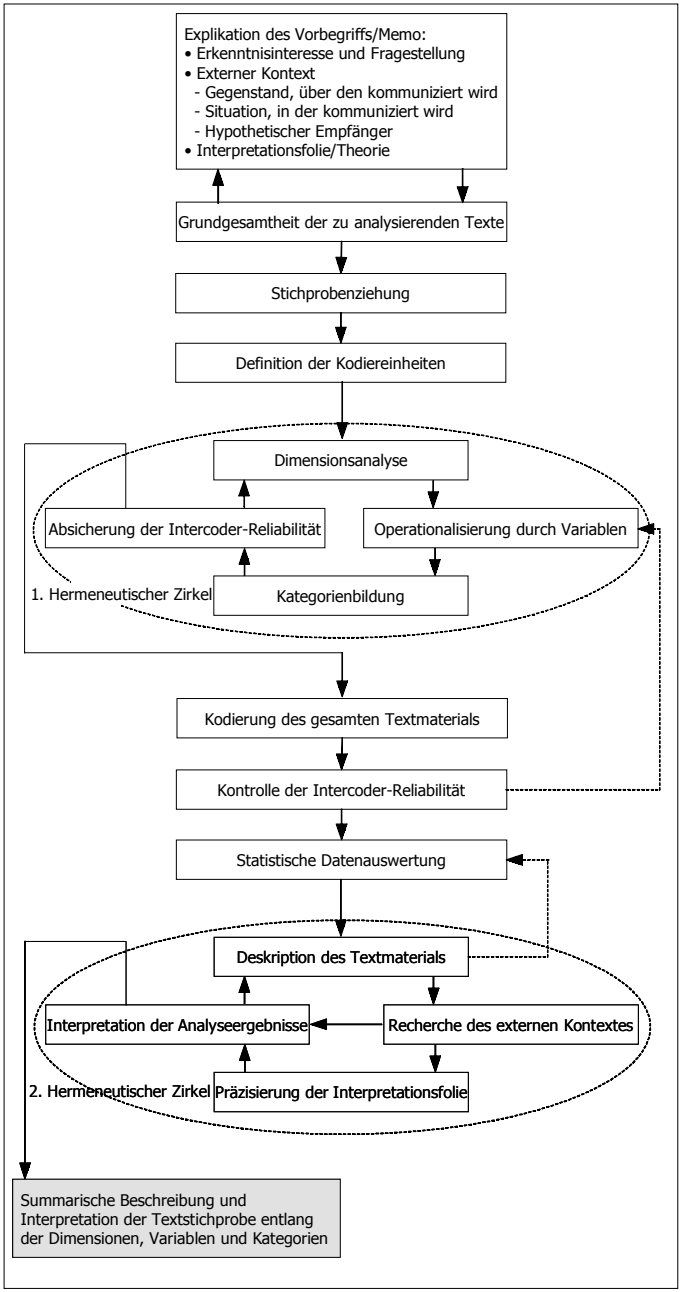
- Welcher Berichtszeitraum?
- Welche thematische Einschränkung der Texte? Z. B. alle Texte, welche die bevorstehende Euroeinführung zum Gegenstand haben, oder nur jene, die von der Präsidentschaft der EZB handeln?
- Nur die deutsche Berichterstattung oder (kontrastierend) auch die französische?
- Welche Medien? Qualitätspresse, Regionalzeitungen, Boulevardzeitungen, Fernsehnachrichten etc,
- Tagespresse bzw. Nachrichtensendungen oder (auch) Wochenzeitungen bzw. Magazinsendungen?
- Medienerzeugnisse, die das gesamte (partei-)politische Spektrum abdecken (im Falle Deutschlands: *taz*, *FR*, *SZ*, *FAZ*, *Welt*) oder solche, die ein besonders breites Publikum erreichen (im Falle Deutschlands: *Bild*, *Tagesschau*, *heute*).
- Welches Textgenre? Nachrichten, Kommentare, Dokumentationen etc.

Stichprobenziehung

Sobald die Grundgesamtheit der zu analysierenden Texte feststeht, stellt sich die Frage der Stichprobenziehung:

- Wie sind die für die Fragestellung der Untersuchung relevanten Texte zu identifizieren? Muss das Thema bereits in der Überschrift erkennbar sein oder erst beim Hineinlesen (z. B. im ersten Absatz)?
- Soll das gesamte relevante Textmaterial analysiert werden oder nur eine Stichprobe?
- Wird eine repräsentative (Zufalls-)Stichprobe angestrebt oder eine systematische Stichprobe, welche die Berichterstattung an bestimmten Stichtagen bzw. zu bestimmten Themen umfasst?
- Sollen alle Texte der systematischen Stichprobe analysiert werden oder eine Zufallsauswahl daraus?

Abb. 1.4.5.1: Ablaufschema der quantitaven Inhaltsanalyse



Wie man sich bezüglich dieser Fragen entscheidet, wird auch von der Berichtdichte im analysierten Zeitraum sowie von der ins Auge gefassten Auswertungsmethode abhängen, die je nach Art des statischen Verfahrens eine bestimmte Mindestgröße der Stichprobe erforderlich macht.

Definition der Kodiereinheiten

Weiters wird die Entscheidung auch von der Art der Kodiereinheiten abhängen, die ebenfalls bereits vorab festzulegen ist, um das Kodierschema darauf abstimmen zu können. Die Frage, was als Kodiereinheit gewählt werden soll, kann nur pragmatisch entschieden werden.

- Ganze Texte (z. B. Nohrstedt, 2001; Mattenschlager & Riedle, 2003)? Das ist naheliegend, hat aber den Nachteil, dass die Texte sehr unterschiedlich lang sein können. Je nach Art des inhaltsanalytischen Kodierschemas können ganze Texte daher mitunter wenig geeignet sein.
- Absätze (z. B. Kempf & Reimann, 1994; Plontz, 2006)? Das hat den Vorteil, dass die Kodiereinheiten in etwa alle gleich lang sind und entspricht zudem der vom Autor intendierten Gliederung.
- Konfliktkonstellationen: Wer mit wem gegen wen? (z. B. Jaeger, Mattenschlager & Meder, 1999). Das ist von Vorteil, wenn man es – wie z. B. bei der Berichterstattung über die post-jugoslawischen Bürgerkriege – mit unübersichtlichen und ständig wechselnden Konfliktlinien zu tun hat und in denselben Texten oft über mehrere Konfliktkonstellationen berichtet wird. Ein Nachteil besteht darin, dass die Kodiereinheiten dann nicht bereits vom Text selbst vorgegeben sind, so dass bereits ihre Identifikation eine Interpretationsleistung beinhaltet.
- Handlungsbeschreibungen: Wer tut wem was? (z. B. Kempf & Palmbach, 1994). Das erlaubt eine gezielte Analyse der Art und Weise, wie das Handeln der Konfliktparteien dargestellt wird, hat aber dieselben Nachteile wie die Definition der Kodiereinheiten durch Konfliktkonstellationen.

Gegebenenfalls kann man die Kodiereinheiten auch durch die einzelnen Sätze definieren. Ob dies sinnvoll ist, wird wegen ihrer relativen Kürze aber davon abhängen, ob das Kodierschema für eine Anwendung auf so kurze Textstellen geeignet ist.

Bei gesprochenem Text ist die Definition der Kodiereinheiten noch ungleich schwieriger. Da die gesprochene Rede ohne Punkt und Komma ist, entfällt die Definition der Kodiereinheiten durch

Sätze oder Absätze von vornherein. Allenfalls könnte man versuchen, den Text in Sequenzen einzuteilen, was aber wieder eine Interpretationsleistung erfordert, die zudem weit komplexer ist als die Identifikation von Konfliktkonstellationen oder Handlungsbeschreibungen, für die sich präzise Regeln angeben lassen. Die Einteilung in Sequenzen lässt daher nur eine geringe Intercoder-Reliabilität erwarten. Am einfachsten ist es noch bei interaktiven Texten, wo man die Kodiereinheiten durch die einzelnen Redebeiträge definieren kann, deren Anfang und Endpunkt durch den Sprecherwechsel markiert sind.

Konstruktion des Kodierschemas

Die zur Konstruktion des Kodierschemas erforderlichen Schritte sind im Wesentlichen dieselben wie bei der qualitativen Inhaltsanalyse auch. Das Kodierschema wird jedoch nur anhand einer Teilstichprobe der zu analysierenden Texte entwickelt und auf seine Intercoder-Reliabilität hin untersucht, bevor es zur Kodierung des gesamten Materials herangezogen wird.

Um es mit hinreichender Intercoder-Reliabilität auf große Textmengen anwenden zu können, wird man ein Kodierschema entwickeln müssen, das weniger interpretativ ist und mehr an der Textoberfläche bleibt als z. B. jenes, das wir in Abschnitt 1.4.4. zur Rekonstruktion des mentalen Modells verwendet haben, dem die Konfliktdarstellung unseres Beispieltextes folgt.

In ihrer quantitativen Vergleichsstudie der deutschen und französischen Presseberichterstattung über den Konflikt um die Präsidentschaft der EZB musste Plontz (2006) das Kodierschema daher erst an die Bedürfnisse einer quantitativen Inhaltsanalyse anpassen, wofür acht Variablen gebildet wurden, die in Tab. 1.4.5.1 dargestellt sind. Sieben der Variablen sind polytom mit je vier Kategorien 'deeskalationsorientiert', 'eskalationsorientiert', 'ambivalent' (wenn in der selben Kodiereinheit sowohl deeskalations- als auch eskalationsorientierte Aspekte der Variable vertreten waren) und 'nicht vorhanden'. Die achte Variable ('Verwendung propagandaähnlicher Argumentationstechniken' mit den Kategorien 'vorhanden' vs. 'nicht vorhanden' wurde binär kodiert.[1]

1. Zu den Ankerbeispielen und Kodierregeln siehe Plontz (2005, 47-103).

Nr.	Variable	Kodierung
1	Wettbewerbsorientierte vs. kooperationsorientierte Stellungnahmen	Polytom
2	Ausdruck von Werten	
3	Darstellung des Antagonismus	
4	Darstellung der Rechte, Absichten und Eigenschaften der anderen Seite	
5	Darstellung der Rechte, Absichten und Eigenschaften der eigenen Seite	
6	Darstellung der Verhaltensweisen der anderen Seite	
7	Darstellung der Verhaltensweisen der eigenen Seite	
8	Verwendung propagandaähnlicher Argumentationstechniken	Dichotom

Tab. 1.4.5.1: Quantitatives Kodierschema zur Erfassung eskalations- und deeskalationsorientierter Aspekte der Berichterstattung über den Konflikt um die Präsidentschaft der EZB

Kontrolle der Intercoder-Reliabilität

Zur Kontrolle der Intercoder-Reliabilität wird eine Teilstichprobe von Kodiereinheiten von zwei unabhängigen Auswertern kodiert und ein Maß für die Beurteilerübereinstimmung berechnet. Dafür steht eine Vielzahl an statistischen Kennzahlen zur Verfügung (vgl. Wirtz & Casper, 2002), unter denen sich insbesondere der Koeffizient κ nach Cohen (1960) durch seine überzeugende Logik auszeichnet.

	0	..	j	..	$m-1$	Σ
0	n_{00}		n_{0j}		$n_{0,m-1}$	n_{0o}
:						
i	n_{i0}		n_{ij}		$n_{i,m-1}$	n_{io}
:						
$m-1$	$n_{m-1,0}$		$n_{m-1,j}$		$n_{m-1,m-1}$	$n_{m-1,o}$
Σ	n_{o0}		n_{oj}		$n_{o,m-1}$	n_{oo}

Tab. 1.4.5.2: Matrix der konkordanten und diskordanten Kodierungen

Berechnung des Koeffizienten κ

Zur Berechnung von κ fertigt man eine quadratische Kontingenztafel an, in welche die Kodiereinheiten eingetragen werden, wobei die Zeilen dem vom 1. Auswerter vergebenen Code und die Spalten dem vom 2. Auswerter vergebenen Code entsprechen. Die Eintragungen n_{ij} in der Kontingenztafel sind die Häufigkeiten, mit denen die Kodiereinheiten vom 1. Auswerter in Kategorie i und vom 2. Auswerter in Kategorie j kodiert wurden (vgl. Tab. 1.4.5.2).

Im Falle einer idealen Urteilseinstimmung müssten alle Eintragungen in der Hauptdiagonale der Kontingenztafel stehen. Entsprechend berechnet man gemäß

$$P_0 = \frac{1}{n} \sum_{i=0}^{m-1} n_{ii} \qquad [1.4.5.1]$$

den Anteil der konkordanten Kodierungen und vergleicht diesen mit dem Anteil der Urteilsübereinstimmungen, der schon allein aufgrund des Zufalls zu erwarten wäre. Dieser errechnet sich gemäß

$$P_e = \frac{1}{n} \sum_{i=0}^{m-1} e_{ii} \qquad [1.4.5.2]$$

aus den erwarteten Häufigkeiten

$$e_{ij} = \frac{n_{io}\, n_{oj}}{n} \quad . \qquad [1.4.5.3]$$

Zum Vergleich der realisierten mit den rein zufällig zu erwartenden Urteilsübereinstimmungen wird schließlich der über den Zufall hinausgehende Anteil an Übereinstimmungen ($P_0 - P_e$) zum maximal möglichen Anteil von überzufälligen Übereinstimmungen ($1 - P_e$) in Beziehung gesetzt, indem der Quotient

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \qquad [1.4.5.4]$$

berechnet wird. Ein Wert von $\kappa > 0.75$ wird nach Wirtz & Casper (2002, 59) als sehr gute Übereinstimmung gewertet, Werte zwischen 0.6 und 0.75 gelten als gute Übereinstimmung und Werte zwischen 0.4 und 0.6 stehen für eine akzeptable Übereinstimmung (siehe Anhang, S. 370).

Kodierung des Textmaterials und erneute Kontrolle der Intercoder-Reliabilität

Wurde eine hinreichende Intercoder-Reliabilität erzielt, so folgt als nächster Schritt die Kodierung des gesamten Textmaterials, an deren Ende man die Intercoder-Reliabilität erneut an einer Teilstichprobe von Texten überprüfen sollte, da die Möglichkeit besteht, dass die Auswerter im Laufe der Kodierarbeit ihren eigenen, von den Kodieranweisungen abweichenden Stil entwickelt haben. Noch besser ist es, schon während des Auswertungsprozesses immer wieder eine Überprüfung der Intercoder-Reliabilität

vorzunehmen, diskordante Urteile zu diskutieren und eventuelle Stilabweichungen zu korrigieren.

Statistische Datenauswertung

Für die statistische Auswertung des so gewonnenen Datenmaterials steht so gut wie das gesamte statistische Methodeninventar für qualitative Variablen zur Verfügung und welche Auswertungsmethoden man im konkreten Fall anwenden wird, hängt selbstverständlich auch von der jeweiligen Fragestellung ab.

Gleichwohl wird man bei der Wahl des Auswertungsverfahrens aber bemüht sein, so weit wie möglich der Kritik entgegenzutreten, welche bereits Kracauer (1952) an der quantitativen Inhaltsanalyse geübt hat und die auch heute noch auf so manche inhaltsanalytische Untersuchung zutrifft:

1. Die zählenden Verfahren der quantitativen Inhaltsanalyse vernachlässigten die Interdependenzen der verschiedenen Teile eines Textes und die Beziehungen zwischen den Variablen;
2. relevant seien nicht die Häufigkeiten, mit denen die verschiedenen Textmerkmale auftreten, sondern die Muster, welche sie bilden; und
3. folglich bestehe ein Konflikt zwischen interner und externer Validität inhaltsanalytischer Studien, der dazu führe, dass Pilotstudien oft stärker qualitativ ausgerichtet seien, die externe Validität der quantitativen Hauptstudie dann aber hinter jener der Pilotstudie zurückbleibe.

Während dem ersten dieser Kritikpunkte bereits mittels der klassischen Verfahren der Kontingenz- und Korrelationsanalyse begegnet werden kann, verlangt der zweite ein statistisches Auswertungsverfahren, welches die Merkmalsmuster, die der Kodierprozess in einzelne Variablen zerlegt hat, wieder rekonstruiert. Oder anders ausgedrückt: Wenn es nicht die einzelnen Textmerkmale sind, die als Indikatoren für die Argumentationsrichtung eines Textes dienen können, sondern die Merkmalsmuster, dann bedürfen wir eines statistischen Verfahrens, welches geeignet ist, die in den analysierten Texten enthaltenen, in den daraus gewonnenen Daten aber nicht mehr manifesten Muster von Textmerkmalen zu identifizieren.

Latent-Class-Analyse

Eben dies leistet die etwa zeitgleich mit Kracauers Kritik von Lazarsfeld (1950) entwickelte Latent-Class-Analyse (vgl. Kap. 2.3 und 2.10), die allerdings erst durch Tarnai & Bos (1989) erstmals Anwendung in der Inhaltsanalyse gefunden hat.

Sinn und Zweck der inhaltsanalytischen Anwendung der Latent-Class-Analyse ist es,

1. die in den Daten (latent) enthaltenen Merkmalsmuster (= latente Klassen) zu identifizieren,
2. zu untersuchen, in welcher Art von Texten die verschiedenen Merkmalsmuster besonders häufig (oder selten) vorkommen,
3. typische Texte zu identifizieren, welche diese Merkmalsmuster gleichsam in Reinform enthalten und
4. diese dann mittels qualitativer Verfahren (Textinterpretation oder Inhaltsanalyse) mehr in die Tiefe gehend zu analysieren, womit die empirische Grundlage geschaffen wird, auf welcher die quantitativ-inhaltsanalytischen Ergebnisse validiert und interpretiert werden können (vgl. Kempf, 2002).

Damit wird nicht nur das Verhältnis zwischen quantitativen und qualitativen textanalytischen Verfahren neu bestimmt, sondern letztlich auch die damit korrelierende Trennung zwischen Vor- und Hauptuntersuchung aufgehoben: Im Resultat jeder der beiden Analysemethoden entstehen Erkenntnisse, auf welche die andere aufbauen kann, und jede der beiden Methoden trägt zum Verständnis der analysierten Texte eine andere Art von Erkenntnissen bei. Während die qualitativen Methoden eine detaillierte Analyse einiger weniger Texte leisten, sorgt die quantitative Inhaltsanalyse dafür, dass die daraus gewonnenen Einsichten nicht nur auf ein mehr oder minder willkürlich ausgewähltes Textbeispiel bezogen, sondern in methodisch geregelter Form auf die Grundgesamtheit der interessierenden Texte verallgemeinert werden können.

Kontrollfragen

- Wie lautet die Leitfrage der Inhaltsanalyse nach Lasswell und wie lassen sich ihre Elemente zu den Grundbegriffen der klassischen Kommunikationstheorie und zu den verschiedenen Forschungsproblemen in Beziehung setzen, zu deren Untersuchung auf inhaltsanalytische Methoden zurückgegriffen wird?
- Diskutieren sie, inwieweit Lasswells Leitfrage über die Analyse des Textinhalts hinausgeht!
- Wie definiert Berelson die (quantitative) Inhaltsanalyse und welche Kritik hat Kracauer daran geübt?
- Diskutieren Sie, wie Kracauer das Verhältnis von quantitativer und qualitativer Inhaltsanalyse sieht und stellen Sie die Vor- und Nachteile der beiden Analysemethoden einander gegenüber!

- Was versteht man unter der inhaltsanalytischen Kodierung und welche Varianten lassen sich dabei innerhalb der quantitativen Inhaltsanalyse unterscheiden?
- Nennen Sie Gemeinsamkeiten und Unterschiede von quantitativen und qualitativen inhaltsanalytischen Variablen!
- Erläutern Sie, wie die Bedeutung der inhaltsanalytischen Variablen normiert werden kann!
- Wie unterscheiden sich quantitative und qualitative Inhaltsanalyse hinsichtlich der Offenheit des Kodierverfahrens und wie unterscheiden sie sich diesbezüglich von der Textinterpretation?
- Wie können Sachkundigkeit und Unvoreingenommenheit der Textanalyse sichergestellt werden?
- Erläutern Sie das Ablaufschema der qualitativen Inhaltsanalyse und diskutieren Sie die beiden hermeneutischen Zirkel, die darin enthalten sind! Wie unterscheidet sich die Inhaltsanalyse diesbezüglich von der Textinterpretation?
- Welche Anforderungen sind an inhaltsanalytische Kodierschemata zu stellen?
- Wie unterscheiden sich quantitative und qualitative Inhaltsanalyse bezüglich der Festlegung der Kodiereinheiten?
- Welche Form hat die Darstellung der Analyseergebnisse in der qualitativen Inhaltsanalyse?
- Welche Arten der inhaltsanalytischen Kategorienbildung lassen sich unterscheiden?
- Erläutern Sie anhand eines Beispiels die Vorgehensweise bei der materialbezogenen Kategorienbildung!
- Wie unterscheiden sich quantitative und qualitative Inhaltsanalyse hinsichtlich der Anwendung der Dimensionen, Variablen und Kategorien und welche Konsequenzen ergeben sich daraus für die quantitative Inhaltsanalyse?
- Was versteht man in der quantitativen Inhaltsanalyse unter dem Codemuster einer Kodiereinheit und welche Form hat es?
- Was beschreibt die Codematrix und welche Form hat sie?
- Welcher Zusammenhang besteht zwischen der Art der Kodiereinheit und der Eignung des Kodierschemas in der quantitativen Inhaltsanalyse und welche Möglichkeiten gibt es, das Kodierschema an das jeweilige Textmaterial anzupassen?
- Diskutieren Sie, inwiefern zwischen materialbezogener und theoriebezogener Kategorienbildung ein fließender Übergang besteht!
- Erläutern Sie anhand eines Beispiels die Vorgehensweise bei der theorienbezogenen Kategorienbildung!
- Erläutern Sie anhand eines Beispiels die Vorgehensweise bei der Interpretation der Ergebnisse der qualitativen Inhaltsanalyse!
- Erläutern Sie das Ablaufschema der quantitativen Inhaltsanalyse und diskutieren Sie, inwiefern es sich von jenem der qualitativen Inhaltsanalyse unterscheidet!
- Diskutieren Sie verschiedene Möglichkeiten der Stichprobenziehung in der quantitativen Inhaltsanalyse!
- Nennen Sie einige Möglichkeiten zur Definition der Kodiereinheiten und diskutieren Sie deren Vor- und Nachteile!

- Welche Schwierigkeiten ergeben sich bei der Analyse von gesprochenem Text bezüglich der Definition der Kodiereinheiten?
- Welche Besonderheiten sind bei der quantitativen Inhaltsanalyse bezüglich der Konstruktion des Kodierschemas zu beachten?
- Erläutern Sie die Vorgehensweise bei der Kontrolle der Intercoder-Reliabilität!
- Welche Kritik hat Kracauer an der quantitativen Inhaltsanalyse geübt, und durch welche Maßnahmen kann dieser Kritik entgegengekommen werden?
- Diskutieren Sie Sinn und Zweck der Anwendung der Latent-Class-Analyse auf inhaltsanalytische Daten!

1.5 Integration quantitativer und qualitativer Analysemethoden

Fragestellung, Grundgesamtheit und Stichprobenauswahl

Zur Demonstration der Vorgehensweise greifen wir auf eine Studie von Kempf & Reimann (2002) zurück, welche die Darstellung verschiedener Konfliktlösungsoptionen in der Berichterstattung über den Golfkrieg zum Gegenstand hatte.

Analysiert wurden europäische und amerikanische Medien sowie zwei äthiopische Prestigezeitungen und der irakische *Baghdad Observer*, der jedoch nach Beginn der Luftangriffe sein Erscheinen einstellte und daher nur während der Vorkriegsphase analysiert werden konnte.

Untersucht wurde die gesamte Golfkriegsberichterstattung dieser Medien an 31 Stichtagen zwischen dem 2. August 1990 und dem 15. Januar 1993. Die Stichprobe der europäischen Medien beinhaltete in jedem der vier Länder Deutschland, Finnland, Norwegen und Schweden jeweils die größte überregionale Qualitätszeitung, die größte Boulevardzeitung, eine der größten Regionalzeitungen und die Abendnachrichten des wichtigsten nichtkommerziellen Fernsehsenders. Die amerikanische Medienstichprobe war ähnlich aufgebaut, enthielt jedoch keine Regionalzeitung und die Nachrichtensendungen der Fernsehstationen *ABC*, *CBS* und *NBC* konnten nur während der heißen Kriegsphase und nur so weit analysiert werden, als sie durch *Sky News* in Europa verbreitet wurden. Insgesamt wurden n = 4096 Nachrichtentexte in die Studie aufgenommen.

In 18.1% der analysierten Texte wurde mindestens eine der nachstehenden Optionen zur Lösung des Golfkonfliktes angesprochen: militärische Gewaltanwendung, wirtschaftliche Sanktionen, irakische Friedensinitiativen, Friedensinitiativen von Mitgliedern der Golfkriegsallianz, Friedensinitiativen der Vereinten Nationen und/oder Friedensinitiativen dritter Parteien. Diese Teilstichprobe von n = 740 Texten lag der unten dargestellten quantitativen Inhaltsanalyse zugrunde.

Das quantitative Kodierschema

Ziel der quantitativen Inhaltsanalyse war es, zu untersuchen, ob und wie die Konfliktlösungsoptionen in den Medien dargestellt wurden: affirmativ (= neutral bis zustimmend) und/oder negativ (= infragestellend bis ablehnend). Durch Kombination der Konfliktlösungsoptionen mit der Art ihrer Darstellung wurde hierfür das in Tab. 1.5.1 dargestellte Kodierschema aus zwölf binären Variablen mit den Kategorien 1 = "kommt vor" und 0 = "kommt nicht vor" gebildet.

Für die später durchzuführende Latent-Class-Analyse war es wegen der Annahme der lokalen Unabhängigkeit der Variablen (vgl. Kap. 2.2.4) maßgeblich, dass die Variablen logisch unabhängig sein mussten. Dies wurde dadurch erzielt, dass derselbe Text bezüglich ein und derselben Konfliktlösungsoption nicht nur alternativ auf der affirmativen oder auf der negativen Variable, sondern ggf. auch auf beiden Variablen in der Kategorie 'kommt vor' oder auf beiden Variablen in der Kategorie 'kommt nicht vor' kodiert werden konnte (Kodierregel).

Tab. 1.5.1: Die Variablen der quantitativen Inhaltsanalyse

Variable	Thema	Bezugnahme
M+	Militärische Gewaltanwendung als einzig mögliche und/oder wirksamste Form des Konfliktaustrags	Affirmativ
M-		Negativ
W+	Wirtschaftlichen Sanktionen sollten mehr Zeit gegeben werden oder hätten mehr Zeit gegeben werden sollen oder um wirksam zu werden	Affirmativ
W-		Negativ
I+	Irakische Friedensinitiativen	Affirmativ
I-		Negativ
A+	Friedensinitiativen von Mitgliedern der Golfkriegs-Allianz (ohne UN)	Affirmativ
A-		Negativ
UN+	Friedensinitiativen der Vereinten Nationen	Affirmativ
UN-		Negativ
D+	Friedensinitiativen dritter Parteien	Affirmativ
D-		Negativ

Häufigkeitsverteilung der Variablen

Die Häufigkeiten, mit denen die Variablen in der Kategorie "kommt vor" kodiert wurden, sind in Abb. 1.5.1 dargestellt: Die militärische Gewaltanwendung stellte die am häufigsten und auch am kontroversesten diskutierte Konfliktlösungsoption dar. Häufig (in 28.4% der analysierten Texte) wurde sie als einzig

mögliche und/oder wirksamste Form des Konfliktaustrags porträtiert, aber auch nicht selten (20.0%) wurde dies angezweifelt oder verneint. Friedensinitiativen dritter Parteien folgten an zweiter Stelle und wurden insgesamt weit positiver dargestellt (39,6% affirmativ, nur 6.8% negativ). Die geringste Erwähnung fanden Friedensinitiativen von Mitgliedern der Golfkriegsallianz, die nur selten affirmativ (9.2%) und so gut wie nie (0.8%) negativ thematisiert wurden.

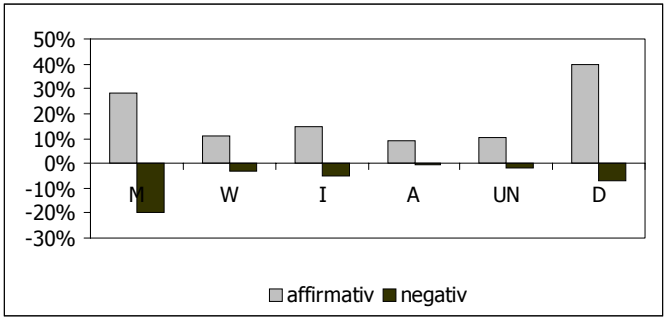


Abb. 1.5.1: Die Darstellung der Konfliktlösungsoptionen in den internationalen Medien M = Militärische Gewaltanwendung; W = Wirtschaftliche Sanktionen; I = Irakische Friedensinitiativen; A = Alliierte Friedensinitiativen; UN = Friedensinitiativen der UNO; D = Friedensinitiativen dritter Parteien

Sprachnormierung

Zur verbalen Beschreibung der Häufigkeitsverteilung – wie auch später bei der Beschreibung der LCA-Ergebnisse – empfiehlt es sich, von einer Art Sprachnormierung Gebrauch zu machen, die verhindern soll, dass dieselbe Prozentzahl z. B. einmal als häufig und ein anderes Mal nur als gelegentlich umschrieben wird. Hierfür hat sich die im Anhang (S. 370) wiedergegebene Sprachregelung bewährt, die wir auch oben angewendet haben.

Da kleine und große Wahrscheinlichkeiten mit der geringsten Fehlerstreuung geschätzt werden können, ist der Häufigkeitsbereich, der durch dieselben Wortmarken beschrieben wird, für mittelgroße Häufigkeiten am breitesten und für sehr kleine oder sehr große Häufigkeiten am schmalsten gewählt.

Interpretation der Häufigkeitsverteilung

Insgesamt bietet die Häufigkeitsverteilung der Variablen in Abb. 1.5.1 das Bild, dass alle Konfliktlösungsoptionen in den Medien berichtet wurden. Selbst irakische Friedensangebote wurden nicht selten (14.5%) affirmativ berichtet und kaum in Frage gestellt (4.9%). Zudem lässt das Verhältnis (affirmativ : negativ) von ca. 3:1 im Vergleich zu dem Verhältnis von etwas weniger als 3:2, mit dem ein militärischer Konfliktaustrag favorisiert wurde, und dem Verhältnis von mehr als 5:1, mit dem Friedensinitiativen

dritter Parteien gewürdigt wurden, den Eindruck entstehen, als hätten die untersuchten Medien dem Golfkrieg eher distanziert und kritisch gegenübergestanden. Dieser Eindruck ist jedoch trügerisch, und die Häufigkeitsverteilung in Abb. 1.5.1 sagt nichts darüber aus, in welche Richtung die Medien tatsächlich argumentiert haben.

Latent-Class-Analyse

Um die tatsächliche(n) Argumentationsrichtunge(n) rekonstruieren zu können, ist es nicht ausreichend, nur die Häufigkeiten zu betrachten, mit denen die verschiedenen Pro- und Contra-Argumente insgesamt vorgebracht wurden, sondern wir müssen die Argumentationsmuster identifizieren, zu denen sie kombiniert sind.

Die Anwendung der Latent-Class-Analyse auf die o.g. Daten resultierte gemäß AIC-Kriterium (vgl. Kap. 2.10.3) in der Identifikation von neun typischen Argumentationsmustern (vgl. Tab. 1.5.2), die in Abb. 1.5.2 dargestellt sind.

Tab. 1.5.2: Goodness-of-Fit-Statistiken der Latent-Class-Analyse (nach Kempf, 2002, 157)

h	ln{mL(X)}	n(P)	df	AIC
1	-2869.618	12	4083	5763.23
2	-2708.930	25	4070	5467.86
3	-2657.564	38	4057	5391.12
4	-2557.404	51	4044	5216.80
5	-2456.286	64	4031	5040.57
6	-2427.489	77	4018	5008.97
7	-2368.420	90	4005	4916.83
8	-2354.040	103	3992	4914.07
9	-2326.360	116	3979	4884.71
10	-2319.250	129	3966	4896.50
Sat.	-2223.318	4095		

Drei dieser Argumentationsmuster (41.1% des analysierten Textmaterials) fokussieren die Anwendung militärischer Gewalt und oder wirtschaftlicher Sanktionen gegen den Irak:

- Militärische Logik (16.7%), die militärische Gewaltanwendung fast ausnahmslos (99.9%) als einzig gangbare und/oder effektivste Form des Konfliktaustrags darstellt und dies kaum (9.9%) je in Frage stellt.
- Infragestellung militärischer Logik (14.3%), welche Argumente zugunsten des militärischen Konfliktaustrag zwar relativ

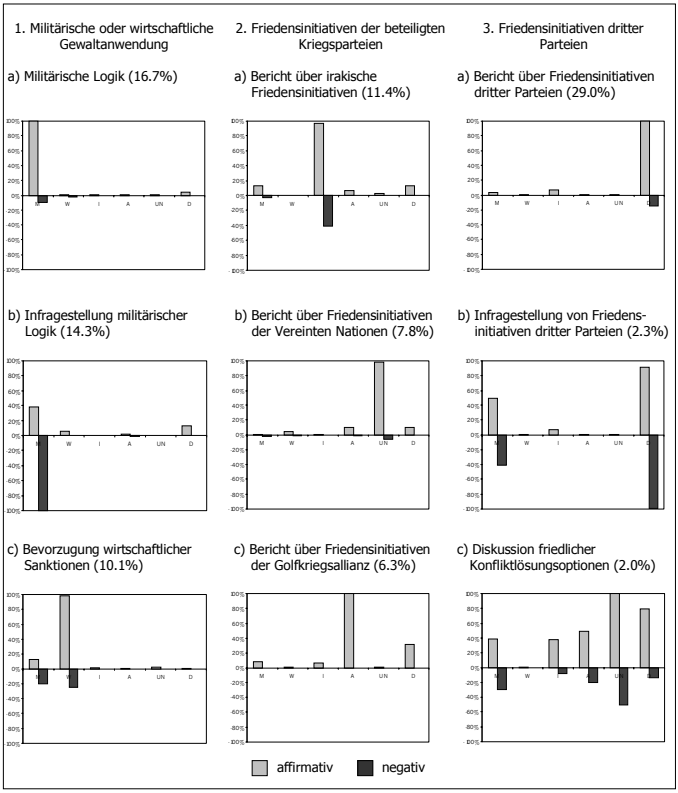


Abb. 1.5.2: Typische Argumentationsmuster
M = Militärische Gewaltanwendung
W = Wirtschaftliche Sanktionen
I = Irakische Friedensinitiativen
A = Alliierte Friedensinitiativen
UN = Friedensinitiativen der UNO
D = Friedensinitiativen dritter Parteien

häufig (38.0%) vorträgt, die Angemessenheit des militärischen Konfliktaustrags jedoch so gut wie ausnahmslos (99.9%) in Frage stellt und nicht selten auch auf Friedensinitiativen dritter Parteien affirmativ Bezug nimmt (13.4%).

- Bevorzugung wirtschaftlicher Sanktionen (10.1%), die in der Regel dafür argumentiert, dass diesen mehr Zeit gegeben werden sollte (98.2%), dabei aber durchaus ambivalent ist und nicht selten auch Gegenargumente vorbringt (25.3%). Immerhin wird die militärische Gewaltanwendung in diesem Argumentationsmuster zwar nur gelegentlich, aber doch häufiger negativ thematisiert (20.2%) als affirmativ (13.0%).

Drei weitere Argumentationsmuster (25.5% des Textmaterials) fokussieren Friedensinitiativen der beteiligten Kriegsparteien:

- 11.4% der analysierten Texte berichten über irakische Frie-

densinitiativen (97%), die sie jedoch häufig in Frage stellen (41.0%) und dabei gelegentlich auf die Unverzichtbarkeit militärischer Gewalt (13.4%) und/oder auf Friedensbemühungen dritter Parteien (13.3%) verweisen.
- 7.8% der Texte berichten über Friedensinitiativen der Vereinten Nationen (98.2%), die – im Unterschied zu den irakischen Initiativen – kaum in Frage gestellt werden (6.2%).
- Nur wenige Texte (6.3%) berichten so gut wie ausnahmslos affirmativ über Friedensinitiativen von Mitgliedern der Golfkriegsallianz (99.9%), stellen diese so gut wie nie in Frage (0.1%) und beziehen sich häufig auch affirmativ auf Friedensinitiativen dritter Parteien (31.0%).

Die verbleibenden drei Argumentationsmuster (33.3% des Textmaterials) fokussieren Friedensinitiativen dritter Parteien:

- 29.0% der analysierten Texte berichten so gut wie ausnahmslos affirmativ über Friedensinitiativen dritter Parteien (99.9%), stellen diese aber gelegentlich auch in Frage (14.1%).
- Einige der Texte (2.3%) stellen Friedensinitiativen dritter Parteien, auf die sie sich in der Regel auch affirmativ beziehen (92.0%), so gut wie ausnahmslos in Frage (99.0%) und beziehen sich auf die militärische Gewaltanwendung eher affirmativ (49.7%) als negativ (40.6%).
- Ein noch kleinerer Anteil der Texte (2.0%) beinhaltet eine umfassende Diskussion friedlicher Konfliktlösungsoptionen, wobei fast ausnahmslos auf Friedensinitiativen der Vereinten Nationen (99.9%), in den meisten Fällen auf Friedensinitiativen dritter Parteien (79.8%) und häufig auch auf alliierte (49.0%) und irakische (38.2%) Friedensinitiativen affirmativ Bezug genommen und militärische Gewaltanwendung relativ häufig abgelehnt (29.3%) wird. Noch häufiger aber wird militärische Gewalt affirmativ thematisiert (39.2%) und keine der Friedensinitiativen bleibt unhinterfragt,[1] so dass sich insgesamt ein eher ambivalentes Bild bietet.

Kontextualisierung der LCA-Ergebnisse

Mit der Identifikation und Beschreibung typischer Argumentationsmuster sind wir gegenüber der reinen Häufigkeitsbeschreibung unserer Datenstichprobe (vgl. Abb. 1.5.1) schon ein ganzes Stück weitergekommen. Für ein besseres Verständnis dieser Argumentationsmuster können wir in einem nächsten Schritt unter-

1. Irakische Initiativen werden zu 8.4%, alliierte zu 20.3%, und Initiativen der Vereinten Nationen sogar zu 50.0% (auch) negativ thematisiert.

suchen, welche der Argumentationsmuster (1) in welcher Art von Medien, (2) während welcher Phasen des Golfkriegs oder (3) in welchen Ländern besonders häufig oder besonders selten zu finden sind, etc.[1]

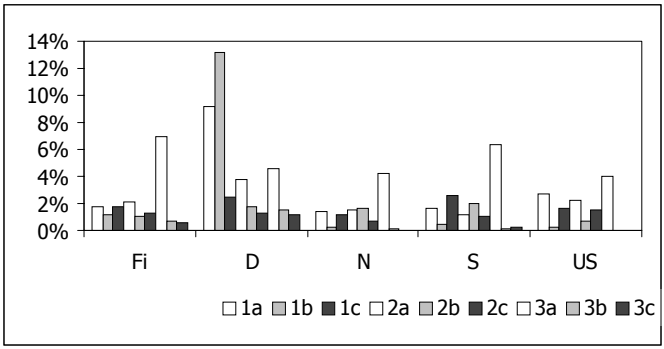


Abb. 1.5.3: Häufigkeit der Argumentationsmuster in den verschiedenen Ländern
Fi = Finnland
D = Deutschland
N = Norwegen
S = Schweden
US = USA

Als Beispiel hierfür betrachten wir die Auftrittshäufigkeit der Argumentationsmuster in den deutschen und amerikanischen Medien insgesamt, d. h. bezogen auf die Gesamtstichprobe aller n = 4096 Texte. Wie wir Abb. 1.5.3 entnehmen können, war die Präsentation alternativer Konfliktlösungsoptionen in den amerikanischen Medien deutlich von den strategischen Interessen der USA geprägt.

- Friedensinitiativen der Vereinten Nationen (2b) oder dritter Parteien (3a) wurde weit weniger Aufmerksamkeit geschenkt als in den europäischen Medien,
- die militärische Logik wurde extrem selten in Frage gestellt (1b) und
- eine umfassende Diskussion friedlicher Konfliktlösungsoptionen (3c) wurde in den amerikanischen Medien überhaupt nicht gefunden.

Die deutschen Medien präsentierten im Gegensatz hierzu die ausführlichste und kontroverseste Diskussion der verschiedenen Konfliktlösungsoptionen:

- Ein der militärischen Logik folgendes Argumentationsmuster (1a) trat dreimal so oft auf wie im Durchschnitt;

1. Zur Berechnung der Kontingenzen auf Grundlage der Membership-Wahrscheinlichkeiten vgl. Kap. 2.10.5.

- das Argumentationsmuster der Zurückweisung militärischer Logik (1b) war sogar fünfmal so oft zu finden und
- eine umfassende Diskussion friedlicher Konfliktlösungsoptionen (3c) gab es ebenfalls dreimal so häufig wie im Durchschnitt.

Bedeutsam wird dieses Ergebnis vor dem historischen Hintergrund des Golfkriegsdiskurses (externer Kontext): Während die USA die Führungsnation der Golfkriegsallianz war, war Deutschland am Golfkrieg nicht militärisch beteiligt. Im Einklang mit der damals (noch) gültigen Interpretation des Grundgesetzes war der Verfassungsauftrag der Bundeswehr strikt auf die Verteidigung von deutschem (oder NATO-)Territorium beschränkt. Der Golfkrieg wurde von deutschen Politikern und Medien jedoch benutzt, um eine Verfassungsdebatte über die künftige Rolle der Bundeswehr nach dem Fall der Berliner Mauer und der deutschen Wiedervereinigung in Gang zu setzen. Dafür war zweierlei vonnöten: Die Beachtung der Argumente der Friedensbewegung, die in Deutschland besonders stark war, und die Entkräftung ihrer Argumente durch die Betonung der Unvermeidbarkeit militärischer Gewalt, um Diktatoren wie Saddam Hussein Einhalt bieten zu können, wo immer in der Welt sie deutsche Interessen beeinträchtigen sollten (vgl. Kempf, 1994; Meder, 1998; Schluroff, 1998).

Entsprechend waren die Informationen Pro- und Kontra der verschiedenen Konfliktlösungsoptionen in den deutschen Medien sowohl vielfältiger als auch häufiger zu finden als in den amerikanischen Medien. Während die amerikanischen Medien das Thema nur in 13.1% aller Texte berührten, taten dies die deutschen Medien fast dreimal so oft (38.9%).

Grenzen der quantitativen Inhaltsanalyse

Da die Variablen der quantitativen Inhaltsanalyse relativ an der Textoberfläche bleiben und lediglich beschreiben, ob die Konfliktlösungsoptionen affirmativ und/oder negativ erwähnt werden, geben auch die identifizierten Argumentationsmuster lediglich Aufschluss über die Informationslage (report aspect), nicht jedoch über die Überredungstechniken (command aspect), mittels derer manche Argumente gestützt, andere entwertet und bestimmte Schlussfolgerungen nahegelegt werden (parade aspect).

Um Aufschluss darüber zu gewinnen, wie die Medien tatsächlich berichtet und die öffentliche Meinung beeinflusst haben, ist eine tiefergehende Analyse der Texte erforderlich. Im Rahmen einer quantitativen Inhaltsanalyse wäre dies allenfalls ein Stück weit

möglich, z. B. indem wir auszählen, wie viele Textstellen pro Text eine bestimmte Konfliktlösungsoption affirmativ thematisieren und wie viele dies negativ tun. Oder wir könnten auf einer Lickert-Skala einschätzen, wie sehr der Text insgesamt eher zum affirmativen oder zum negativen Pol tendiert. Beides erfordert jedoch bereits eine Interpretation der zu analysierenden Texte[1] und dürfte kaum mit der erforderlichen Intercoder-Reliabilität möglich und arbeitstechnisch zu aufwändig sein, um eine Stichprobe von immerhin 740 Texten mit vertretbarem Arbeitsaufwand kodieren zu können. Hinzu kommt, dass auch diese Auszählungen oder Einschätzungen nur relativ grobe Indikatoren für den command und parade aspect der Texte abgeben würden.

Kombination von quantitativer und qualitativer Inhaltsanalyse

Ein Ausweg aus diesem Dilemma bietet sich, indem wir quantitative und qualitative Analyseverfahren miteinander verbinden und eine Auswahl von typischen Textbeispielen einer qualitativen Inhaltsanalyse unterziehen. Welche Texte für ein bestimmtes Argumentationsmuster typisch sind, kann aufgrund ihrer Membership-Wahrscheinlichkeiten (vgl. Kap. 2.10.1) entschieden werden, wobei man jene als typisch auswählt, deren Membership-Wahrscheinlichkeit bezüglich des jeweiligen Argumentationsmusters sich in der Größenordnung von Eins bewegt. Aus der Gesamtmenge der typischen Textbeispiele, die immer noch sehr groß sein kann, kann man sodann eine systematische Auswahl treffen, indem man z. B. darauf achtet, dass alle Länder und alle Arten von Medien, in denen sich typische Beispiele finden lassen, sowie alle Ereignisse, in deren Kontext diese auftreten, durch die zu analysierende Textstichprobe abgedeckt werden.

Qualitative Analyse der Berichterstattung über Friedensinitiativen dritter Parteien

Wie fruchtbar eine solche Kombination von quantitativer und qualitativer Inhaltsanalyse sein kann, demonstrieren wir im Folgenden anhand des Argumentationsmusters 3a, das über Friedensinitiativen dritter Parteien zu 99.9% affirmativ berichtet, sie nur gelegentlich (14.1%) auch in Frage stellt und damit eine gewisse Affinität zu den Friedensinitiativen vermuten lässt.

Ob diese Beachtung der Friedensbemühungen dritter Parteien aber schon so etwas wie eine Distanz der Medien zum Golfkrieg bedeutet oder ob ihre Thematisierung nicht ganz im Gegenteil dazu dient, den Krieg zu rechtfertigen und/oder die Motivation zur Fortsetzung des Krieges zu stützen etc., kann nur mittels ei-

1. Schon die Abgrenzung verschiedener Textstellen voneinander ist nicht interpretationsfrei möglich.

ner tiefergehenden Analyse der entsprechenden Textbeispiele eruiert werden.

Stichprobenauswahl

Als Grundlage für die qualitative Inhaltsanalyse wählen wir je ein typisches Textbeispiel aus der überregionalen Qualitätspresse der USA und der vier europäischen Länder, deren Golfkriegsberichterstattung Gegenstand der Studie waren. Zwei der analysierten Texte waren ganz zu Beginn der Luftangriffe auf Bagdad in der *Washington Post* (USA) und in der *Süddeutschen Zeitung* (Deutschland) erschienen. Das dritte Beispiel erschien im *Helsingin Sanomat* (Finnland) am selben Tag, an dem die Medien mit der Nachricht von der Bombardierung des Amirya Bunkers erstmals über irakische Zivilopfer berichtet hatten. Zwei weitere Texte erschienen in *Aftenposten* (Norwegen) und *Dagens Nyheter* (Schweden) während Gorbatschows letzter Friedensbemühungen vor Beginn der Bodenoffensive.

Das qualitative Kodierschema

Die Analyse der Beispiele erfolgte auf Grundlage des von Kempf, Reimann und Luostarinen (1996) entwickelten Kodierschemas zur qualitativen Inhaltsanalyse eskalations- und deeskalationsorientierter Aspekte der Kriegsberichterstattung, das wir in gekürzter und modifizierter Form bereits in Abschnitt 1.4.4 zur Analyse des Textbeispiels *Widersinnig* verwendet haben.

Das Ergebnis der qualitativen Inhaltsanalyse

Das Ergebnis der qualitativen Inhaltsanalyse zeigte eine deutliche Eskalationsorientierung der europäischen und amerikanischen Medien. Obwohl sie der Darstellung von Friedensinitiativen dritter Parteien breiten Raum einräumten, gab es nur wenig kritischen Journalismus, der den Friedensbemühungen eine Chance ließ.

Weitgehend neutrale Berichterstattung

Das einzige Textbeispiel, welches die Friedensinitiativen nicht in einen eskalationsorientierten Kontext stellte, war der Artikel aus dem *Helsingin Sanomat*. Auf der Ebene der aktuellen Tagesereignisse über zwei Friedensinitiativen – den Besuch von Primakov in Bagdad und das Treffen der blockfreien Staaten – berichtend, enthält dieser Text nur wenige eskalationsorientierte Textmerkmale, die in erster Linie die Vergeblichkeit der beiden Friedensinitiativen zum Inhalt haben und insofern als unvermeidbar erscheinen. Indem der Text der Darstellung des Versuches, Saddam Hussein zum Rückzug aus Kuwait zu bewegen, einigen Platz einräumt, enthält er auch eine Reihe von deeskalationsorientierten Merkmalen: Infragestellung militärischer Gewalt als geeignetem Mittel der Konfliktlösung, Forderungen nach einer friedlichen Streitbeilegung, Perspektiven der Versöhnung und eine kritische

Distanz gegenüber beiden Konfliktparteien. Auch diese Textmerkmale resultieren aber aus den Tatsachen (report aspect), über welche berichtet wird, und sind nicht das Ergebnis einer spezifischen journalistischen Darstellungsweise (command und parade aspect).

Ungeschminkte Propaganda

Die beiden Texte aus *Aftenposten* und *Dagens Nyheter* sind im Gegensatz dazu mehr oder minder ungeschminkte Propaganda, die Gorbatschovs Friedensplan dadurch diskreditiert, dass sie die Perspektve eines Machterhalts Saddam Husseins als untragbar zurückweist (*Aftenposten*) und/oder davor warnt, dass "die neue irakische Nachgiebigkeit ein bloßer Trick" sei. Beide Texte sind durch eskalationsorientierte Textmerkmale dominiert: Betonung militärischer Werte, Konstruktion des Konfliktes als Win-Lose-Prozess, negative Bezugnahme auf Vermittlungsbemühungen und Zurückweisung friedlicher Alternativen, Betonung der Bedrohung, welche von Saddam Hussein ausgeht, und Stimulierung von Misstrauen gegen ihn, Dämonisierung seiner Intentionen und Anreize, sich mit der eigenen (alliierten) Elite zu identifizieren.

Elaborierte Propagandatechniken

Die Artikel aus der *Washington Post* und der *Süddeutschen Zeitung*, welche die Friedensinitiativen nur erwähnen, um die Kriegsschuld auf Saddam Hussein zu attribuieren, bedienen sich dagegen diffizilerer Propagandatechniken. Beide Texte greifen eine letzte Friedensinitiative Gorbatschows auf, um den Luftkrieg als unvermeidlbar darzustellen, und berichten über Gorbatschows Initiative, den Angriff auf den Irak um wenige Stunden zu verschieben, um einen letzten Vermittlungsversuch unternehmen zu können, erst nachdem sie der Leserschaft klar gemacht haben, dass die Sowjetunion den Angriff unterstützt und nachdem sie die sowjetische Reaktion auf den alliierten Angriff in den Kontext weltweiter positiver Stellungnahmen gestellt haben. Während manche Nationen zwar Bedenken äußerten, unterstützen sie den Angriff dennoch. Nur Kuba, Nordkorea und der Iran verurteilten die Bombenangriffe.

Während der Text aus der *Washington Post* durch eskalationsorientierte Textmerkmale dominiert (insbesondere militärische Logik und Konstruktion des Konfliktes als kompetitiver Prozess) und durch zweiseitige Botschaften (Lumsdaine & Janis, 1953; Reimann, 2002) geprägt ist, folgt die *Süddeutsche Zeitung* einer sogar noch elaborierteren Strategie. Der Artikel gibt Gorbatschows Initiative und kritischen Stimmen, welche den Ausbruch des Krie-

ges als tragisches Ereignis bedauern (das jedoch unvermeidbar gewesen sei, nachdem sich der Irak als unnachgiebig gezeigt habe), einen breiten Raum. Die sowjetischen Friedensbemühungen werden positiv und detailreich dargestellt. Ja, es werden sogar Anreize zur Identifikation mit Gorbatschow gesetzt. Auf den ersten Blick scheint der Text daher eine wilde Mischung aus eskalations- und deeskalationsorientierten Elementen zu sein. Bei genauerer Betrachtung zerfällt er jedoch in fünf Sequenzen, welche die Empörung über den Krieg in fünf argumentativen Schritten in Empörung über den Feind verwandeln.

Der ganze Artikel hat die Form einer zweiseitigen Botschaft, für welche die positive Darstellung Gorbatschows und seiner Friedensinitiative eine zentrale Rolle spielt. Kritik an der alliierten Politik, die Entrüstung über den Kriegsausbruch und das Versagen der internationalen Staatengemeinschaft, den Krieg abzuwenden, werden gegen Saddam Hussein gerichtet. Innerhalb dieses Szenarios dienen die detaillierte Beschreibung der sowjetischen Vermittlungsbemühungen und der Anreiz zur Identifikation mit Gorbatschow dazu, die Empörung über Saddam Hussein zu bekräftigen. Je mehr sich die sowjetischen Vermittler um eine friedliche Streitbeilegung bemühen und je mehr dies durch den Artikel gewürdigt wird, desto mehr ist es gerechtfertigt, Saddam Hussein die Schuld für das Scheitern der Vermittlungsbemühungen zu geben, desto mehr kann die Kriegsschuld auf Saddam Hussein attribuiert werden und desto mehr erscheint es als gerechtfertigt, auf einen militärische Konfliktaustrag zu setzen.

Kontextualisierung der qualitativen Analyseergebnisse

Die in den analysierten Texten zutage tretenden Unterschiede bezüglich ihrer Eskalationsorientierung und der Verwendung von Propagandatechniken sind nicht nur den Ereignissen geschuldet, über welche berichtet wird, sondern auch der Redaktionslinie der jeweiligen Zeitungen und/oder den jeweiligen nationalen Mediendiskursen.

Der Text aus dem *Helsingin Sanomat* erschien in einer Phase des Golfkriegs, während derer es erstmals offenkundig geworden war, dass der Krieg nicht bloß ein sauberer und nur technischer Krieg zwischen Waffensystemen war, sondern ein wirklicher Krieg, der wirkliche Opfer forderte und auch die irakische Zivilbevölkerung in Mitleidenschaft zog. Zu diesem Zeitpunkt wäre die Propaganda nicht gut beraten gewesen, sich als solche zu erkennen zu geben, und außerdem waren die finnischen Medien in ihrer

Unterstützung des Golfkrieges weit zögerlicher und berichteten darüber weit weniger parteilich als die anderen Medien, die in dieser Studie untersucht wurden (vgl. Kempf, 2001).

Die Texte aus *Aftenposten* und *Dagens Nyheter* wurden dagegen in einer Phase des Krieges publiziert, während derer das Feindbild Saddam Hussein bereits etabliert war. Am Vorabend der Bodenoffensive war es für die Propaganda essentiell, jegliche Initiative, welche die strategischen Pläne des Militärs stören könnte, zu demontieren, und sie ging kaum ein Risiko ein, wenn sie dies ganz offen tat. Die öffentliche Meinung hatte sich mit dem Krieg inzwischen arrangiert und die Friedensbewegung (die in Norwegen und Schweden ohnedies nicht sehr stark war) war bereits zusammengebrochen.

Zu Beginn des Krieges, als die Artikel aus der *Washington Post* und der *Süddeutschen Zeitung* erschienen, war die Situation eine ganz andere. Es gab eine starke Friedensbewegung (nicht nur in Deutschland, sondern auch in den USA), das Feindbild Saddam Hussein war viel schwächer ausgeprägt und ein großer Teil der Öffentlichkeit (insbesondere in Deutschland) war dem Krieg abgeneigt und gegenüber den Kriegszielen von Skepsis erfüllt. Zu diesem Zeitpunkt war es für die Propaganda essentiell, kritische Stimmen in ihre Strategie einzubinden, das Feindbild zu verschärfen, die alliierten Angriffe als *ultima ratio* zu legitimieren und Saddam Hussein die Schuld daran zu geben. Entsprechend musste die Propaganda viel vorsichtiger und mit mehr Raffinesse vorgehen.

Kontrollfragen

- Warum müssen die inhaltsanalytischen Variablen logisch unabhängig sein, wenn man die Daten mittels Latent-Class-Analyse auswerten will?
- Warum ist zur Beschreibung der LCA-Ergebnisse eine Sprachnormierung sinnvoll und warum ist dabei der Häufigkeitsbereich, der durch dieselben Wortmarken beschrieben wird, für mittelgroße Häufigkeiten am breitesten und für sehr kleine oder sehr große Häufigkeiten am schmalsten zu wählen?
- Was leistet die Anwendung der Latent-Class-Analyse, das durch die Betrachtung der (Gesamt-)Häufigkeitsverteilung der Daten nicht geleistet werden kann?
- Diskutieren Sie, inwiefern und warum auch die LCA-Ergebnisse nur begrenzt Auskunft über die Argumentationsrichtung der Texte geben können!
- Welchen Ausweg aus diesem Dilemma eröffnet die Latent-Class-Analyse und wie geht man dabei vor?

Kapitel 2
Psychometrie

2.1
Grundlagen

2.1.1. Die Anfänge der Testpsychologie

Der erste Intelligenztest

Der Gedanke, dass man psychische Eigenschaften messen könne, ist sehr alt. Bereits 1883 veröffentlichte Francis Galton (1822–1911) ein Buch, in dem er eine Reihe von Fragen und Aufgaben vorschlug, mit deren Hilfe man den Grad der geistigen Begabung messen könne. 1894 entwickelte Alfred Binet (1857–1911) den ersten Intelligenztest für Kinder, den er zehn Jahre später gemeinsam mit Theodore Simon (1873–1961) durch eine Normierung der Aufgaben stark verbesserte. Für jedes Lebensjahr wurde eine Reihe von Aufgaben (Items) zusammengestellt und an einer sehr großen Stichprobe von Kindern überprüft. Wenn 75% aller Kinder einer Altersgruppe (aber nur 25% der Kinder der nächstjüngeren Altersgruppe) die Aufgaben lösen konnten, so galten sie als geeicht (d. h. der Altersstufe in ihrem Schwierigkeitsgrad entsprechend).

Für jedes Lebensjahr wurde eine Serie von fünf Aufgaben zusammengestellt. Z. B. für die Zehnjährigen: Fünf Gewichte vom leichtesten (6g) bis zum schwersten (18g) ordnen; die Zeichnung einer komplizierten geometrischen Figur aus dem Gedächtnis wiedergeben; mit drei vorgegebenen Worten (z. B. "Schnee, spielen, Schlitten") zwei Sätze bilden; absurde Sätze kritisieren (z. B. "Ich habe drei Brüder – Paul, Ernst und ich."); sowie fünf praktische Verstandesfragen (z. B. "Was muss man machen, wenn man in die Schule geht und man merkt unterwegs, dass es schon später ist als gewöhnlich?").

Das Intelligenzalter

Aus den Antworten der Kinder wurde ihr Intelligenzalter (IA) errechnet: Ausgehend von dem Jahr, bis zu dem ein Kind alle Aufgaben lösen konnte, wurde für jede gelöste Aufgabe 1/5 Jahr dazugezählt. Das Intelligenzalter wurde dann zum Lebensalter (LA) in Beziehung gesetzt, z. B.:

IA – LA = 10.4 – 10.0 = 0.4 Jahre Intelligenz-Vorsprung

War das Intelligenzalter kleiner als das Lebensalter, so sprach man von einem Intelligenz-Rückstand.

Der Intelligenzquotient

William Stern (1871–1938) kritisierte das Verfahren, da die Differenz zwischen Intelligenzalter und Lebensalter keinen eindeutigen Sinn hat, sondern ganz Verschiedenes bedeutet, je nach der Altersstufe des geprüften Kindes.

"Ein Rückstand von zwei Jahren ist beim sechsjährigen Kind ein Zeichen starker geistiger Minderwertigkeit, er verrät beim neunjährigen einen sehr viel geringeren Schwachsinnsgrad; und für das zwölfjährige Kind braucht die durch ihn ausgedrückte Schwäche überhaupt noch nicht jenseits der Grenze der Normalität zu liegen. Ein bestimmter Differenzwert hat also um so geringere Bedeutung, je höher des Alter des Kindes ist. Oder anders ausgedrückt: bei gleichem Intelligenzgrad muß die Differenz wachsen, wenn das Lebensalter wächst" (zit. n. Rohracher, 1971, 383).

Als Maß für die Intelligenz schlug Stern daher den *Intelligenzquotienten* (IQ = IA/LA) vor. Später hat es sich eingebürgert, den Intelligenzquotienten mit 100 zu multiplizieren:

$$IQ = 100*IA/LA. \qquad [2.1.1.1]$$

Scorebildung

Als man später begann, auch Intelligenztests für Erwachsene zu entwickeln, musste man das Vorgehen etwas modifizieren. Im Erwachsenenalter bleibt die intellektuelle Leistungsfähigkeit über einen großen Zeitraum hin konstant. Alterstypische Aufgaben sind daher nicht mehr konstruierbar. Stattdessen ging man zur Auswahl von Aufgaben über, die in ihrer Schwierigkeit hinreichend stark variieren, so dass leichte Aufgaben von vielen, schwierige Aufgaben dagegen nur von wenigen Vpn gelöst werden.

- Für jede gelöste Aufgabe erhält die Vp einen Punkt (dichotome Testitems).
- Mitunter werden die Antworten auf die Testitems auch auf einer mehrstufigen Skala bewertet (polytome Testitems); z. B. 0 Punkte = falsch; 1 Punkt = teilweise richtig; 2 Punkte = vollständig richtig.

Die Testleistung der Vp wird schließlich in ihrem Testscore ausgedrückt: Meistens durch die Gesamtzahl der im Test erreichten Punkte. Manchmal werden die Items aber auch unterschiedlich gewichtet.

Arten von Testitems

Der große Erfolg der Intelligenztests hat dazu geführt, dass man bald schon begonnen hat, auch andere psychologische Eigenschaften mittels Tests messen zu wollen, z. B. Einstellungen oder Persönlichkeitseigenschaften. Diese Tests sind im Prinzip genauso aufgebaut. Die Items sind dort aber meist keine Aufgaben, die es zu lösen gilt, sondern Fragen, welche die Vpn beantworten.

In der klinischen Psychologie gibt es Tests (Symptomchecklisten), die nicht vom Patienten selbst ausgefüllt werden, sondern der behandelnde Therapeut kreuzt an, welche Symptome ein Patient zeigt. Die Items sind hier also Symptome, die auftreten können oder nicht.

2.1.2 Die Fragestellungen der Testtheorie

Grundbegriffe

Allgemein lässt sich festhalten, dass psychologische Tests Instrumente zur Erhebung einer standardisierten Verhaltensstichprobe sind. Sie bestehen aus einer mehr oder minder großen Anzahl von Items (i = 1,...,k). Die Antwort einer Vp v auf Item i bezeichnen wir mit x_{vi}. Sie wird auf einer zwei- oder mehrstufigen Skala beurteilt. Bei dichotomen Items ist x_{vi} = 0 oder 1. Bei polytomen Items unterscheiden wir m Antwortkategorien: x_{vi} = 0,1,..., m-1.

Aus den Itemantworten wird der Testscore der Vp (x_{vo}) errechnet, indem man die bei den einzelnen Items erzielten Punktzahlen aufsummiert

$$x_{vo} = \sum_{i=1}^{k} x_{vi} \qquad \text{(Summenscore).} \qquad [2.1.2.1]$$

Mitunter werden die Items unterschiedlich gewichtet

$$x_{v\bullet} = \sum_{i=1}^{k} \alpha_i x_{vi} \qquad \text{(gewichteter Summenscore),} \qquad [2.1.2.2]$$

wobei α_i das Gewicht bezeichnet, mit dem Item i in den Testscore eingeht. Um den Score zu bezeichnen, den eine Vp in einem bestimmten Test t erhalten hat, schreibt man ggf. auch x_{vt}.

Eine vollständige Beschreibung des Testergebnisses einer Vp liefert das Antwortmuster, d. h. der Antwortvektor der Vp

$$x_v = (x_{v1}, ..., x_{vk}). \qquad [2.1.2.3]$$

Die Antworten von n Vpn auf k Testitems bilden die Antwortmatrix

$$X = ((x_{vi})), \qquad [2.1.2.4]$$

die man erhält, indem man die Antwortvektoren der Vpn zeilenweise untereinander schreibt (vgl. Tabelle 2.1.2.1):

Tab. 2.1.2.1: Die Antwortmatrix

	1	...	i	...	k	Σ
1	x_{11}	...	x_{1i}	...	x_{1k}	x_{1o}
:	:		:		:	:
v	x_{v1}	...	x_{vi}	...	x_{vk}	x_{vo}
:	:		:		:	:
n	x_{n1}	...	x_{ni}	...	x_{nk}	x_{no}
Σ	x_{o1}	...	x_{oi}	...	x_{ok}	

Die Zeilensummen der Antwortmatrix ergeben den Personenrandvektor $(x_{1o},...,x_{no})$, dessen Eintragungen x_{vo} die Summenscores der Vpn sind.

Die Spaltensummen der Antwortmatrix ergeben den Itemrandvektor $(x_{o1},...,x_{ok})$, dessen Eintragungen x_{oi} die Gesamtzahl der Punkte ist, welche von den Vpn bei Item i erzielt wurden

$$x_{oi} = \sum_{v=1}^{n} x_{vi} . \qquad [2.1.2.5]$$

Bei dichotomen Items sind dies die (absoluten) Lösungshäufigkeiten der Items.

Aus der so charakterisierten Verhaltensstichprobe soll auf eine zugrunde liegende Eigenschaft der Vpn geschlossen werden, z. B. auf ihre Intelligenz, auf bestimmte Einstellungen oder Persönlichkeitsmerkmale oder auch auf psychische Erkrankungen.

Fragestellungen der Testtheorie

Für die Psychologische Testtheorie ergibt sich daraus eine Reihe von grundlegenden Fragestellungen. Diese betreffen:

1. Das Problem der Messung latenter Eigenschaften: Wie ist es denn überhaupt möglich, von den beobachteten (manifesten) Daten auf zugrunde liegende (latente) Eigenschaften der Vpn zu schließen?
2. Das Problem der Homogenität von Tests: Wie kann sichergestellt werden, dass die Items eines Tests tatsächlich alle die selbe latente Eigenschaft der Vpn messen. Oder, pragmatisch formuliert:

a. Wie kann die Verwendung des Testscores als Maß für die latente Eigenschaft der Vpn gerechtfertigt werden, und
b. wie können wir Psychodiagnostik betreiben, wenn dies nicht der Fall ist?

3. Das Problem der Reliabilität: Wie können wir die Messgenauigkeit von Tests beurteilen und eine hinreichend hohe Messgenauigkeit sicherstellen?
4. Das Problem der Validität: Wie können wir gewährleisten, dass der Test auch tatsächlich jene Eigenschaft misst, die er zu messen vorgibt, bzw. die wir mit Hilfe des Tests erfassen wollen?

2.1.3 Die Beurteilung von Testleistungen

Vergleichsnormen zur Beurteilung des Intelligenzquotienten

Um das Testergebnis einer Vp interpretieren und daraus diagnostische Urteile ableiten zu können, benötigen wir einen Vergleichsmaßstab.
Im Falle des klassischen Intelligenzquotienten ist dieser relativ einfach zu gewinnen bzw. ergibt er sich unmittelbar aus der Art und Weise, wie der IQ konstruiert ist:

Aufgrund der Definition des IQ durch Gleichung [2.1.1.1] ergibt sich unmittelbar, dass ein IQ = 100 eine durchschnittliche Testleistung beschreibt.

Als durchschnittlich werden wir aber nicht nur eine Testleistung von exakt 100 IQ-Punkten ansehen, sondern wir werden einen Wertebereich angeben, innerhalb dessen die Testleistung als durchschnittlich gilt. Wir werden jene ca. 2/3 der Vpn als durchschnittlich begabt bezeichnen, deren IQ am wenigsten von 100 abweicht.

Da der Intelligenzquotient innerhalb jeder gegebenen Altersstufe näherungsweise normalverteilt ist, mit einem Erwartungswert von $\mu = 100$ und einer Standardabweichung von $\sigma = 15$

$$IQ \ldots N(100;\ 15^2) \qquad [2.1.3.1]$$

ergibt sich daraus ein Durchschnittsbereich von 85 bis 115 IQ-Punkten. Die Normalverteilung hat nämlich die Eigenschaft, dass 68.27% der Werte im Bereich $\mu \pm \sigma$ liegen.

Entsprechend werden wir jene 95.45% der Bevölkerung, deren IQ im Bereich zwischen 70 und 130 Punkten liegt, d. h. im Bereich $\mu \pm 2\sigma$, als normal begabt bezeichnen und erst die Vpn, deren

Testleistung unterhalb dieses Bereichs liegt, als minder begabt, sowie jene, deren Testleistung oberhalb dieses Bereichs liegt, als hoch begabt einstufen.

Innerhalb dieser Bereiche kann man dann noch weiter differenzieren. Z. B. unterscheidet das Diagnostische und Statistische Manual Psychischer Störungen (DSM-IV) vier verschieden Schweregrade der intellektuellen Beeinträchtigung (vgl. Tabelle 2.1.3.1)

Tab. 2.1.3.1: Schweregrade intellektueller Beeinträchtigung

Schweregrad	IQ
Leicht	55–50 bis etwa 70
Mäßig	35–40 bis etwa 50–55
Schwer	20–25 bis etwa 35–40
Schwerst	Unter 20–25

Dass die Grenzen dabei nur in etwa angegeben werden, verweist darauf, dass diagnostische Urteile nicht einfach automatisch und schematisch aus den Testergebnissen abgelesen werden können. Geistige Behinderung ist laut DSM-IV nicht nur durch deutlich unterdurchschnittliche allgemeine intellektuelle Leistungsfähigkeit (IQ unter 70) definiert, sondern auch durch gleichzeitige Defizite bzw. Beeinträchtigungen der sozialen Anpassungsfähigkeit, worunter man die Leistungsfähigkeit bei der Erfüllung sozialer Normen versteht, die das soziale Umfeld bezüglich Anpassung und Verantwortung, Kommunikation, alltagspraktischer Fähigkeiten sowie persönlicher Unabhängigkeit von Personen des entsprechenden Alters erwartet.

Die Normalverteilung des Intelligenzquotienten

Dass Intelligenzquotienten (näherungsweise) normalverteilt sind, hat im Laufe der Psychologiegeschichte zu allerhand Spekulationen Anlass gegeben, dass – und warum – "die Intelligenz" normalverteilt sei. Tatsächlich handelt es sich dabei zwar um eine empirisch feststellbare Regelmäßigkeit; *nicht* jedoch um eine empirische, *sondern* um eine strukturelle Gesetzmäßigkeit – oder wie man auch sagen kann –, um ein *methodisches Artefakt*, das sich schlicht und einfach aus der Art und Weise ergibt, wie das Intelligenzalter festgestellt und wie daraus der Intelligenzquotient berechnet wird.

Das Intelligenzalter ist nämlich nichts anderes als der Summenscore der Vpn, dividiert durch die Anzahl der Testaufgaben je Altersstufe

$$IA = \frac{x_{vo}}{5}.$$

Der Summenscore wiederum errechnet sich aus der Summe der Itemantworten,

$$x_{vo} = \sum_{i=1}^{k} x_{vi},$$

wobei eine sehr große Anzahl von Antwortvariablen aufsummiert wird.

Die Verteilung der Summe von k Zufallsvariablen strebt jedoch mit wachsendem k gegen eine Normalverteilung – und zwar nicht nur dann, wenn die k Variablen voneinander unabhängig sind und alle derselben Verteilungsfunktion folgen, wie dies im zentralen Grenzwertsatz (vgl. Kreyszig, 1968, 201) vorausgesetzt ist.[1] Wenn keine der Variablen die anderen dominiert, ergibt sich die asymptotische Normalverteilung der Summe von k Zufallsvariablen unter relativ schwachen Bedingungen auch dann, wenn die Variablen unterschiedlich verteilt[2] und/oder mäßig hoch korreliert[3] sind (siehe hierzu weiterführend auch Tijms, 2004).

Wegen der Invarianz der Normalverteilung gegenüber linearen Transformationen (vgl. Kreyszig, 1968, 190) bleibt die Normalverteilungseigenschaft erhalten, wenn man den Summenscore durch eine Konstante dividiert und so das Intelligenzalter berechnet.

Innerhalb jeder Altersgruppe ist aber auch das Lebensalter der Probanden eine Konstante. Der Intelligenzquotient errechnet sich aus dem Intelligenzalter daher ebenfalls durch eine lineare Transformation, weshalb er bei großem k (näherungsweise) normalverteilt ist.

Umrechnung des Summenscores in IQ-Punkte

Schon bei Intelligenztests für Erwachsene, bei denen wir keine alterstypischen Testaufgaben mehr konstruieren können, ist die Berechnung eines Intelligenzquotienten nicht mehr möglich. Das Prinzip, eine Vergleichspopulation (= beim IQ die Gruppe der

1. Zum Beweis des zentralen Grenzwertsatzes siehe z. B. Schmetterer (1966).
2. Eine hinreichende Bedingung hierfür wurde erstmals von Ljapunov (1901) hergeleitet. Eine notwendige und hinreichende Bedingung enthält der Satz von Lindeberg-Feller, der erstmals von Lindeberg (1922) bewiesen wurde; vgl. Fisz (1970, 241ff.).
3. Bedingungen hierfür wurden u. a. von Hall & Heyde (1980) hergeleitet.

Gleichaltrigen) als Maßstab für die Beurteilung der Testleistung heranzuziehen, können wir aber beibehalten, indem wir den Test an einer sehr großen Stichprobe von Vpn eichen, d. h., zunächst Schätzwerte für den Erwartungswert μ und die Standardabweichung σ des Summenscores berechnen, und dann den Summenscore mittels der linearen Transformationen

$$z = \frac{x_{vo} - \mu}{\sigma} \qquad [2.1.3.2]$$

und

$$IQ = 100 + 15z \qquad [2.1.3.3]$$

in IQ-Punkte umrechnen, für welche dieselben Vergleichsmaßstäbe anwendbar sind wie für den klassischen Intelligenzquotienten – jetzt aber nicht mehr per definitionem relativ zur Altersgruppe, sondern relativ zu einer wie auch immer gearteten Vergleichspopulation (Referenzpopulation), für welche die Eichstichprobe repräsentativ ist.

Kontrollfragen

- Beschreiben Sie die Vorgehensweise von Binet & Simon bei der Konstruktion alterstypischer Testaufgaben!
- Wie berechnet sich nach Binet & Simon das Intelligenzalter?
- Wann sprechen Binet & Simon
  a) von einem Intelligenz-Vorsprung
  b) von einem Intelligenz-Rückstand?
- Welche Kritik hat William Stern an dieser Vorgehensweise geübt?
- Wie ist der Intelligenzquotient definiert?
- Warum muss man die Vorgehensweise zur Konstruktion von Testaufgaben modifizieren, wenn man Intelligenztests für Erwachsene konstruiert?
- Wie geht man dabei vor?
- Was versteht man unter dem Testscore, wie wird er berechnet?
- Worin bestehen die Items eines Tests
  a) bei Intelligenztests
  b) bei Einstellungs- und Persönlichkeitstests
  c) bei Symtomchecklisten?
- Worin besteht der Unterschied zwischen dichotomen und polytomen Testitems, wie viele Antwortkategorien werden dabei unterschieden und mit welchen Zahlenwerten werden sie bezeichnet?
- Wie berechnet sich
  a) der Summenscore einer Vp
  b) der gewichtete Summenscore einer Vp?

- Wie ist der Antwortvektor einer Vp definiert?
- Was versteht man unter der Antwortmatrix, wie errechnen sich daraus
- a) der Itemrandvektor
  b) der Personenrandvektor
  und was bilden sie ab?
- Welches sind die vier Hauptprobleme der psychologischen Testtheorie?
- Wie ist der Intelligenzquotient verteilt?
- In welchem Bereich muss der IQ einer Vp liegen, damit sie
  a) als durchschnittlich begabt
  b) als normalbegabt
  c) als minderbegabt
  d) als hochbegabt
  gilt?
- Welcher IQ-Bereich entspricht laut DSM-IV einer
  a) leichten
  b) mäßigen
  c) schweren
  d) schwersten
  intellektuellen Beeinträchtigung?
- Was muß laut DSM-IV alles gegeben sein, um eine geistige Behinderung zu diagnostizieren?
- Worauf beruht die Normalverteilung des Intelligenzquotienten?
- Was versteht man unter der Eichung eines Tests, und wie werden die Testscores dabei in IQ-Punkte umgerechnet?

Übungsaufgaben

1. Ein Kind ist 9 Jahre und 4 Monate alt. Im Binet-Simon-Test hat es den folgenden Antwortvektor:

Altersgruppe	7 Jahre	8 Jahre	9 Jahre	10 Jahre	11 Jahre
Aufgabe Nr.	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5
Antwortvektor	1 1 1 1 1	1 1 0 1 1	1 0 0 0 0	0 0 0 1 0	0 0 0 0 0

   Da das Kind alle Aufgaben für 7-Jährige lösen konnte, wird vorausgesetzt, dass es auch alle Aufgaben für jüngere Kinder hätte lösen können.
   Da es keine der Aufgaben für 11-Jährige lösen konnte, wird vorausgesetzt, dass es auch die Aufgaben für ältere Kinder nicht hätte lösen können.
   a) Berechnen Sie das Lebensalter des Kindes!
   b) Berechnen Sie sein Intelligenzalter!
   c) Berechnen Sie den Intelligenzquotienten!
   d) Beurteilen Sie, ob ein Intelligenz-Vorsprung oder ein Intelligenz-Rückstand besteht!
2. Stellen Sie den Antwortvektor des Kindes in Aufgabe 1 dar!
3. Berechnen Sie den Summenscore des Kindes in Aufgabe 1!
4. Berechnen Sie
   a) den Itemrandvektor und

b) den Personenrandvektor
der nachstehenden Antwortmatrix und erläutern Sie, was diese beschreiben!

Items				
1	2	3	4	5
0	0	0	0	1
0	0	0	1	1
0	0	1	1	1
0	1	1	1	1
1	0	0	0	0
1	1	0	0	0
0	1	1	1	1
0	0	1	1	1
0	0	1	1	0
0	0	1	1	1

5. Der Summenscore in einem Intelligenztest aus 80 Items ist in der Gruppe der 20-25 Jährigen nach $N(45.2; 4.2^2)$ verteilt. Ein 22-jähriger Abiturient hat das Testergebnis $x_{vt} = 54$ erzielt. Rechnen Sie den Score in IQ-Punkte um!

2.2
Testen und Messen

2.2.1 Grundannahmen der klassischen Testtheorie[1]

Der Kalkül der klassischen Testtheorie

Seit der Konstruktion des ersten Intelligenztests durch Binet & Simon (1905) stellt die Qualitätsprüfung psychologischer Testuntersuchungen ein Problem von nicht unerheblicher praktischer Bedeutung dar. Man hat daher schon früh damit begonnen, statistische Rechenformeln zu entwickeln, welche es erlauben sollten, die Verlässlichkeit von Tests bzw. Testergebnissen empirisch zu untersuchen. Diese wurden erstmals von Gulliksen (1950) in eine geschlossene Theorie psychologischer Testergebnisse integriert, die sogenannte klassische Testtheorie.

Dabei folgte Gulliksen strikt den Vorgaben, wie sie von Hilbert (zit.n. Meschkowski, 1986) für die axiomatische Begründung formaler Kalküle formuliert wurden: Die Grundbegriffe bleiben undefiniert, die Axiome sind widerspruchsfrei und voneinander unabhängig, und das Axiomensystem ist vollständig.

Maßzahl, True-Score und Messfehler

Am Anfang der klassischen Testtheorie steht die Unterscheidung von zwei Arten von Zufallsvariablen X_{ot} und T_{ot}, die auf derselben Population von Merkmalsträgern definiert sind. Diese nennen wir die *Referenzpopulation*. Die Elemente der Referenzpopulation bezeichnen wir mit v = 1,…,n und nennen sie *Vpn* oder *Probanden*.

Die Realisationen der Zufallsvariable X_{ot} bezeichnen wir mit x_{vt} und nennen sie den *Score* oder *Messwert* der Vp v in Test t. Entsprechend heißt X_{ot} die *Score-Variable* oder *Maßzahlvariable* des Tests t in der Referenzpopulation.

Die Realisationen der Zufallsvariable T_{ot} bezeichnen wir mit τ_{vt} und nennen sie den *True-Score* oder *wahren Messwert* der Vp v in Test t. Entsprechend heißt T_{ot} die *True-Score-Variable* des Tests t in der Referenzpopulation.

1. Vgl. hierzu ausführlicher Bd. 1, Kap. 2.1.1–2.1.4.

Die Differenz zwischen Maßzahlvariable und True-Score-Variable bezeichnen wir mit F_{ot} und nennen sie die *Fehlervariable* des Tests t in der Referenzpopulation

$$F_{ot} = X_{ot} - T_{ot} \quad . \qquad [2.2.1.1]$$

Ihre Realisationen bezeichnen wir mit f_{vt} und nennen sie den *Messfehler* der Vp v in Test t

$$f_{vt} = x_{vt} - \tau_{vt} \, . \qquad [2.2.1.2]$$

Die Axiome von Gulliksen

Darauf aufbauend formulieren wir die Axiome:

$A1: E(F_{ot}) = 0$	Der mittlere Messfehler ist gleich Null.
$A2: \rho(T_{ot}, F_{ot}) = 0$	Der Messfehler korreliert nicht mit dem True-Score.
$A3: \rho(F_{ot}, F_{ot'}) = 0$	Die Messfehler aus verschiedenen Tests t und t' korrelieren nicht miteinander.
$A4: \rho(F_{ot}, T_{ot'}) = 0$	Der Messfehler in einem Test t korreliert nicht mit dem True-Score in einem anderen Test t'.

Wie Gulliksen zeigt, können aus diesen Axiomen tatsächlich die gesamten Ergebnisse der klassischen Testtheorie hergeleitet werden.

Ein Modell der klassischen Testtheorie

Um ihre Anwendung auf psychologische Tests zu rechtfertigen, muss entsprechend der Arbeitsteilung zwischen Formalwissenschaften und empirischen Wissenschaften aber erst noch gezeigt werden, dass psychologische Tests tatsächlich ein Modell dieses Kalküls darstellen. D. h. es muss untersucht werden, wie der wahre Messwert oder True-Score definiert werden kann, so dass die Geltung der Axiome von Gulliksen sichergestellt ist.

Die Grundannahmen von Novick

Ein solches Modell wurde unabhängig von Koutsopoulos (1962) und Novick (1966) konstruiert. Wir folgen hier der Darstellung bei Lord & Novick (1968). Die Modellannahmen lauten:

1. Jeder Testung t eines Probanden v entspricht eine zufällige Variable möglicher Testergebnisse X_{vt} mit endlichem Erwartungswert $E(X_{vt})$ und endlicher Varianz $\sigma^2(X_{vt})$. Diese nennen wir die (individuelle) *Score-Variable* der Vp v in Test t.
2. Das Testergebnis x_{vt}, welches der Proband erzielt, ist eine unabhängige Realisation dieser Score-Variable.

3. Der True-Score des Probanden τ_{vt} ist *per definitionem* gleich dem Erwartungswert der Score-Variable

$$\tau_{vt} = E(X_{vt}). \qquad [2.2.1.3]$$

Legt man diese Modellannahmen zugrunde, so ist die Geltung der Axiome von Gulliksen sichergestellt.

Der Messfehler der klassischen Testtheorie

Gleichzeitig wird mit der expliziten Definition des True-Scores als Erwartungswert der Score-Variable aber auch deutlich, dass der Messfehler der klassischen Testtheorie ausschließlich *reine Zufallsfehler* beschreibt, welche sich daraus ergeben, dass einem Probanden eben nicht ein bestimmtes Testergebnis entspricht, sondern eine Wahrscheinlichkeitsverteilung möglicher Testergebnisse, die um den – als True-Score bezeichneten – Erwartungswert dieser Verteilung streuen.

Wie aus der mathematischen Herleitung der Axiome aus den Grundannahmen von Novick (vgl. Bd. 1, Kap. 2.1.4) hervorgeht, beschreibt der Messfehler der klassischen Testtheorie jedoch *nicht* etwa *Störeinflüsse* wie Müdigkeit oder mangelnde Aufmerksamkeit. Solche Störeinflüsse schlagen sich nicht auf den Messfehler, sondern auf die tatsächliche Leistungsfähigkeit (True-Score) der Probanden nieder. Entsprechend beschreibt der True-Score der klassischen Testtheorie auch nicht *die* Leistungsfähigkeit eines Probanden in dem jeweiligen Test schlechthin, sondern lediglich seine Leistungsfähigkeit unter den jeweiligen Versuchs- und Störbedingungen, welche das Testergebnis beeinflussen können.

Präzisierung der Modellannahmen von Novick

Um dies explizit zu machen, können die Modellannahmen von Novick wie folgt präzisiert werden:

1. Jeder Testung t eines Probanden v *unter gegebenen Versuchs- und Störbedingungen* entspricht eine zufällige Variable möglicher Testergebnisse X_{vt} mit endlichem Erwartungswert $E(X_{vt})$ und endlicher Varianz $\sigma^2(X_{vt})$. Diese nennen wir die (individuelle) Score-Variable der Vp v in Test t *unter den jeweiligen Versuchs- und Störbedingungen*.
2. Das Testergebnis x_{vt}, welches der Proband *unter den jeweiligen Versuchs- und Störbedingungen* erzielt, ist eine unabhängige Realisation dieser Score-Variable.
3. Der True-Score des Probanden τ_{vt} *unter den jeweiligen Versuchs- und Störbedingungen* ist per definitionem gleich dem Erwartungswert dieser Score-Variable.

Aus Gründen der Sprachökonomie werden wir auf den Zusatz "unter den jeweiligen Versuchs- und Störbedingungen" im Folgenden verzichten. Er ist jedoch stets mitzudenken – nicht nur im Falle der klassischen Testtheorie, sondern auch bei allen anderen testtheoretischen Modellen, die in diesem Buch behandelt werden. Was wir messen können, sind immer nur die jeweilige Leistungsfähigkeit, die jeweiligen Einstellungen oder die jeweilige Symptomatik unserer Probanden.

2.2.2 Reliabilität und Validität

Ziele der klassischen Testtheorie

Die klassische Testtheorie verfolgt vor allem zwei Ziele:

1. den Grad der Genauigkeit zu bestimmen, mit der ein Test jenes Verhaltensmerkmal misst, das er misst, was immer das sein mag (Reliabilität), und
2. den Grad der Genauigkeit zu bestimmen, mit der er jenes Verhaltensmerkmal zu erfassen vermag, welches er messen soll (Validität).

Die Reliabilität

Um diese beiden Eigenschaften eines Tests zu beschreiben, bedient sich die klassische Testtheorie der Korrelationsrechnung und definiert die Reliabilität als *Quadrat der Korrelation* zwischen den beobachteten Testergebnissen (X_{ot}) und den True-Scores der Probanden (T_{ot})

$$\text{Reliabilität} = \rho^2(X_{ot}, T_{ot}) = \rho^2_{XT}. \qquad [2.2.2.1]$$

Die Validität

Die Validität dagegen wird als *unquadrierte* Korrelation der Testergebnisse (X_{ot}) mit einer davon unabhängigen Messung des Kriteriums ($Y_{ot'}$) definiert, welches mit Hilfe des Tests erfasst werden soll

$$\text{Validität} = \rho(X_{ot}, Y_{ot'}) = \rho_{XY}. \qquad [2.2.2.2]$$

Sowohl der Test selbst als auch die Messung der Kriteriumsvariablen können mit einem Messfehler behaftet sein, so dass $X_{ot} = T_{ot} + F_{ot}$ und $Y_{ot'} = T_{ot'} + F_{ot'}$.

Methoden der Reliabilitätsbestimmung

Anders als die Validität, die als Korrelation zwischen beobachtungssprachlich definierten Variablen definiert ist, und die daher jederzeit aus empirischen Daten berechnet werden kann, ist die Reliabilität als Quadrat der Korrelation der Testergebnisse mit

den True-Scores definiert, die ein theoretisches Konstrukt darstellen, das nicht beobachtbar ist. Die Definitionsgleichung [2.2.2.1] stellt daher *keine Rechenformel* dar, nach der man die Reliabilität eines Tests berechnen könnte. Gleichwohl gibt es eine Reihe von Methoden, wie man die Reliabilität eines Tests statistisch schätzen kann. Darauf werden wir weiter unten eingehen.

Beziehungen zwischen Reliabilität und Validität

Eine dieser Methoden resultiert unmittelbar aus der Gleichung

$$\rho(X_{ot}, Y_{ot'}) = \rho(X_{ot}, T_{ot})\,\rho(T_{ot}, T_{ot'})\,\rho(Y_{ot'}, T_{ot'}), \qquad [2.2.2.3]$$

wonach die Validität in drei Faktoren aufgespalten werden kann,

- deren einer die Quadratwurzel aus der Reliabilität des Tests ist,
- deren zweiter die Korrelation zwischen den True-Scores in Test und Kriterium darstellt und
- deren dritter die Quadratwurzel aus der Reliabilität der Kriteriumsmessung ist.

Parallelität

Die Reliabilität eines Tests t ist daher gleich seiner Validität bezüglich eines gleichwertigen Tests t', der dieselbe True-Score-Variable $T_{ot} \equiv T_{ot'}$ mit derselben Messgenauigkeit erfasst. Tests, die diese Voraussetzungen erfüllen, bezeichnet man als *Paralleltests*.

Beweis

Seien t und t' parallele Testungen, so ist in Gleichung [2.2.2.3] $\rho(T_{ot}, T_{ot'}) = 1$ und $\rho(X_{ot}, T_{ot}) = \rho(Y_{ot'}, T_{ot'})$ und daher

$$\rho(X_{ot}, Y_{ot'}) = \rho(X_{ot}, T_{ot})\,\rho(Y_{ot'}, T_{ot'}) = \rho^2(X_{ot}, T_{ot}) = \rho^2_{XT}.$$

Die Paralleltest-Methode

Damit haben wir bereits eine erste Rechenformel gewonnen, wie man die Reliabilität eines Tests bestimmen kann, wenn man über eine parallele Testung desselben Verhaltensmerkmals verfügt. Die Reliabilität eines Tests t ist gleich seiner Korrelation mit einem Paralleltest t'

$$\rho^2_{XT} = \rho(X_{ot}, Y_{ot'}). \qquad [2.2.2.4]$$

Essentielle Parallelität

Wegen der Invarianz der Korrelation gegenüber linearen Transformationen ist die Paralleltest-Methode auch dann noch anwendbar, wenn die beiden Tests zwar dasselbe Verhaltensmerkmal gleich gut messen, sich aber in ihrer Schwierigkeit un-

terscheiden, so dass $T_{ot} = T_{ot'} + c_{tt'}$. Man spricht in diesem Falle von *essentiell parallelen* Testungen.

τ-Äquivalenz und essentielle τ-Äquivalenz

Unterscheiden sich die beiden Testungen in ihrer Messgenauigkeit, so spricht man

- von *τ-äquivalenten Tests*, wenn $T_{ot} \equiv T_{ot'}$,
- bzw. von *essentiell τ-äquivalenten Tests*, wenn $T_{ot} = T_{ot'} + c_{tt'}$.

Die τ-Äquivalenz ist folglich ein Spezialfall der essentiellen τ-Äquivalenz für $c_{tt'} = 0$.

In beiden Fällen ergibt die Paralleltest-Methode nicht mehr die Reliabilität des Test t, sondern das geometrische Mittel aus den Reliabilitäten der beiden Tests t und t'

$$\rho(X_{ot}, Y_{ot'}) = \rho(X_{ot}, T_{ot})\,\rho(Y_{ot'}, T_{ot'}) = \sqrt{\rho^2_{XT}\ \rho^2_{YT}}\,. \qquad [2.2.2.5]$$

Verletzung der essentiellen τ-Äquivalenz

Gravierendere Auswirkungen hat die Verletzung der essentiellen τ-Äquivalenz. Die beiden Tests t und t' messen dann nicht mehr das selbe Merkmal und für die Korrelation zwischen den True-Score-Variablen gilt $\rho(T_{ot}, T_{ot'}) < 1$.

Wenn die beiden Tests dieselbe Messgenauigkeit aufweisen, so unterschätzt die Paralleltest-Methode in diesem Falle die Reliabilität des Tests, so dass

$$\rho(X_{ot}, Y_{ot'}) < \rho^2_{XT}\,, \qquad [2.2.2.6]$$

Unterscheiden sich die beiden Tests in ihrer Messgenauigkeit, so unterschätzt sie das geometrische Mittel ihrer Reliabilitäten

$$\rho(X_{ot}, Y_{ot'}) < \sqrt{\rho^2_{XT}\ \rho^2_{YT}}\,. \qquad [2.2.2.7]$$

Zusammenfassung

Zusammenfassend lässt sich festhalten, dass die Paralleltest-Methode eine *untere Schranke* für das geometrische Mittel aus den Reliabilitäten der beiden Tests liefert

$$\rho(X_{ot}, Y_{ot'}) \le \sqrt{\rho^2_{XT}\ \rho^2_{YT}}\,, \qquad [2.2.2.8]$$

wobei sich im einzelnen die in Tabelle 2.2.2.1 dargestellten Eigenschaften der Paralleltest-Methode ergeben:

	$T_{ot} \equiv T_{ot}$	$T_{ot} = T_{ot'} + c_{tt'}$	$T_{ot} \neq T_{ot'} + c_{tt'}$
$\rho^2_{XT} = \rho^2_{YT}$	$\rho(X_{ot}, Y_{ot'}) = \rho^2_{XT}$		$\rho(X_{ot}, Y_{ot'}) < \rho^2_{XT}$
$\rho^2_{YT} \neq \rho^2_{YT}$	$\rho(X_{ot}, Y_{ot'}) = \sqrt{\rho^2_{XT}\ \rho^2_{YT}}$		$\rho(X_{ot}, Y_{ot'}) < \sqrt{\rho^2_{XT}\ \rho^2_{YT}}$

Tab. 2.2.2.1: Eigenschaften der Paralleltest-Methode

Varianten der Paralleltest-Methode

Bezüglich der Anwendung der Paralleltest-Methode lassen sich zwei Varianten unterscheiden:

- Bei der *Retest-Reliabilität* wird derselbe Test den selben Probanden zweimal vorgegeben. Zwischen Test und Testwiederholung liegt in der Regel ein größerer Zeitraum. Dies birgt die Gefahr in sich, dass sich das gemessene Verhaltensmerkmal in der Zwischenzeit verändert haben kann, so dass die (essentielle) τ-Äquivalenz der beiden Testungen nicht gegeben ist.
  Die Retest-Reliabilität ist daher nur auf Tests anwendbar, die relativ stabile Verhaltensmerkmale – wie z. B. die Intelligenz – messen. Bei der Messung von instabilen Verhaltensmerkmalen dagegen – wie z. B. der Depressivität, die periodischen Schwankungen unterliegt –, führt die Anwendung der Retest-Reliabilität zu einer Unterschätzung der Reliabilität.
- Bei der *Paralleltest-Reliabilität* wird dieses Problem umgangen, indem man das selbe Verhaltensmerkmal zeitgleich mit zwei verschiedenen Tests misst. Da es aber sehr schwierig ist, zwei Tests zu konstruieren, die nicht nur das selbe Verhaltensmerkmal messen, sondern die auch die selbe Messgenauigkeit aufweisen, wird man auch damit die Reliabilität des Tests in der Regel nicht exakt bestimmen können.

Herleitung der Gleichung [2.2.2.3]

Zunächst zeigen wir, dass der Testscore nicht mit dem Messfehler des Kriteriums korreliert ist, so dass die Kovarianz

$$\sigma(X_{ot}, F_{ot'}) = 0. \qquad [2.2.2.9]$$

Laut Axiom 1 ist $E(F_{ot'}) = 0$ und

$$\sigma(X_{ot}, F_{ot'}) = E(X_{ot}F_{ot'}) - E(X_{ot})E(F_{ot'})$$

$$= E(X_{ot}F_{ot'})$$

$$= E((T_{ot} + F_{ot})F_{ot'})$$

$$= E(T_{ot}F_{ot'}) + E(F_{ot}F_{ot'})$$

$$= E(T_{ot}F_{ot'}) - E(T_{ot})E(F_{ot'}) + E(F_{ot}F_{ot'}) - E(F_{ot})E(F_{ot'})$$

$$= \sigma(T_{ot}, F_{ot'}) + \sigma(F_{ot}, F_{ot'}).$$

Aufgrund von Axiom 3 ist aber $\sigma(F_{ot}, F_{ot'}) = 0$ und wegen Axiom 4 ist $\sigma(T_{ot}, F_{ot'}) = 0$, so dass die Kovarianz $\sigma(X_{ot}, F_{ot'})$ verschwindet.

Als nächstes zeigen wir, dass die Validität in zwei multiplikativ verknüpfte Komponenten

$$\rho(X_{ot}, Y_{ot'}) = \rho(X_{ot}, T_{ot'})\,\rho(Y_{ot'}, T_{ot'}) \qquad [2.2.2.10]$$

aufgespalten werden kann, von denen die eine die Korrelation zwischen Testscore (X_{ot}) und True-Score des Kriteriums ($T_{ot'}$) und die andere die Quadratwurzel aus der Reliabilität der Kriteriumsmessung darstellt:

$$\rho(X_{ot}, Y_{ot'}) = \frac{\sigma(X_{ot}, Y_{ot'})}{\sigma(X_{ot})\,\sigma(Y_{ot'})}$$

$$= \frac{E(X_{ot}Y_{ot'}) - E(X_{ot})E(Y_{ot'})}{\sigma(X_{ot})\,\sigma(Y_{ot'})}$$

$$= \frac{E(X_{ot}(T_{ot'} + F_{ot'})) - E(X_{ot})E(T_{ot'} + F_{ot'})}{\sigma(X_{ot})\,\sigma(Y_{ot'})}$$

$$= \frac{E(X_{ot}T_{ot'}) + E(X_{ot}F_{ot'}) - E(X_{ot})E(T_{ot'}) - E(X_{ot})E(F_{ot'})}{\sigma(X_{ot})\,\sigma(Y_{ot'})}$$

$$= \frac{\sigma(X_{ot}, T_{ot'}) + \sigma(X_{ot}, F_{ot'})}{\sigma(X_{ot})\,\sigma(Y_{ot'})}.$$

Da aber $\sigma(X_{ot}, F_{ot'}) = 0$ (vgl. Gleichung [2.2.2.9]) erhalten wir:

$$\rho(X_{ot}, Y_{ot'}) = \frac{\sigma(X_{ot}, T_{ot'})}{\sigma(X_{ot})\,\sigma(Y_{ot'})}$$

$$= \frac{\sigma(X_{ot}, T_{ot'})}{\sigma(X_{ot})\,\sigma(T_{ot'})}\,\frac{\sigma^2(T_{ot'})}{\sigma(Y_{ot'})\,\sigma(T_{ot'})}.$$

Und da True-Score und Messfehler laut Axiom 2 nicht miteinander korrelieren, ist $\sigma(Y_{ot'}, T_{ot'}) = \sigma^2(T_{ot'})$, und wir können auch schreiben:

$$\rho(X_{ot}, Y_{ot'}) = \frac{\sigma(X_{ot}, T_{ot'})}{\sigma(X_{ot})\,\sigma(T_{ot'})}\,\frac{\sigma(Y_{ot'}, T_{ot'})}{\sigma(Y_{ot'})\,\sigma(T_{ot'})}$$

$$= \rho(X_{ot}, T_{ot'})\,\rho(Y_{ot'}, T_{ot'}).$$

Auf analogem Wege zeigen wir, dass auch $\rho(X_{ot}, T_{ot'})$ in zwei multiplikativ verknüpfte Komponenten

$$\rho(X_{ot}, T_{ot'}) = \rho(X_{ot}, T_{ot})\,\rho(T_{ot}, T_{ot'}) \qquad [2.2.2.11]$$

zerlegt werden kann, von denen die eine die Quadratwurzel aus der Reliabilität des Tests und die andere die Korrelation zwischen den True-Scores von Test und Kriterium ist:

$$\rho(X_{ot}, T_{ot'}) = \frac{\sigma(X_{ot}, T_{ot'})}{\sigma(X_{ot})\,\sigma(T_{ot'})}$$

$$= \frac{E(X_{ot}T_{ot'}) - E(X_{ot})E(T_{ot'})}{\sigma(X_{ot})\,\sigma(T_{ot'})}$$

$$= \frac{E((T_{ot} + F_{ot})T_{ot'}) - E(T_{ot} + F_{ot})E(T_{ot'})}{\sigma(X_{ot})\,\sigma(T_{ot'})}$$

$$= \frac{E(T_{ot}T_{ot'}) + E(F_{ot}T_{ot'}) - E(T_{ot})E(T_{ot'}) - E(F_{ot})E(T_{ot'})}{\sigma(X_{ot})\,\sigma(T_{ot'})}$$

$$= \frac{\sigma(T_{ot}, T_{ot'}) + \sigma(F_{ot}, T_{ot'})}{\sigma(X_{ot})\,\sigma(T_{ot'})} \quad .$$

Da aber laut Axiom 4 der Messfehler des Kriteriums nicht mit dem True-Score des Tests korreliert, folgt daraus

$$\rho(X_{ot}, T_{ot'}) = \frac{\sigma(T_{ot}, T_{ot'})}{\sigma(X_{ot})\,\sigma(T_{ot'})}$$

$$= \frac{\sigma^2(T_{ot})}{\sigma(X_{ot})\,\sigma(T_{ot})}\,\frac{\sigma(T_{ot}, T_{ot'})}{\sigma(T_{ot})\,\sigma(T_{ot'})} .$$

Da True-Score und Messfehler laut Axiom 2 nicht miteinander korrelieren, ist $\sigma(X_{ot}, T_{ot}) = \sigma^2(T_{ot})$, und wir können auch schreiben:

$$\rho(X_{ot}, T_{ot'}) = \frac{\sigma(X_{ot}, T_{ot})}{\sigma(X_{ot})\,\sigma(T_{ot})}\,\frac{\sigma(T_{ot}, T_{ot'})}{\sigma(T_{ot})\,\sigma(T_{ot'})}$$

$$= \rho(X_{ot}, T_{ot})\,\rho(T_{ot}, T_{ot'}) \quad .$$

Einsetzen in Gleichung [2.2.2.10] ergibt schließlich:

$$\rho(X_{ot}, Y_{ot'}) = \rho(X_{ot}, T_{ot})\,\rho(T_{ot}, T_{ot'})\,\rho(Y_{ot'}, T_{ot'}) \quad .$$

Die größtmögliche Validität eines Tests

Als weitere Schlussfolgerung aus Gleichung [2.2.2.3] ergibt sich, dass die Validität eines Tests nie größer sein kann als die Quadratwurzel seiner Reliabilität

$$\rho_{XY} \le \sqrt{\rho^2_{XT}} \quad . \qquad [2.2.2.12]$$

Beweis

Da Korrelationen keine Werte größer Eins annehmen können, folgt aus Gleichung [2.2.2.3]:

$$\rho(X_{ot}, Y_{ot'}) = \rho(X_{ot}, T_{ot})\,\rho(T_{ot}, T_{ot'})\,\rho(Y_{ot'}, T_{ot'})$$

$$\leq \rho(X_{ot}, T_{ot}) = \sqrt{\rho^2_{XT}}$$

Da es sich bei der Reliabilität um eine Zahl zwischen Null und Eins handelt, ist die Quadratwurzel aus der Reliabilität allerdings größer als die Reliabilität selbst

$$\sqrt{\rho^2_{XT}} > \rho^2_{XT} \;,$$

so dass die Validität eines Tests seine Reliabilität durchaus übersteigen kann (vgl. Abb. 2.2.2.1). Weil die Validität aber nicht größer werden kann als die Quadratwurzel aus der Reliabilität ist dies nur in beschränktem Maße möglich: Die Reliabilität eines Tests ist eine notwendige Voraussetzung für seine Validität. Nur reliable Tests können auch valide sein.

Abb. 2.2.2.1: Die größtmögliche Validität eines Tests als Funktion seiner Reliabilität

Zur Interpretation des Reliabilitätskoeffizienten

Grundlegend für die Interpretation des Reliabilitätskoeffizienten ist die in Gleichung [2.2.2.13] dargestellte Beziehung

$$\rho^2_{XT} = \frac{\sigma^2(T_{ot})}{\sigma^2(X_{ot})} = \frac{\sigma^2(T_{ot})}{\sigma^2(T_{ot}) + \sigma^2(F_{ot})} \;, \qquad [2.2.2.13]$$

wonach die Reliabilität den systematischen Varianzanteil der Testergebnisse beschreibt.

Dies ist ein durchaus plausibles Maß für die Messgenauigkeit eines Tests. Bei genauerer Betrachtung der Gleichung [2.2.2.13] zeigt sich jedoch, dass die Reliabilität nicht die Messgenauigkeit des Tests schlechthin beschreibt, sondern lediglich seine Messgenauigkeit *relativ zu einer gegebenen Referenzpopulation*:

- Je homogener die Referenzpopulation ist, d. h. je weniger sich die Probanden in ihren True-Scores unterscheiden, desto kleiner ist die True-Score-Varianz und desto geringer fällt die Reliabilität aus.
- Im Extremfall, wenn die Probanden alle denselben True-Score haben, nimmt die Reliabilität den Wert Null an – und zwar unabhängig davon, wie genau der Test misst und wie klein die Fehlerstreuung auch sein mag.

Herleitung von Gleichung [2.2.2.13]

Für die Quadratwurzel aus der Reliabilität gilt die Gleichung

$$\rho_{XT} = \frac{\sigma(X_{ot}, T_{ot})}{\sigma(X_{ot})\sigma(T_{ot})} \qquad [2.2.2.14]$$

worin

$$\sigma(X_{ot}, T_{ot}) = E(X_{ot}T_{ot}) - E(X_{ot})E(T_{ot}) .$$

Aufgrund der linearen Grundgleichung $X_{ot} = T_{ot} + F_{ot}$ und des 1. Axioms von Gulliksen, wonach $E(F_{ot}) = 0$ ist aber

$$E(X_{ot}) = E(T_{ot}) + E(F_{ot}) = E(T_{ot}) ,$$

so dass wir für die Kovarianz der Testergebnisse mit ihrem True-Score die folgende Gleichung erhalten:

$$\sigma(X_{ot}, T_{ot}) = E((T_{ot} + F_{ot})T_{ot}) - E(T_{ot})^2$$

$$= E(T_{ot}^2) + E(F_{ot}T_{ot}) - E(T_{ot})^2 .$$

Wegen Axiom 1 können wir diese Gleichung auch in der Form

$$\sigma(X_{ot}, T_{ot}) = E(T_{ot}^2) - E(T_{ot})^2 + E(F_{ot}T_{ot}) - E(F_{ot})E(T_{ot})$$

$$= \sigma^2(T_{ot}) + \sigma(F_{ot}, T_{ot})$$

schreiben. Wegen Axiom 2 ist aber $\sigma(F_{ot}, T_{ot}) = 0$, so dass wir die Gleichung

$$\sigma(X_{ot}, T_{ot}) = \sigma^2(T_{ot}) \qquad [2.2.2.15]$$

erhalten. Setzen wir dies in Gleichung [2.2.2.14] ein, so ergibt sich

$$\rho_{XT} = \frac{\sigma^2(T_{ot})}{\sigma(X_{ot})\sigma(T_{ot})} = \frac{\sigma(T_{ot})}{\sigma(X_{ot})}$$

und entsprechend

$$\rho_{XT}^2 = \frac{\sigma^2(T_{ot})}{\sigma^2(X_{ot})} \ .$$

Wegen $X_{ot} = T_{ot} + F_{ot}$ und aufgrund von Axiom 2 folgt schließlich für den Nenner

$$\sigma^2(X_{ot}) = \sigma^2(T_{ot}) + \sigma^2(F_{ot}) \ .$$

Der Standardmessfehler

Wünschenswert wäre daher ein Maß für die Fehlerstreuung des Tests, welches von der Homogenität der Referenzpopulation unabhängig ist.

Ein solches Maß ist der als Standardabweichung der Fehlervariable definierte *Standardmessfehler*, der sich nach einfacher Umformung der Gleichung [2.2.2.13] berechnen lässt, wenn die Reliabilität des Tests und die Maßzahlvarianz in der Referenzpopulation bekannt sind

$$\text{Standardmeßfehler} = \sigma(F_{ot}) = \sigma(X_{ot})\sqrt{1 - \rho_{XT}^2} \ . \qquad [2.2.2.16]$$

Auch der Standardmessfehler ist aber kein populationsunabhängiges Maß für die Messgenauigkeit eines Tests.

- Er beschreibt nicht, wie genau der True-Score einer bestimmten Vp gemessen werden kann,
- sondern nur die *durchschnittliche* Fehlerstreuung in der Referenzpopulation.

Hinzu kommt, dass man zur Bewertung des Standardmessfehlers einen Vergleichsmaßstab braucht.

- Ob z. B. ein Standardmessfehler von $\sigma(F_{ot}) = 1.5$ als geringfügige oder als erhebliche Ungenauigkeit der Testergebnisse zu bewerten ist, wird davon abhängen, wie breit die Testergebnisse streuen.
- Sind die Testergebnisse breit gestreut, z. B. mit $\sigma(X_{ot}) = 15$, so ist eine Fehlerstreuung von $\sigma(F_{ot}) = 1.5$ offensichtlich ge-

ringfügiger, als wenn schon die Testergebnisse selbst nur mit $\sigma(X_{ot}) = 3$ streuen.

Vergleichen wir jedoch die Fehlerstreuung mit der Streuung der Testscores, dann gelangen wir wieder zur Reliabilität:

$$\frac{\sigma(F_{ot})}{\sigma(X_{ot})} = \sqrt{1 - \rho_{XT}^2}\,.$$

Die Populationsabhängigkeit der Beurteilung der Messgenauigkeit von Tests ist im Rahmen der klassischen Testtheorie nicht behebbar.

Herleitung des Standardmessfehlers

Wegen $\sigma^2(X_{ot}) = \sigma^2(T_{ot}) + \sigma^2(F_{ot})$ kann Gleichung [2.2.2.13] auch in der Form

$$\rho_{XT}^2 = \frac{\sigma^2(T_{ot})}{\sigma^2(X_{ot})} = \frac{\sigma^2(X_{ot}) - \sigma^2(F_{ot})}{\sigma^2(X_{ot})}$$

dargestellt werden. Multiplikation der Gleichung mit $\sigma^2(X_{ot})$ ergibt

$$\sigma^2(X_{ot})\rho_{XT}^2 = \sigma^2(X_{ot}) - \sigma^2(F_{ot})$$

und daher

$$\sigma^2(F_{ot}) = \sigma^2(X_{ot}) - \sigma^2(X_{ot})\rho_{XT}^2 = \sigma^2(X_{ot})\left(1 - \rho_{XT}^2\right).$$

Diese Gleichung ist der Rechenformel für den Standardmessfehler äquivalent.

Eine populationsunabhängige Beurteilung der Messgenauigkeit einer Testung wird erst dann möglich, wenn man den testtheoretischen Kalkül um zusätzliche Postulate erweitert. Zumindest aber bei langen Tests, die aus einer Vielzahl von Testitems bestehen, liefert der Standardmessfehler im mittleren Scorebereich eine durchaus brauchbare Näherungsformel.

Zusammengesetzte Tests

In der Praxis wird man möglichst reliable Tests anstreben. Dies wirft die Frage auf, wie man die Reliabilität eines Tests verbessern kann. Wenn wir uns vergegenwärtigen, dass Tests standardisierte Verhaltensstichproben sind, so liegt es nahe, die Reliabilität durch Vergrößerung der Stichprobe zu verbessern.

Angenommen, wir verfügen über zwei essentiell parallele Tests $j = 1$ und $j = 2$ mit den Scorevariablen X_{o1} und X_{o2}, die dasselbe Verhaltensmerkmal $T_{o1} = T_{o2} + c_{12}$ gleich reliabel messen, so dass

$$\rho^2(X_{o1}, T_{o1}) = \rho^2(X_{o2}, T_{o2}) = \rho^2_{XT_o},$$

dann könnten wir daraus einen (zusammengesetzten) Test t bilden, der die Items aus beiden (Teil-)Tests enthält, so dass sich der Umfang unserer Verhaltensstichprobe verdoppelt und $X_{ot} = X_{o1} + X_{o2}$.

Reliabilität bei doppelter Testlänge

Für die Reliabilität des zusammengesetzten Tests gilt dann die bereits auf Spearman (1907) zurückgehende Gleichung

$$\rho^2_{XT} = \frac{2\rho^2_{XT_o}}{1 + \rho^2_{XT_o}}, \qquad [2.2.2.17]$$

die in der Literatur als Spearman-Brown-Formel bekannt ist. Sie beschreibt die Zunahme der Reliabilität eines Tests, wenn man ihn durch Hinzufügen eines essentiell parallelen Tests auf die doppelte Testlänge erweitert.

Abb. 2.2.2.2: Zunahme der Reliabilität bei Verdoppelung der Testlänge

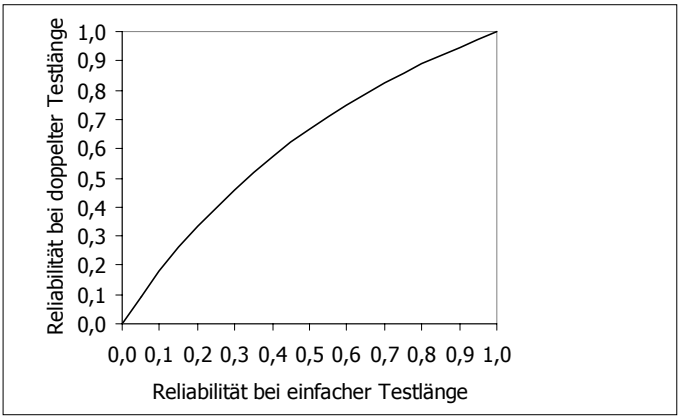


Abb. 2.2.2.2 zeigt, dass die Reliabilität eines Tests dabei umso stärker zunimmt, je kleiner die ursprüngliche Reliabilität war. Ist die Reliabilität schon bei einfacher Testlänge relativ hoch, so kann sie durch Verdoppelung der Testlänge nur relativ geringfügig verbessert werden.

Da die Spearman-Brown-Formel nur für (zumindest essentiell) parallele Teiltests gilt, wird die Reliabilität des zusammengesetzten Tests den prognostizierten Wert allerdings nur dann erreichen, wenn die Teiltests diese Eigenschaft auch tatsächlich aufweisen.

Herleitung der Spearman-Brown-Formel

Für die Reliabilität des zusammengesetzten Tests gilt wegen Gleichung [2.2.2.13] die Beziehung

$$\rho_{XT}^2 = \frac{\sigma^2(T_{ot})}{\sigma^2(X_{ot})} = \frac{\sigma^2(T_{o1} + T_{o2})}{\sigma^2(X_{o1} + X_{o2})}. \qquad [2.2.2.18]$$

Darin ist für beliebige Teiltests

$$\sigma^2(T_{o1} + T_{o2}) = \sigma^2(T_{o1}) + \sigma^2(T_{o2}) + 2\sigma(T_{o1}, T_{o2}) \qquad [2.2.2.19]$$

und

$$\sigma^2(X_{01} + X_{o2}) = \sigma^2(X_{o1}) + \sigma^2(X_{o2}) + 2\sigma(X_{o1}, X_{o2}) \qquad [2.2.2.20]$$

$$= \sigma^2(X_{o1}) + \sigma^2(X_{o2}) + 2\rho(X_{o1}, X_{o2})\sigma(X_{o1})\sigma(X_{o2}).$$

Sind die Teiltests (zumindest essentiell) parallel, so ist $\sigma^2(T_{o1}) = \sigma^2(T_{o2}) = \sigma^2(T_{oo})$ und $\sigma(T_{o1}, T_{o2}) = \sigma^2(T_{oo})$, und Gleichung [2.2.2.19] nimmt die Form

$$\sigma^2(T_{o1} + T_{o2}) = 4\sigma^2(T_{oo}) \qquad [2.2.2.21]$$

an. In Gleichung [2.2.2.20] ist für (zumindest essentiell) parallele Teiltests $\sigma^2(X_{o1}) = \sigma^2(X_{o2}) = \sigma^2(X_{oo})$ und die Korrelation der beiden Teiltests ist gleich ihrer Reliabilität, so dass Gleichung [2.2.2.20] die Form

$$\sigma^2(X_{o1} + X_{o2}) = 2\sigma^2(X_{oo}) + 2\rho_{XT_o}^2\sigma^2(X_{oo}) \qquad [2.2.2.22]$$

$$= 2\sigma^2(X_{oo})\left(1 + \rho_{XT_o}^2\right)$$

annimmt. Einsetzen der Gleichungen [2.2.2.21] und [2.2.2.22] in Gleichung [2.2.2.18] ergibt schließlich

$$\rho_{XT}^2 = \frac{\sigma^2(T_{o1} + T_{o2})}{\sigma^2(X_{o1} + X_{o2})}$$

$$= \frac{4\sigma^2(T_{oo})}{2\sigma^2(X_{oo})\left(1 + \rho_{XT_o}^2\right)}$$

$$= \frac{2\dfrac{\sigma^2(T_{oo})}{\sigma^2(X_{oo})}}{1 + \rho_{XT_o}^2} = \frac{2\rho_{XT_o}^2}{1 + \rho_{XT_o}^2}.$$

Reliabilität bei n-facher Testlänge

Wie sich zeigen lässt, kann die Spearman-Brown-Formel für jede beliebige Testlänge verallgemeinert werden und nimmt für die Reliabilität eines aus n (zumindest essentiell) parallelen Testtei-

len j = 1,...,n zusammengesetzten Test mit der Scorevariable

$$X_{ot} = \sum_{j=1}^{n} X_{oj}$$

die Form

$$\rho^2_{XT} = \frac{n\,\rho^2_{XT_o}}{1 + (n-1)\,\rho^2_{XT_o}} \qquad [2.2.2.23]$$

an.[1] Abb. 2.2.2.3 zeigt, dass die Reliabilität mit zunehmender Testlänge umso rascher wächst, je geringer die Ausgangsreliabilität war.

Die verallgemeinerte Spearman-Brown-Formel in Gleichung [2.2.2.23] ist dabei nicht nur für ganzzahlige n anwendbar, so dass sich damit auch berechnen lässt,

- mit welcher Reliabilitätsverbesserung zu rechnen ist, wenn man z. B. einen Test aus k=20 Items auf 30 Items (= 1.5-fache Testlänge) verlängert,
- oder mit welcher Verschlechterung der Reliabilität zu rechnen ist, wenn man eine Kurzform des Tests verwendet, die nur 10 Items (= halbe Testlänge) umfasst.

Abb.2.2.2.3: Zunahme der Reliabilität bei n-facher Testlänge

1. Für den Beweis der Gleichung [2.2.2.23] siehe das Standardwerk von Lord & Novick (1968).

Berechnung der erforderlichen Testlänge

Durch Auflösen der Gleichung [2.2.2.23] nach n können wir eine Rechenformel gewinnen, mittels welcher sich berechnen lässt, auf eine wie vielfache Testlänge ein Test mit der Reliabilität $\rho^2_{XT} = r$ verlängert werden müsste, um eine gewünschte Reliabilität $\rho^2_{XT} = R$ zu erzielen:

$$n = \frac{R(1-r)}{(1-R)r}. \qquad [2.2.2.24]$$

Wie Abb. 2.2.2.4 zeigt, wächst die erforderliche Testlänge sehr rasch an, und zwar umso rascher, je geringer die Ausgangsreliabilität r und je größer die gewünschte Reliabilität R ist.

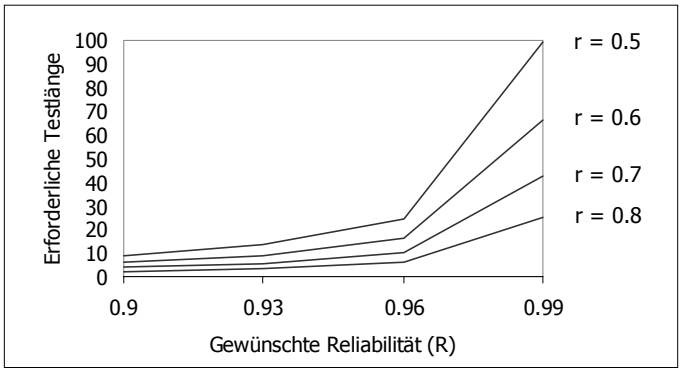


Abb.2.2.2.4: Erforderliche Testlänge für verschiedene Werte von r und R

Bei einer Ausgangsreliabilität von r = 0.5 muss der Test bereits auf 9-fache Testlänge erweitert werden, um eine Reliabilität von R = 0.9 zu erzielen und für eine Reliabilität von R = 0.99 benötigt man gar die 99-fache Testlänge. Selbst bei einer relativ guten Ausgangsreliabilität von 0.8 ist mehr als die zweifache Testlänge erforderlich, um R = 0.9 zu erzielen und um eine Reliabilität von R = 0.99 zu erreichen muss der Test fast auf die 25-fache Testlänge erweitert werden.

Herleitung der Gleichung [2.2.2.24]

Einsetzen von $\rho^2_{XT} = R$ und $\rho^2_{XT_o} = r$ in Gleichung [2.2.2.23] ergibt

$$R = \frac{n\,r}{1+(n-1)\,r}$$

$$= \frac{n\,r}{1+nr-r}.$$

Durch Multiplikation der Gleichung mit (1 + nr – r) folgt daraus

$$R + Rnr - Rr = nr$$

$$Rnr - nr = Rr - R$$

$$n(R-1)r = R(r-1)$$

und nach Division der Gleichung durch den Ausdruck (R – 1) r erhalten wir

$$n = \frac{R(r-1)}{(R-1)r}\ .$$

Die Multiplikation von Zähler und Nenner mit (-1) ergibt schließlich

$$n = \frac{R(1-r)}{(1-R)r}\ .$$

Testteilungs-Methoden der Reliabilitäts-bestimmung

Bereits oben haben wir mit der *Paralleltest-Methode* zwei Verfahrensweisen kennen gelernt, wie man die Reliabilität eines Tests bestimmen kann, wenn das gemessene Verhaltensmerkmal über die Zeit hin stabil ist, so dass man den Vpn den selben Test zweimal vorgeben kann, ohne dass sich ihr True-Score zwischen Test und Testwiederholung verändert (*Retest-Reliabilität*), oder wenn man über einen Paralleltest verfügt, der dasselbe Verhaltensmerkmal gleich gut misst (*Paralleltest-Reliabilität*).

Mit Hilfe der Spearman-Brown-Formel können wir die Reliabilität eines Tests aber auch dann bestimmen, wenn diese Voraussetzungen nicht gegeben sind, es uns jedoch gelingt, den Test in eine Anzahl von n (zumindest essentiell) parallelen Subtests j = 1,...,n zu zerlegen.

Ist dies der Fall, so besitzen alle diese Subtests die selbe Reliabilität $\rho^2_{XT_j} = \rho^2_{XT_o}$, die man mittels der Paralleltest-Methode bestimmen kann, indem man die paarweisen Korrelationen zwischen den Subtests berechnet, wobei für je zwei verschiedene Subtests j und j'

$$\rho^2_{XT_o} = \rho(X_{oj}, X_{oj'}).$$

Mittels der Spearman-Brown-Formel [2.2.2.23]

$$\rho^2_{XT} = \frac{n\ \rho^2_{XT_o}}{1 + (n-1)\ \rho^2_{XT_o}}$$

kann die Reliabilität des Tests (ρ^2_{XT}) schließlich aus der Reliabilität der Subtests hochgerechnet werden.

Split-Half-Reliabilität

Den Spezialfall mit n = 2 bezeichnet man auch als *Split-Half-Reliabilität*, wobei man als Schätzung für die Reliabilität der beiden Subtests deren empirische Korrelation einsetzt

$$\rho^2_{XT_o} \cong r(X_{o1}, X_{o2}) \ . \qquad [2.2.2.25]$$

Split-n-Reliabilität

Für n > 2 Subtests gibt es n(n-1)/2 paarweise Korrelationen zwischen den Subtests. Als Schätzung für die Subtest-Reliabilität verwendet man in diesem Falle daher das geometrische Mittel dieser Korrelationen

$$\rho^2_{XT_o} \cong \sqrt[\frac{n(n-1)}{2}]{\prod_{j=1}^{n-1} \prod_{j'=j+1}^{n} r(X_{oj}, X_{oj'})} \ . \qquad [2.2.2.26]$$

Der Koeffizient Alpha

Ebenfalls auf einer Zerlegung des Tests in Subtests beruht der bereits auf Guttman (1945) zurückgehende und von Cronbach (1951) so benannte *Koeffizient Alpha*, der eine untere Schranke für die Reliabilität angibt, so dass

$$\rho^2_{XT} \geq \alpha . \qquad [2.2.2.27]$$

Die Berechnung des Koeffizienten Alpha, der für n = 2 durch

$$\alpha_2 = 2\left(1 - \frac{\sigma^2(X_{o1}) + \sigma^2(X_{o2})}{\sigma^2(X_{ot})}\right) \qquad [2.2.2.28]$$

definiert ist, wird in der Literatur auch als interne Konsistenzanalyse bezeichnet.

Interne Konsistenzanalyse

Gegenüber der Testteilungsmethode weist die interne Konsistenzanalyse den Vorteil auf, dass ihre Anwendung nicht an die Voraussetzung der essentiellen Parallelität der Subtests gebunden ist.

- Das Gleichheitszeichen in [2.2.2.27] gilt bereits für essentiell τ-äquivalente Subtests,
- und selbst wenn diese Voraussetzung nicht erfüllt ist, kann die Reliabilität des Tests keinesfalls kleiner sein als der Koeffizient Alpha.

Im allgemeinen Fall für n ≥ 2 Subtests hat der Koeffizient Alpha die Form

$$\alpha_n = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} \sigma^2(X_{oj})}{\sigma^2(X_{ot})}\right). \qquad [2.2.2.29]$$

Die Kuder-Richardson-Formeln

Wird der Test in seine einzelnen Items zerlegt, so nimmt die Gleichung [2.2.2.29] für dichotome Testitems die Form

$$\alpha_k = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} p_{oi}(1-p_{oi})}{\sigma^2(X_{ot})}\right). \qquad [2.2.2.30]$$

an, die erstmals von Kuder & Richardson (1937) hergeleitet wurde und in der Literatur unter dem Namen KR-20 bekannt ist.

Sind die Testitems alle gleich schwierig, so besitzen sie dieselbe Lösungswahrscheinlichkeit $p_{oi} = p_{oo}$ und die Gleichung [2.2.2.30] vereinfacht sich zu der unter dem Namen KR-21 bekannten Rechenformel

$$\alpha_k = \frac{k}{k-1}\left(1 - \frac{k\, p_{oo}(1-p_{oo})}{\sigma^2(X_{ot})}\right). \qquad [2.2.2.31]$$

Von deren Anwendung ist jedoch abzuraten, da sie eine geringfügige Rechenvereinfachung um den Preis einer Zusatzannahme erkauft, bei deren Verletzung der Ausdruck auf der rechten Seite von [2.2.2.31] den Koeffizienten Alpha nicht mehr korrekt wiedergibt.

Herleitung des Koeffizienten Alpha[1]

Wegen

$$(\sigma(T_{o1}) - \sigma(T_{o2}))^2 \geq 0$$

gilt

$$\sigma^2(T_{o1}) + \sigma^2(T_{o2}) \geq 2\sigma(T_{o1})\sigma(T_{o2}).$$

Da aber

1. Aus Gründen der mathematischen Einfachheit leiten wir den Koeffizienten Alpha hier nur für n = 2 her. Für den allgemeinen Fall siehe Lord & Novick (1968).

$$\left|\rho(T_{o1}, T_{o2})\right| = \frac{\left|\sigma(T_{o1}, T_{o2})\right|}{\sigma(T_{o1})\sigma(T_{o2})} \leq 1$$

und daher

$$\left|\sigma(T_{o1}, T_{o2})\right| \leq \sigma(T_{o1})\sigma(T_{o2})$$

folgt

$$\sigma^2(T_{o1}) + \sigma^2(T_{o2}) \geq 2\left|\sigma(T_{o1}, T_{o2})\right|.$$

Wegen

$$\sigma^2(T_{ot}) = \sigma^2(T_{o1} + T_{o2}) = \sigma^2(T_{o1}) + \sigma^2(T_{o2}) + 2\sigma(T_{o1}, T_{o2})$$

folgt weiters, dass

$$\sigma^2(T_{ot}) \geq 4\sigma(T_{o1}, T_{o2})$$

und daher

$$\rho_{XT}^2 = \frac{\sigma^2(T_{ot})}{\sigma^2(X_{ot})} \geq \frac{4\sigma(T_{o1}, T_{o2})}{\sigma^2(X_{ot})}. \qquad [2.2.2.32]$$

Nun ist aber wegen der Axiome 3 und 4

$$E(X_{o1}X_{o2}) = E((T_{o1} + F_{o1})(T_{o2} + F_{o2}))$$

$$= E(T_{o1}T_{o2}) + E(F_{o1}T_{o2}) + E(T_{o1}F_{o2}) + E(F_{o1}F_{o2})$$

$$= E(T_{o1}T_{o2})$$

und wegen Axiom 1 ist

$$E(X_{o1}) = E(T_{o1} + F_{o1}) = E(T_{o1}).$$

Aus demselben Grund ist $E(X_{o2}) = E(T_{o2})$, so dass

$$\sigma(X_{o1}, X_{o2}) = E(X_{o1}X_{o2}) - E(X_{o1})E(X_{o2}) \qquad [2.2.2.33]$$

$$= E(T_{o1}T_{o2}) - E(T_{o1})E(T_{o2}) = \sigma(T_{o1}, T_{o2}).$$

Durch Einsetzen der Gleichung [2.2.2.33] in die Ungleichung [2.2.2.32] erhalten wir

$$\rho_{XT}^2 \geq 2\frac{2\sigma(X_{o1}, X_{o2})}{\sigma^2(X_{ot})}$$

und wegen

$$\sigma^2(X_{ot}) = \sigma^2(X_{01} + X_{o2}) = \sigma^2(X_{o1}) + \sigma^2(X_{o2}) + 2\sigma(X_{o1}, X_{o2})$$

folgt

$$\rho_{XT}^2 \geq 2\frac{\sigma^2(X_{ot}) - \sigma^2(X_{o1}) - \sigma^2(X_{o2})}{\sigma^2(X_{ot})}$$

und daher

$$\rho_{XT}^2 \geq 2\left(1 - \frac{\sigma^2(X_{o1}) + \sigma^2(X_{o2})}{\sigma^2(X_{ot})}\right) \quad .$$

Parallele Subtests

Wird der Test in (zumindest essentiell) parallele Subtests zerlegt, dann sind der Koeffizient Alpha und die Spearman-Brown-Formel äquivalent, und beide Berechnungsweisen geben die Reliabilität des Tests korrekt wieder.

Beweis

Seien j = 1 und j = 2 essentiell parallele Subtests, dann ist $\sigma^2(X_{o1}) = \sigma^2(X_{o2}) = \sigma^2(X_{oo})$ und $\rho(X_{o1}, X_{o2}) = \rho_{XT_o}^2$. Für die Kovarianz der beiden Subtests erhalten wir folglich

$$\sigma(X_{o1}, X_{o2}) = \rho(X_{o1}, X_{o2})\sigma(X_{o1})\sigma(X_{o2})$$

$$= \rho_{XT_o}^2 \sigma^2(X_{oo})$$

und für die Varianz des Gesamtscores folgt

$$\sigma^2(X_{ot}) = \sigma^2(X_{o1}) + \sigma^2(X_{o2}) + 2\sigma(X_{o1}, X_{o2})$$

$$= 2\sigma^2(X_{oo}) + 2\rho_{XT_o}^2 \sigma^2(X_{oo})$$

$$= 2\sigma^2(X_{oo})(1 + \rho_{XT_o}^2).$$

Setzen wir dies in den Koeffizienten Alpha ein, so ergibt sich

$$\alpha = 2\left(1 - \frac{\sigma^2(X_{o1}) + \sigma^2(X_{o2})}{\sigma^2(X_{ot})}\right)$$

$$= 2\left(1 - \frac{2\sigma^2(X_{oo})}{2\sigma^2(X_{oo})(1 + \rho_{XT_o}^2)}\right)$$

$$= 2\left(1 - \frac{1}{1 + \rho_{XT_o}^2}\right)$$

$$= 2\left(\frac{1 + \rho_{XT_o}^2 - 1}{1 + \rho_{XT_o}^2}\right)$$

$$= \frac{2\,\rho^2_{XT_o}}{1+\rho^2_{XT_o}}\ .$$

Dies ist aber nichts anderes als die Spearman-Brown-Formel, die für essentiell parallele Subtests gleich der Reliabilität des zusammengesetzten Tests ist.

Anwendungsvoraussetzungen des Koeffizienten Alpha

Anders als bei den auf der Spearman-Brown-Formel beruhenden Testteilungs-Methoden ist die Anwendbarkeit des Koeffizienten Alpha jedoch nicht an die Zerlegung des Tests in essentiell parallele Subtests gebunden. Der Koeffizient Alpha gibt die Reliabilität des Tests bereits dann korrekt wieder, wenn die Subtests zwar essentiell t-äquivalent sind, sich jedoch in ihrer Messgenauigkeit unterscheiden.

Beweis[1]

Seien j = 1 und j = 2 essentiell τ-äquivalente Subtests mit $T_{o1} = T_{oo}$ und $T_{o2} = T_{oo} + c_{12}$, so dass

$$\sigma^2(X_{oj}) = \sigma^2(T_{oo}) + \sigma^2(F_{oj})\ , \qquad [2.2.2.34]$$

dann gelten für die True-Score-Variable des Gesamttests

$$T_{ot} = T_{o1} + T_{o2} = 2T_{oo} + c_{12}$$

und die zugehörige Maßzahlvariable

$$X_{ot} = X_{o1} + X_{o2} = 2T_{oo} + F_{o1} + F_{o2} + c_{12}$$

die Gleichungen

$$\sigma^2(T_{ot}) = 4\sigma^2(T_{oo}) \qquad [2.2.2.35]$$

und

$$\sigma^2(X_{ot}) = 4\sigma^2(T_{oo}) + \sigma^2(F_{o1}) + \sigma^2(F_{o2})\ , \qquad [2.2.2.36]$$

so dass wir durch Einsetzen von [2.2.2.35] und [2.2.2.36] in Gleichung [2.2.2.13] für die Reliabilität des Gesamttests den Ausdruck

$$\rho^2_{XT} = \frac{4\sigma^2(T_{oo})}{4\sigma^2(T_{oo}) + \sigma^2(F_{o1}) + \sigma^2(F_{o2})} \qquad [2.2.2.37]$$

erhalten. Einsetzen von [2.2.2.34] und [2.2.2.36] in den Koeffizienten Alpha (Gleichung [2.2.2.28]) ergibt

1. Aus Gründen der mathematischen Einfachheit führen wir den Beweis hier nur für n = 2 Subtests.

$$\alpha_2 = 2\left(1 - \frac{2\sigma^2(T_{oo}) + \sigma^2(F_{o1}) + \sigma^2(F_{o2})}{4\sigma^2(T_{oo}) + \sigma^2(F_{o1}) + \sigma^2(F_{o2})}\right)$$

$$= 2\left(\frac{4\sigma^2(T_{oo}) + \sigma^2(F_{o1}) + \sigma^2(F_{o2})}{4\sigma^2(T_{oo}) + \sigma^2(F_{o1}) + \sigma^2(F_{o2})} - \frac{2\sigma^2(T_{oo}) + \sigma^2(F_{o1}) + \sigma^2(F_{o2})}{4\sigma^2(T_{oo}) + \sigma^2(F_{o1}) + \sigma^2(F_{o2})}\right)$$

$$= \frac{4\sigma^2(T_{oo})}{4\sigma^2(T_{oo}) + \sigma^2(F_{o1}) + \sigma^2(F_{o2})},$$

so dass der Koeffizient Alpha mit der Reliabilität (Gleichung [2.2.2.37]) übereinstimmt.

Anwendungsvoraussetzungen der Testteilungs-Methoden

Die auf der Spearman-Brown-Formel beruhenden Testteilungs-Methoden geben die Reliabilität des Tests dagegen nur dann korrekt wieder, wenn die zur Reliabilitätsbestimmung gebildeten Subtests zumindest essentiell parallel sind. Essentielle τ-Äquivalenz reicht nicht aus. Die Subtests müssen auch dieselbe Messgenauigkeit besitzen.

Beweis

Seien $j = 1$ und $j = 2$ essentiell τ-äquivalente Subtests mit $T_{o1} = T_{oo}$ und $T_{o2} = T_{oo} + c_{12}$, dann gilt für die Korrelation der beiden Subtests wegen Gleichung [2.2.2.3] und [2.2.2.13] der Ausdruck

$$\rho(X_{o1}, X_{o2}) = \rho(X_{o1}, T_{o1})\rho(X_{o2}, T_{o2}) \qquad [2.2.2.38]$$

$$= \sqrt{\rho^2_{XT_1}\,\rho^2_{XT_2}}$$

$$= \sqrt{\frac{\sigma^2(T_{oo})}{\sigma^2(T_{oo}) + \sigma^2(F_{o1})}\,\frac{\sigma^2(T_{oo})}{\sigma^2(T_{oo}) + \sigma^2(F_{o2})}}$$

$$= \frac{\sigma^2(T_{oo})}{\sqrt{(\sigma^2(T_{oo}) + \sigma^2(F_{o1}))(\sigma^2(T_{oo}) + \sigma^2(F_{o2}))}}$$

Einsetzen von [2.2.2.38] für $\rho^2_{XT_o}$ in die Spearman-Brown-Formel [2.2.2.17] ergibt

$$\frac{2\rho(X_{o1}, X_{o2})}{1 + \rho(X_{o1}, X_{o2})} = \qquad [2.2.2.39]$$

$$= \frac{\dfrac{2\sigma^2(T_{oo})}{\sqrt{(\sigma^2(T_{oo}) + \sigma^2(F_{o1}))(\sigma^2(T_{oo}) + \sigma^2(F_{o2}))}}}{1 + \dfrac{\sigma^2(T_{oo})}{\sqrt{(\sigma^2(T_{oo}) + \sigma^2(F_{o1}))(\sigma^2(T_{oo}) + \sigma^2(F_{o2}))}}}$$

$$= \frac{\dfrac{2\sigma^2(T_{oo})}{\sqrt{(\sigma^2(T_{oo})+\sigma^2(F_{o1}))(\sigma^2(T_{oo})+\sigma^2(F_{o2}))}}}{\dfrac{\sigma^2(T_{oo})+\sqrt{(\sigma^2(T_{oo})+\sigma^2(F_{o1}))(\sigma^2(T_{oo})+\sigma^2(F_{o2}))}}{\sqrt{(\sigma^2(T_{oo})+\sigma^2(F_{o1}))(\sigma^2(T_{oo})+\sigma^2(F_{o2}))}}}$$

$$= \frac{2\sigma^2(T_{oo})}{\sigma^2(T_{oo})+\sqrt{(\sigma^2(T_{oo})+\sigma^2(F_{o1}))(\sigma^2(T_{oo})+\sigma^2(F_{o2}))}}$$

$$= \frac{4\sigma^2(T_{oo})}{2\sigma^2(T_{oo})+2\sqrt{(\sigma^2(T_{oo})+\sigma^2(F_{o1}))(\sigma^2(T_{oo})+\sigma^2(F_{o2}))}}$$

Wegen Gleichung [2.2.2.37] ist dieser Ausdruck genau dann gleich der Reliabilität des Gesamttests, wenn die Nenner der Brüche in [2.2.2.37] und [2.2.2.39] übereinstimmen und also

$$2\sqrt{(\sigma^2(T_{oo})+\sigma^2(F_{o1}))(\sigma^2(T_{oo})+\sigma^2(F_{o2}))} = \qquad [2.2.2.40]$$

$$= 2\sigma^2(T_{oo})+\sigma^2(F_{o1})+\sigma^2(F_{o2})$$

Wie man sich durch quadrieren und ausmultiplizieren der linken und der rechten Seite der Gleichung [2.2.2.40] leicht überzeugen kann, ist dies genau dann der Fall, wenn

$$4\sigma^2(F_{o1})\sigma^2(F_{o2}) =$$

$$= \sigma^4(F_{o1})+2\sigma^2(F_{o1})\sigma^2(F_{o2})+\sigma^4(F_{o2})$$

und also

$$\sigma^4(F_{o1})-2\sigma^2(F_{o1})\sigma^2(F_{o2})+\sigma^4(F_{o2}) = 0$$

so dass

$$(\sigma^2(F_{o1})-\sigma^2(F_{o2}))^2 = 0$$

und die beiden Subtests daher dieselbe Fehlervarianz

$$\sigma^2(F_{o1}) = \sigma^2(F_{o2})$$

besitzen müssen.

Überschätzung der Reliabilität durch die Split-Half-Methode

Sind die Subtests zwar essentiell τ-äquivalent, jedoch unterschiedlich messgenau, dann überschätzt die Split-Half-Methode die Reliabilität des Tests.

Beweis

Seien j = 1 und j = 2 essentiell τ-äquivalente Subtests, die sich jedoch in ihrer Messgenauigkeit unterscheiden, so dass

$$\sigma^2(F_{o1}) \neq \sigma^2(F_{o2}) \ ,$$

so ist der Ausdruck im Nenner des Bruches in Gleichung [2.2.2.39] kleiner als der Ausdruck im Nenner von [2.2.2.37] und die Reliabilität des Tests wird durch die Split-Half-Reliabilität überschätzt.

Unterschätzung der Reliabilität durch die Split-Half-Methode

Besitzen die Subtests dagegen die selbe Messgenauigkeit ohne essentiell τ-äquivalent zu sein, so wird die Reliabilität durch die Split-Half Methode unterschätzt.

Beweis

Seien j = 1 und j = 2 zwei Subtests, welche zwar die selbe Messgenauigkeit besitzen, so dass

$$\rho(X_{o1}, T_{o1}) = \rho(X_{o2}, T_{o2}) = \sqrt{\rho^2_{XT_o}} \ ,$$

ohne dass die Subtests jedoch (zumindest essentiell) t-äquivalent sind und daher

$$\rho(T_{o1}, T_{o2}) < 1 \ ,$$

so ist wegen Gleichung [2.2.2.3]

$$\rho(X_{o1}, X_{o2}) < \rho^2_{XT_o} \ ,$$

und Einsetzen in die Spearman-Brown-Formel [2.2.2.23] ergibt einen Wert, der die Reliabilität des Gesamttests unterschätzt:

$$\frac{2\rho(X_{o1}, X_{o2})}{1+\rho(X_{o1}, X_{o2})} < \frac{2\rho^2_{XT_o}}{1+\rho^2_{XT_o}} = \rho^2_{XT} \ .$$

Verletzung der Anwendungsvoraussetzungen

Im Unterschied zum Koeffizienten Alpha, der in jedem Fall eine untere Schranke der Reliabilität bildet, ist dies bei den Testteilungs-Methoden nicht der Fall. Sind ihre Anwendungsvoraussetzungen verletzt, so können die Testteilungs-Methoden sowohl zu einer Unterschätzung als auch zu einer Überschätzung der Reliabilität führen.

Vergleich der verschiedenen Reliabilitätsmaße

Einen Überblick über die verschiedenen Methoden der Reliabilitätsbestimmung und ihrer Anwendungsvoraussetzungen gibt Tabelle 2.2.2.2.

Methode	Version	Voraussetzung	Falls nicht gegeben
Paralleltest-Methode	Retest-Reliabilität	Essentiell parallele Wiederholbarkeit	$< \sqrt{\rho^2_{XT}\rho^2_{YT}}$
	Paralleltest-Reliabilität	Zwei essentiell parallele Testformen	
Testteilungs-Methode	Split-Half (SP-B)	Zwei essentiell parallele Testhälften	Nicht anwendbar
	Split n (SP-B)	n essentiell parallele Subtests	
Interne Konsistenz-analyse	α_2	Zwei essentiell τ-äquivalente Testhälften	$< \rho^2_{XT}$
	α_n	n essentiell τ-äquivalente Subtests	
	α_k (KR-20)	k parallele, binäre Testitems	
	(KR-21)		Nicht anwendbar

Tab. 2.2.2.2: Methoden der Reliabilitätsbestimmung und ihre Anwendungsvorausetzungen

Sind ihre Anwendungsvoraussetzungen erfüllt, so muss jede dieser Methoden zum selben Ergebnis führen, das die Reliabilität des Tests exakt wiedergibt. Weichen die Ergebnisse voneinander ab, so erfahren wir daraus über die Reliabilität des Tests immerhin soviel, dass sie nicht kleiner als das größte Alpha ist:

$$\rho^2_{XT} \geq \max(\alpha_2, \alpha_n, \alpha_k) . \qquad [2.2.2.41]$$

Umgekehrt gibt der Vergleich der verschiedenen Reliabilitätsmaße auch Hinweise darauf, welche der Anwendungsvoraussetzungen empirisch verletzt sind.

Methode	Ergebnis
Paralleltest-Methode	0.58
Split-Half (Spearman-Brown)	0.94
Cronbach-Alpha (α_2)	0.94
Kuder-Richardson (α_k)	0.89

Tab. 2.2.2.3: Reliabilitätskoeffizienten des Subtests Zahlenreihen nach Kempf & Meder (1993)

So ergab z. B. die Reliabilitätsanalyse des Subtests "Zahlenreihen"

der Form A des Intelligenz-Struktur-Tests (IST-70) von Amthauer (1973) die in Tabelle 2.2.2.3 dargestellten Ergebnisse.

Daraus können wir nun darauf schließen,

- dass die Reliabilität des Subtests Zahlenreihen der Form A des IST-70 keinesfalls kleiner ist als 0.94 (= der größere der beiden Werte α_2 und α_k),
- dass die angeblich parallellen Subtests Zahlenreihen der beiden Testformen A und B des IST-70 in Wirklichkeit nicht einmal essentiell parallel sind (da die Paralleltest-Reliabilität kleiner als 0.94 ausfällt),
- dass die Testitems des Subtests Zahlenreihen der Form A des IST-70 nicht parallel sind (wegen $\alpha_k < 0.94$), und
- dass die zur Berechnung von Split-Half und Cronbach-Alpha gebildeten Testhälften möglicherweise (zumindest essentiell) parallel sind (da beide Methoden zum selben Ergebnis führen),
- so dass es keinerlei empirische Anhaltspunkte gibt, die dagegen sprechen, dass die Reliabilität des Subtests Zahlenreihen der Form A des IST-70 (in der untersuchten Referenzpopulation) tatsächlich $\rho^2_{XT} = 0.94$ beträgt.

2.2.3 Stochastische Testmodelle

Hatten wir in der klassischen Testtheorie ganz pragmatisch vorausgesetzt, mittels psychologischer Tests bestimmte Verhaltensmerkmale "messen" zu können, so wollen wir nun den Begriff der Messung selbst problematisieren. Dazu wiederholen wir zunächst einige der Hauptergebnisse, die wir schon in Bd. 1, Kap. 2.4 erarbeitet haben.

Der Begriff der Messung

Unter einer Messung versteht man die Zuordnung von (Maß)Zahlen zu Messobjekten:

$$A \rightarrow x_A.$$

Diese Zuordnung ist nicht Selbstzweck, sondern geschieht so, dass bestimmte empirisch feststellbare Relationen zwischen den Messobjekten auf entsprechende numerische Relationen zwischen den Maßzahlen abgebildet werden, so dass man von den Maßzahlen und den zwischen ihnen bestehenden Relationen jederzeit wieder auf die entsprechenden empirischen Relationen

zwischen den Messobjekten zurückschließen kann (Homomorphie).

Qualitative Variablen

Grundsätzlich kann bei den Messgrößen zwischen qualitativen und quantitativen Variablen unterschieden werden.

Auch qualitative Variablen können gemessen werden, indem man den verschiedenen Ausprägungen der Variable verschiedene Zahlen zuordnet. Man spricht dann von einer Nominalskala (welche die Messobjekte lediglich mit Zahlen benennt). Welche Zahlen man zur Benennung der verschiedenen Qualitäten verwendet, ist völlig gleichgültig. Wichtig ist nur, dass die beobachtete Information über die Messobjekte erhalten bleibt, d. h. aus den Maßzahlen ablesbar ist.

Dichotome und polytome Variablen

Unterscheidet eine Nominalskala nur zwei Skalenwerte, so spricht man von einer dichotomen Variable; unterscheidet sie mehr als zwei Skalenwerte, so spricht man von einer polytomen Variable. In der Praxis verwendet man als Skalenwerte meist die Zahlen 0,1,2,3... Das ist aber nicht zwingend, und auch auf die Reihenfolge kommt es nicht an, da die Nominalskala lediglich die Gleichheit (=°) bzw. Ungleichheit (≠°) der Messobjekte auf die entsprechenden numerischen Relationen (= bzw. ≠) abbildet. Eine dichotome Variable mit den Skalenwerten Null und Eins nennt man auch binär.

Quantitative Variablen

Von einer quantitativen Variable spricht man dann, wenn die Zahlenwerte zumindest eine Rangordnung zwischen den Messobjekten abbilden. Voraussetzung dafür ist, dass wir eine empirische Ordnungsrelation zwischen den Messobjekten (z.B. >°) beobachten können. Letzteres ist jedoch keineswegs trivial, da (zweistellige) empirische Relationen die Eigenschaften der Irreflexivität, Assymetrie und Transitivität erfüllen müssen, damit sie eine Ordnungsrelation sind und auf die entsprechende numerische Ordnungsrelation (z. B. >) abgebildet werden können.

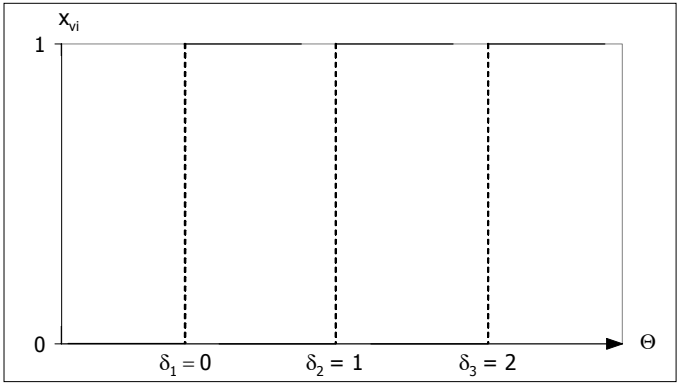
Die Guttman-Skala

Hinreichende Bedingungen dafür, dass die Probanden aufgrund manifester (binärer) Itemantworten auf einer Eigenschaftsdimension (z. B. Intelligenz oder Rassismus) angeordnet werden können, wurden erstmals von Louis Guttman (1950) ausgearbeitet, wobei die empirische Ordnungsrelation v >° w zwischen je zwei Probanden v und w durch die Regel

$$([x_{vi} = 1] \wedge [x_{wi} = 0]) \Rightarrow v >^{\circ} w \qquad [2.2.3.1]$$

definiert wird. Wird ein Item von beiden oder von keiner der Vpn gelöst (bzw. bejaht), so sagt es dagegen nichts darüber aus, wie sich die Probanden bezüglich der gemessenen Eigenschaftsdimension voneinander unterscheiden.

Abb. 2.2.3.1: Guttman-Skala mit 3 Items. Nichtnegative ganze Zahlen als Skalenwerte



Somit definiert jedes Item eine Ordnungsrelation zwischen je zwei Probanden, und die Probanden können aufgrund ihrer Antworten genau dann auf einer Eigenschaftsdimension angeordnet werden, wenn diese Relationen einander nicht widersprechen. D.h. es darf nicht vorkommen, dass ein Item die Relation v >° w, ein anderes Item jedoch die Relation w >° v definiert. Diese Forderung, die sich unmittelbar aus der Assymetrie von Ordnungsrelationen ergibt, hat zur Folge, dass man die Items so anordnen können muss, dass

$$x_{vi} = \begin{cases} 1 & \text{wenn } \theta_v > \delta_i \\ 0 & \text{wenn } \theta_v \leq \delta_i \end{cases}, \qquad [2.2.3.2]$$

worin θ_v den Rangplatz des Probanden (v) und δ_i den (Schwierigkeits-)Rangplatz des Items (i) bezeichnet. Die Eigenschaft der Probanden und die Schwierigkeit der Items werden damit auf der selben Rangskala gemessen (vgl. Abb. 2.2.3.1).

Wenn wir die Rangplätze mit nichtnegativen ganzen Zahlen durchnummerieren, ist der Rangplatz eines Probanden auf der Guttman-Skala gleich der Anzahl der von ihm gegebenen positiven Antworten

$$\theta_v = \sum_{i=1}^{k} x_{vi} \qquad [2.2.3.3]$$

und die Itemrangplätze resultieren aus der Häufigkeit, mit der sie positiv beantwortet werden, wobei jenes Item, das am häufigsten gelöst wurde, den Schwierigkeitsrangplatz Null erhält, u.s.w.

Eine metrische Information ist in der Guttman-Skala jedoch nicht enthalten: Die Abstände zwischen den Skalenwerten zweier Items sagen nichts darüber aus, wie stark sich die Items in ihrer Schwierigkeit unterscheiden, und die Abstände zwischen den Skalenwerten zweier Probanden sagen nichts darüber aus, wie sehr sie sich bezüglich der gemessenen Eigenschaftsdimension voneinander unterscheiden.

Skalentypen

Die Skaleneigenschaft einer Messung ist durch die Klasse der zulässigen Maßzahltransformationen definiert, welche die homomorphe Abbildung der empirischen Relationen auf die numerischen Relationen nicht zerstören und somit eine gleichwertige Skala liefern. Je aussagekräftiger eine Skala ist, desto stärkeren Restriktionen unterliegt die Klasse der zulässigen Maßzahltransformationen.

- So ist bei Nominalskalen jede eineindeutige Maßzahltransformation zulässig, welche gleiche Maßzahlen auf gleiche und verschiedene Maßzahlen auf verschiedene abbildet.
- Bei der Ordinalskala dagegen sind nur noch streng monoton wachsende Transformationen zulässig, so dass die Größer-kleiner-Relation erhalten bleibt.
- Bei der Intervallskala sind nur noch positive lineare Transformationen der Form

  $$X \rightarrow Y : x \rightarrow y = a + cx \quad \text{mit } c > 0 \qquad [2.2.3.4]$$

  zulässig, welche den Nullpunkt der Skala um den Betrag a verschieben und die Maßeinheit um den Faktor c vergrößern bzw. verkleinern.
- Bei der Differenzenskala ist die lineare Transformation so eingeschränkt, dass c = 1. D. h., es ist hier nur noch die Addition einer Konstanten zulässig, welche den Nullpunkt der Skala verschiebt, während die Maßeinheit unverändert bleibt.
- Bei der Rationalskala dagegen ist die lineare Transformation so eingeschränkt, dass a = 0. D. h. es ist hier nur noch die Multiplikation mit einer positiven Konstanten zulässig, welche die

Maßeinheit verändert, während der Nullpunkt erhalten bleibt.
- Bei der Absolutskala ist keinerlei Transformation der Skalenwerte mehr zulässig bzw. nur die sog. identische Transformation, welche jeden Skalenwert in sich selbst überführt, so dass Y = X.

Wird auf eine höhere Skala eine Transformation angewendet, die nur für einen niedrigeren Skalentyp zulässig ist, so wird die Skala dadurch nicht gänzlich zerstört. Der Skalentyp sinkt jedoch auf das Niveau der in der Hierarchie tiefer stehenden Skala ab.

Aus diesem Grunde ist die Messung psychologischer Variablen zumeist nur nominal oder ordinal, und zwar selbst dann, wenn als Indikatoren der psychologischen Variablen auf Beobachtungsgrößen zurückgegriffen werden kann, die auf einer höheren Skala gemessen werden.

So ist z. B. "Hunger" eine subjektseitig definierte (latente) Variable, für die sich verschiedene objektseitig definierte (manifeste) Indikatoren angeben lassen, die sich – wie z. B. die Dauer des Nahrungsentzuges – z. T. sogar auf einer Rationalskala messen lassen. Zwischen diesem Indikator und dem Hunger wird ein streng monoton wachsender Zusammenhang postuliert. So lange über die genaue Form dieses Zusammenhanges jedoch nichts bekannt ist, können wir die Dauer des Nahrungsentzuges noch so genau messen, über den Hunger werden wir daraus nur ordinale Informationen beziehen.

Manifeste und latente Variablen

Manifeste Variablen sind solche, die sich direkt beobachten lassen, wie z. B. die Testleistung in einem Intelligenztest. Latente Variablen können dagegen nicht direkt beobachtet werden, sondern sind aus beobachtbaren Variablen abgeleitet, z. B. Intelligenz.

Die Itemcharakteristikfunktion

Um von manifesten auf latente Variablen schließen zu können, muss eine gesetzmäßige Beziehung angegeben werden, wie die zu erschließende latente Variable mit den beobachteten manifesten Indikatoren zusammenhängt.

Z. B. postuliert man in der differentiellen Psychologie:

- Je intelligenter eine Vp (v) ist, desto eher wird sie Aufgaben (Items) einer bestimmten Art (I) lösen können.
- Und umgekehrt schließt man dann: Je wahrscheinlicher es ist, dass eine Vp eine Aufgabe $i \in I$ lösen kann, desto intelligenter ist sie.

Die Wahrscheinlichkeit des manifesten Indikators Lösung der Testaufgabe (p_{vi}) kann somit also als streng monoton wachsende Funktion der latenten Dimension (Θ) dargestellt werden, die durch den Test gemessen wird, wobei θ_v die Position der Vp v auf dem latenten Kontinuum bezeichnet:

$$\forall_v \forall_{i \in I} : \; p_{vi} = f_i(\theta_v). \qquad [2.2.3.5]$$

Diese Funktion wird in der Psychometrie als Itemcharakteristik bezeichnet.

Über die Form der Itemcharakteristik ist zunächst nur so viel vorausgesetzt, dass es sich um eine streng monoton wachsende Funktion handelt, so dass

$$\theta_v > \theta_w \Leftrightarrow p_{vi} > p_{wi}. \qquad [2.2.3.6]$$

Die Skaleneigenschaften der beteiligten Variablen

Dabei haben wir es mit drei verschiedenen Arten von Variablen zu tun:

1. Mit manifesten Antwortvariablen X_{vi}. Diese beschreiben die Antwort, welche Vp v auf Item i gibt, und sind qualitative (dichotome) Variablen (gelöst/nicht gelöst), die auf einer Nominalskala gemessen werden:

$$x_{vi} = \begin{cases} 1 & \text{falls Item i gelöst} \\ 0 & \text{falls Item i nicht gelöst} \end{cases}$$

2. Mit Lösungswahrscheinlichkeiten p_{vi}. Diese sind nicht direkt beobachtbar (also latent) und werden auf einer Absolutskala gemessen.
3. Mit der durch den Test gemessenen Eigenschaft Θ, die ebenfalls latent ist und deren Skaleneigenschaft davon abhängt, welche Form die Itemcharakteristiken haben.

Genauer gesagt, wird die Skaleneigenschaft der latenten Dimension Θ wieder durch die Klasse der zulässigen Maßzahltransformationen bestimmt, welche die Form des funktionalen Zusammenhangs zwischen den Lösungswahrscheinlichkeiten p_{vi} und der latenten Dimension Θ nicht zerstören.

Die Spezifikation der Itemcharakteristiken

Bezüglich der Itemcharakteristiken ist bisher nur soviel gesgt, dass es sich um streng monoton wachsende Funktionen handelt. Die dadurch definierte, zweistellige empirische Ordnungsrelation

[2.2.3.6] ist jedoch (allenfalls) für die Konstruktion einer Ordinalskala ausreichend. Daraus ergeben sich die Fragen,

1. wie man die Itemcharakteristiken zu spezifizieren hat, um eine leistungsfähigere Skala zu ermöglichen, und
2. wie man überprüfen kann, ob die derart getroffenen Modellannahmen auch mit der empirischen Wirklichkeit verträglich sind.

Lineare Itemcharakteristiken

Was die Spezifikation der Itemcharakteristiken betrifft, so wurden im Laufe der Geschichte verschiedene Vorschläge gemacht. So experimentierte der Begründer der stochastischen Testtheorie, Paul Lazarsfeld (1901–1976), zunächst mit linearen Itemcharakteristiken (Lazarsfeld, 1950) und stellte die Itemlösungswahrscheinlichkeiten durch eine Geradengleichung dar, die wir hier in der Form $p_{vi} = \alpha_i(\theta_v - \delta_i)$ mit $\alpha_i > 0$ schreiben wollen. Voraussetzungsgemäß ist p_{vi} danach umso größer, je größer der Personenparameter θ_v ist. Wir können θ_v daher auch als einen Fähigkeitsparameter bezeichnen. Je größer δ_i ist, desto kleiner ist dagegen p_{vi}, womit δ_i offensichtlich die Itemschwierigkeit beschreibt. Und je größer α_i ist, desto steiler ist der Anstieg der Itemcharakteristik, so dass wir den Parameter α_i als die Trennschärfe des Items interpretieren können: Je größer der Trennschärfeparameter ist, desto geringere Fähigkeitsunterschiede zwischen zwei Vpn v und w resultieren bereits in deutlich verschiedene Lösungswahrscheinlichkeiten p_{vi} und p_{wi} (vgl. Abb. 2.2.3.2).

Obwohl Lazarsfeld mit der Geradengleichung eine möglichst einfache Funktionsform gewählt hatte, erwies sich die Annahme linearer Itemcharakteristiken am Ende als nicht handhabbar. Da Wahrscheinlichkeiten Zahlen zwischen Null und Eins sind, der Ausdruck $\alpha_i(\theta_v - \delta_i)$ jedoch für jedes endliche δ_i und α_i Werte zwischen $-\infty$ und $+\infty$ annehmen kann, werden die Zusatzannahmen erforderlich, wonach $p_{vi} = 0$ falls $\alpha_i(\theta_v - \delta_i) \leq 0$ bzw. $p_{vi} = 1$ falls $\alpha_i(\theta_v - \delta_i) \geq 1$. Die lineare Itemcharakteristik kann dann in der Form

$$f_i(\theta_v) = \begin{cases} 1 & \text{für } \alpha_i(\theta_v - \delta_i) \geq 1 \\ \alpha_i(\theta_v - \delta_i) & \text{für } 0 \leq \alpha_i(\theta_v - \delta_i) \leq 1 \\ 0 & \text{für } \alpha_i(\theta_v - \delta_i) \leq 0 \end{cases} \qquad [2.2.3.7]$$

geschrieben werden. Sie ist aber keine stetige Funktion mehr.

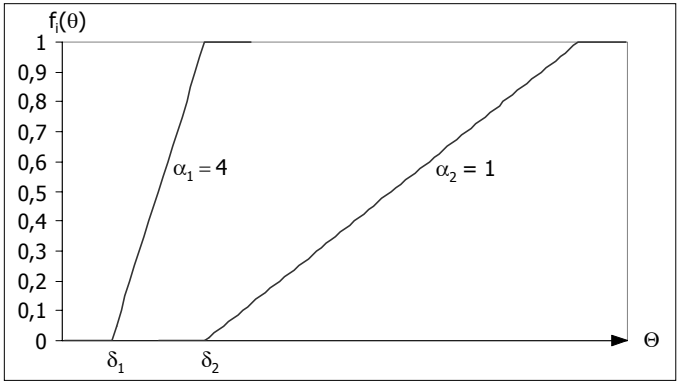


Abb. 2.2.3.2: Lineare Itemcharakteristiken

Das Binomialmodell

Behält man die Geradengleichung als Funktionsform bei, so lässt sich dieses Problem nur dadurch umgehen, dass man von lauter gleich schwierigen und gleich trennscharfen Items mit $\delta_i = \delta_o$ und $\alpha_i = \alpha_o$ ausgeht, so dass die Lösungswahrscheinlichkeiten $p_{vi} = p_{v\bullet}$ alle gleich groß sind. O.B.d.A. können wir die Modellparameter dann zu $\delta_o = 0$ und $\alpha_o = 1$ normieren und die linearen Itemcharakteristiken nehmen alle die selbe itemunabhängige Form

$$f_i(\theta_v) = f_\bullet(\theta_v) = p_{v\bullet} = \theta_v \qquad [2.2.3.8]$$

an (vgl. Abb. 2.2.3.3). Dieses Modell, das u. a. von Klauer (1983, 1987) untersucht wurde, wird als Binomialmodell bezeichnet.

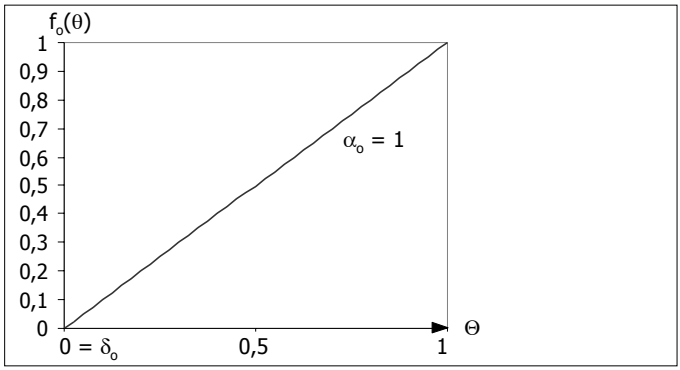


Abb. 2.2.3.3: Die Itemcharakteristik des Binomialmodells

Das Binomialmodell bringt zwar den Vorteil mit sich, dass die Fähigkeitsdimension Θ sogar auf einer Absolutskala gemessen wird; da es jedoch die Ununterscheidbarkeit der Items bezüglich Trennschärfe und Schwierigkeit voraussetzt, erweist es sich kaum als realistisch. Zumindest ist es in seiner Anwendbarkeit äußerst beschränkt.

Das Normal-Ogive-Modell

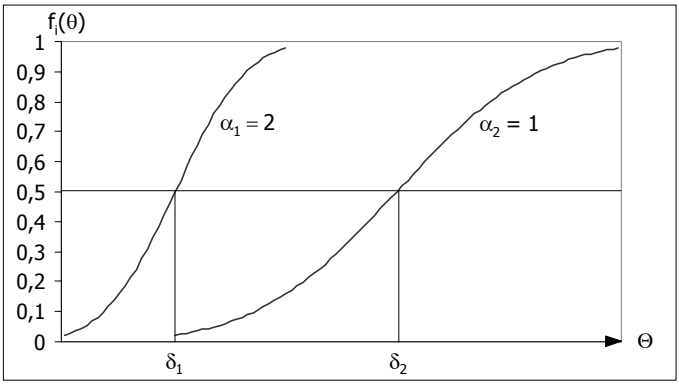
Eleganter und erfolgversprechender ist es, die Annahme linearer Itemcharakteristiken fallen zu lassen und stattdessen eine Funktion zu wählen, deren Wertebereich von vorneherein zwischen Null und Eins liegt, also keine Zusatzannahmen erfordert. Dafür kommt jede stetige Verteilungsfunktion in Frage. Konkret wurde dieser Gedanke bereits von Frederick Lord (1953) umgesetzt, der die Itemcharakteristiken durch die Verteilungsfunktion der Standardnormalverteilung darzustellen versuchte:

$$f_i(\theta_v) = \Phi(\alpha_i(\theta_v - \delta_i)) \qquad [2.2.3.9]$$

$$= \int_{-\infty}^{\alpha_i(\theta_v - \delta_i)} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz.$$

Darin beschreibt θ_v wieder die latente Fähigkeit der Vp, δ_i die Itemschwierigkeit und α_i die Trennschärfe des Items (vgl. Abb. 2.2.3.4).

Abb. 2.2.3.4: Normal-ogive Itemcharakteristiken



Logistische Itemcharakteristiken

Der entscheidende Durchbruch auf der Suche nach einer geeigneten Itemcharakteristikfunktion gelang schließlich Georg Rasch (1901–1980), der die Verteilungsfunktion der Normalverteilung durch die der logistischen Verteilung ersetzte.

Während Rasch (1960) davon ausging, dass sich die Items lediglich in ihrer Schwierigkeit, nicht jedoch in ihrer Trennschärfe voneinander unterscheiden, erweiterte Birnbaum (1968) das logistische Modell um einen zusätzlichen Trennschärfeparameter (vgl. Abb. 2.2.3.5).

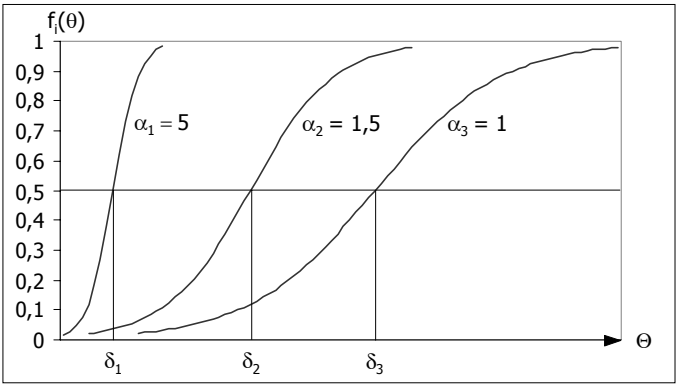


Abb. 2.2.3.5: Logistische Itemcharakteristiken

Das Birnbaum-Modell

$$f_i(\theta_v) = \frac{e^{\alpha_i(\theta_v - \delta_i)}}{1 + e^{\alpha_i(\theta_v - \delta_i)}} \qquad [2.2.3.10]$$

ist vor allem deshalb von Interesse, weil es sowohl die Guttman-Skala als auch das Rasch-Modell und das Binomialmodell als Spezialfälle mit einschließt.

Die Guttman-Skala

Lassen wir in Gleichung [2.2.3.10] den Trennschärfeparameter α_i gegen unendlich streben, dann erhalten wir

$$\lim_{\alpha_i \to \infty} f_i(\theta_v) = 1 \quad \text{für } \theta_v > \delta_i$$

und

$$\lim_{\alpha_i \to \infty} f_i(\theta_v) = 0 \quad \text{für } \theta_v < \delta_i.$$

Da $\theta_V = \delta_i$ wegen prob($\theta_V = \delta_i$) = 0 praktisch ausgeschlossen werden kann, ergibt sich die Guttman-Skala somit als Spezialfall des Birnbaum-Modells bei unendlich großer Itemtrennschärfe.

Das Rasch-Modell

Haben wir es dagegen mit lauter gleich trennscharfen Items zu tun, so dass $\alpha_1 = \alpha_2 = \ldots \alpha_k = 1$, reduziert sich Gleichung [2.2.3.10] auf die Modellgleichung des Rasch-Modells

$$f_i(\theta_v) = \frac{e^{\theta_v - \delta_i}}{1 + e^{\theta_v - \delta_i}}. \qquad [2.2.3.11]$$

Dass die Trennschärfeparameter dabei zu Eins normiert werden können, ist der Tatsache geschuldet, dass die latente Eigenschaft der Vpn (θ_v) und die Itemschwierigkeit (δ_i) im Birnbaum-Modell auf derselben Intervallskala gemessen werden, so dass

$$\Theta \rightarrow \Theta': \theta_v \rightarrow \theta'_v = a + b\,\theta_v$$

$$\delta_i \rightarrow \delta'_i = a + b\,\delta_i$$

eine gleichwertige Skala mit

$$\alpha'_i\,(\theta'_v - \delta'_i) = \alpha_i\,(\theta_v - \delta_i)$$

liefert, wenn die Rationalskala der Itemtrennschärfe gleichzeitig nach

$$A \rightarrow A': \alpha_i \rightarrow \alpha'_i = (b^{-1})\,\alpha_i$$

transformiert wird.

Haben die Itemcharakteristiken die Form des Rasch-Modells, dann werden die latente Eigenschaft der Vpn (θ_v) und die Schwierigkeit der Items (δ_i) dagegen auf einer Differenzenskala gemessen. Im Rasch-Modell ist nur noch die Maßzahltransformation

$$\Theta \rightarrow \Theta': \theta_v \rightarrow \theta'_v = a + \theta_v$$

$$\delta_i \rightarrow \delta'_i = a + \delta_i$$

zulässig. Sie liefert eine gleichwertige Skala mit

$$\theta'_v - \delta'_i = \theta_v - \delta_i,$$

so dass die Funktionsgleichung [2.2.3.11] erhalten bleibt, wenn man die Skalenwerte durch die transformierten Werte ersetzt. Die Skala ist somit bis auf eine additive Konstante bestimmt und lediglich der Nullpunkt kann willkürlich festgesetzt werden.

Das Binomialmodell

Sind die Items darüber hinaus alle gleich schwierig, so können wir o.B.d.A. $\delta_1 = \delta_2 = \ldots = \delta_k = 0$ setzen und Gleichung [2.2.3.11] nimmt die itemunabhängige Form

$$f_i(\theta_v) = f_\bullet(\theta_v) = \frac{e^{\theta_v}}{1 + e^{\theta_v}} \qquad [2.2.3.12]$$

an. Diese Gleichung ist dem Binomialmodell äquivalent. Sie verwendet jedoch eine andere Parametrisierung, wobei der Personenparameter des Rasch-Modells ($\theta_v^{(R)}$) mittels der streng monoton wachsenden Transformation

$$\Theta^{(B)} \to \Theta^{(R)} : \theta_v^{(B)} = p_{v\bullet} \to \theta_v^{(R)} = \ln\left(\frac{p_{v\bullet}}{1 - p_{v\bullet}}\right) \qquad [2.2.3.13]$$

aus dem Fähigkeitsparameter des Binomialmodells ($\theta_v^{(B)}$) hervorgeht. Der Beweis sei dem Leser als Übungsaufgabe überlassen.

Latent-Class-Modelle

Testtheoretische Modelle, welche die Itemantworten mittels Itemcharakteristikfunktion auf eine quantitative latente Eigenschaft (engl. *trait*) der Probanden zurückführen, bezeichnet man in der Psychometrie als Latent-Trait-Modelle. Ist die latente Variable dagegen qualitativer (klassifikatorischer) Art, so spricht man von Latent-Class-Modellen.

Dass man in der Psychodiagnostik nicht unbedingt davon ausgehen kann, dass sich die Probanden auf einer latenten Dimension anordnen lassen, haben wir bereits in Band 1 diskutiert. Namentlich in der klinischen Psychologie geht man traditionell davon aus, dass es nicht die Anzahl der Krankheitssymptome ist, welche ein spezifisches Krankheitsbild ausmacht, sondern ein typisches Symptommuster (Syndrom).

Depressive Syndrome

So beschreiben z. B. Immich et al. (1971) die depressive Symptomatik durch die fünf Symptome

1. Psychomotorik herabgesetzt (PH),
2. Denkablauf gehemmt (DG),
3. ängstlich-gequält-klagsame Stimmung (ÄGK),
4. vital-traurige Stimmung (VT) und
5. Hypochondrien, Schuld- und Verarmungsgedanken (HSV),

deren Verteilung in einer Stichprobe von 238 klinischen Patienten in Abb. 2.2.3.6 wiedergegeben ist.

Die häufigsten Depressionssymptome sind demnach ängstlich-gequält-klagsame Stimmung (75.5% der Patienten), gehemmter Denkablauf (71.4%) und herabgesetzte Psychomotorik (70.6%). Auch Hypochondrien, Schuld- und Verarmungsgedanken finden

sich bei mehr als der Hälfte der Patienten (56.3%), wohingegen eine vital-traurige Stimmung nur bei etwa einem Viertel der Patienten auftritt (25.2%).

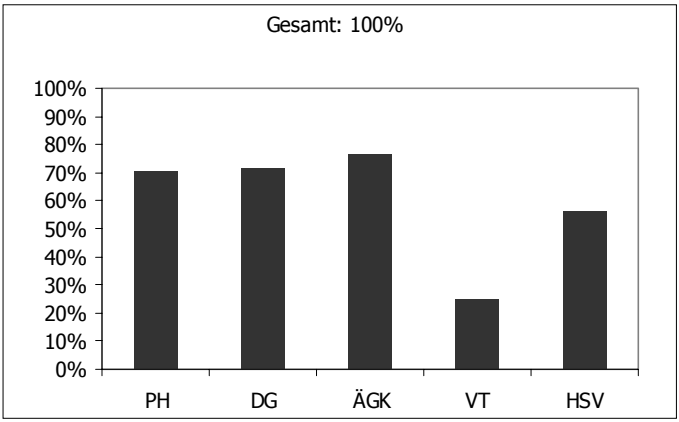


Abb. 2.2.3.6: Verteilung von fünf depressiven Symptomen in einer Stichprobe von 238 klinischen Patienten

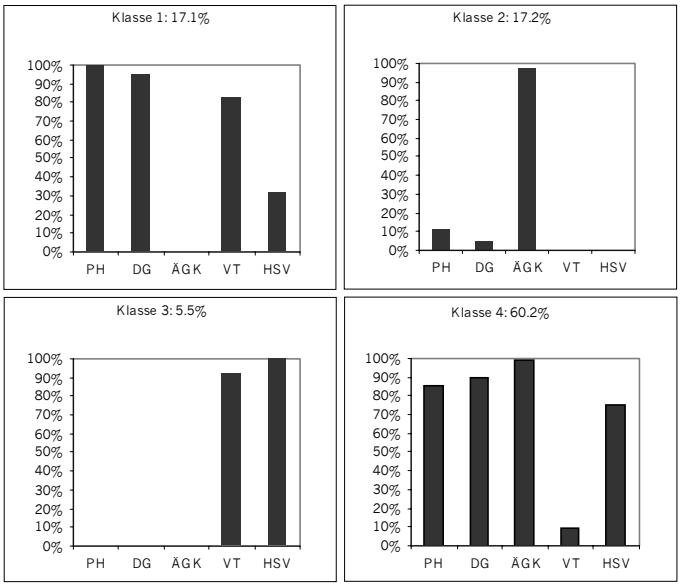


Abb. 2.2.3.7: Die Symptomatik verschiedener Patientengruppen

In der Häufigkeitsverteilung in Abb. 2.2.3.6 sind jedoch die Symptome unterschiedlicher Patientengruppen (Klassen von Patienten) miteinander vermischt, die sich in ihrer Symptomatik deutlich voneinander unterscheiden (vgl. Abb. 2.2.3.7).

- Die Symptomatik der beiden größten Patientengruppen ist durch die ängstlich-gequält-klagsame Stimmung der Patienten geprägt.
- In der kleineren der beiden Gruppen (Klasse 2, 17.2% der Patienten) steht dieses Symptom in der Regel allein. Nur gelegentlich ist auch die Psychomotorik der Patienten herabgesetzt und in sehr seltenen Fällen treten Denkhemmungen auf. Im Unterschied dazu ist die größere der beiden Klassen (Klasse 4, 60% der Patienten) insgesamt symptomreich. Fast alle Symptome treten mit großer Wahrscheinlichkeit auf. Gelegentlich treten zur ängstlich-gequält-klagsamen Stimmung sogar vital-traurige Stimmungslagen hinzu.
- In den anderen beiden Patientengruppen ist eine ängstlich-gequält-klagsame Stimmung dagegen überhaupt nicht zu finden. Ihre Stimmung ist meist vital-traurig.
- Auch hier ist die größere der beiden Gruppen (Klasse 1, 17.1%) insgesamt symptomreich, wenngleich Hypochondrien, Schuld- und Verarmungsgedanken bei diesen Patienten deutlich seltener auftreten als in der Gesamtgruppe. Die Gruppe unterscheidet sich damit von der kleinsten Patientengruppe (Klasse 3, 5.5%), die zu 100% durch Hypochondrien, Schuld- und Verarmungsgedanken geprägt ist, in der herabgesetzte Psychomotorik und gehemmter Denkablauf jedoch überhaupt nicht zu finden sind.

Das Modell der Latent-Class-Analyse

Die empirische Identifikation solcher Syndrome ist Gegenstand der auf Lazarsfeld (1950) zurückgehenden Latent-Class-Analyse (LCA).

Allgemein formuliert, haben wir es dabei mit k manifesten Variablen ($i = 1,..,k$) und h latenten Klassen ($g = 1,...,h$) zu tun:

- Die manifesten Variablen sind qualitativ und können jeweils m_i verschiedene Werte annehmen, die wir o.B.d.A. mit $x = 0,1,...,m_i-1$ bezeichnen.
- Das Charakteristikum, wodurch sich die latenten Klassen voneinander unterscheiden, besteht in den klassenspezifischen Kategorienwahrscheinlichkeiten $p_{gix} = \text{prob}\{X_{vi}=x|v\in g\}$, welche die Wahrscheinlichkeit beschreiben, mit der eine Vp aus Klasse g in Variable i den Wert x erzielt.
- Die Wahrscheinlichkeit $\text{prob}\{v\in g\}$, mit der eine zufällig herausgegriffene Vp der Klasse g angehört, entspricht der (relativen) Klassengröße p_g.

Für die Wahrscheinlichkeit der Datenmatrix (X) folgt daraus die Modellgleichung der LCA:

$$\text{prob}(\mathbf{X}) = \prod_{v=1}^{n} \text{prob}\{(x_{v1},\ldots,x_{vk})\} \qquad [2.2.3.14]$$

$$= \prod_{v=1}^{n}\left(\sum_{g=1}^{h} \text{prob}\{v \in g\}\ \text{prob}\{(x_{v1},\ldots,x_{vk})|v\in g\}\right)$$

$$= \prod_{v=1}^{n}\left(\sum_{g=1}^{h} \text{prob}\{v \in g\} \prod_{i=1}^{k} \text{prob}\{X_{vi} = x_{vi}|v\in g\}\right)$$

Ein-Klassen-Lösung und saturiertes Modell

Weder die Klassengrößen p_g noch die klassenspezifischen Kategorienwahrscheinlichkeiten p_{gix} sind bekannt, und auch von der Anzahl der Klassen, die zur Beschreibung der Daten erforderlich sind, wissen wir lediglich, dass sie zwischen $1 \le h \le h_{max}$ liegt, worin

$$h_{max} = \prod_{i=1}^{k} m_i \qquad [2.2.3.15]$$

die Anzahl der möglichen Antwortmuster $(x_{v1},\ldots,x_{vk})$ darstellt. Die gesuchte Lösung der LCA liegt also irgendwo zwischen zwei Extremen, wonach entweder

- alle Daten aus der selben homogenen Verteilung stammen (sog. Ein-Klassen-Lösung; h = 1), wie dies bei der traditionellen statistischen Analyse angenommen wird, oder
- jedes mögliche Antwortmuster eine eigene Klasse darstellt (sog. saturiertes Modell; $h = h_{max}$).

2.2.4 Das Konzept der lokalen stochastischen Unabhängigkeit

Definition

Bereits in Kapitel 2.2.3, Gleichung [2.2.3.14] haben wir stillschweigend vorausgesetzt, dass die Items lokal unabhängig sind, so dass die Wahrscheinlichkeit des Antwortvektors in der Form

$$\text{prob}(x_v) = \text{prob}(x_{v1},\ldots,x_{vk}) = \prod_{i=1}^{k} \text{prob}(x_{vi}) \qquad [2.2.4.1]$$

dargestellt werden kann. Diese Voraussetzung, die in der stochastischen Testtheorie (fast) durchgehend getroffen wird, bedarf einiger Erläuterungen, da sie auf den ersten Blick der Erfahrung zu

widersprechen scheint, dass Testitems (je nach Personenstichprobe mehr oder minder hoch) miteinander korrelieren.

Item-interkorrelationen

Wie man zeigen kann, stehen die Iteminterkorrelationen jedoch nicht in Widerspruch zu Gleichung [2.2.4.1], welche die Unabhängigkeit der Testitems nur für eine feste Vp v behauptet. Man spricht deshalb von *lokaler* stochastischer Unabhängigkeit der Items. Auch lokal unabhängige Items korrelieren miteinander, und zwar desto höher, je homogener der Test ist und je heterogener die Personenstichprobe ist, in welcher die Korrelation berechnet wird.

Sind die Personenstichproben ideal homogen, so verschwindet die Korrelation.

Von einem homogenen Test sprechen wir dann, wenn die Items alle die selbe latente Dimension messen, d. h. die selbe Ordnungsrelation zwischen je zwei Vpn definieren. Ist dies nicht der Fall, so ist der Test heterogen.

Eine Personenstichprobe bezeichnen wir dagegen als desto heterogener, je mehr sich ihre Mitglieder auf der latenten Dimension voneinander unterscheiden. Umgekehrt bezeichnen wir sie als desto homogener, je geringer die durch den Test gemessene Eigenschaft streut.

Homogene Items

Z. B. betrachten wir zwei homogene Items i = 1 und i = 2, so dass für je zwei Vpn v und w

$$p_{v1} > p_{w1} \leftrightarrow p_{v2} > p_{w2} ,$$

und drei in sich homogene Stichproben von Vpn mit

- Stichprobe 1: $p_{11} = 0.9$ und $p_{12} = 0.8$,
- Stichprobe 2: $p_{21} = 0.8$ und $p_{22} = 0.7$ und
- Stichprobe 3: $p_{31} = 0.3$ und $p_{32} = 0.2$.

Die Items seien lokal unabhängig, so dass sich bei einem Stichprobenumfang von jeweils $n_1 = n_2 = n_3 = 100$ die in Tab. 2.2.4.1 dargestellten Vierfeldertafeln ergeben und die Item-Interkorrelationen innerhalb der Stichproben verschwinden.

Betrachten wir dagegen die Gesamtstichprobe aller n = 300 Vpn, so ist die Stichprobe nicht mehr homogen und es ergibt sich die Vierfeldertafel in Tab. 2.2.4.2, in welcher die Items zu

$$r_\Phi = \frac{134 \times 64 - 66 \times 36}{\sqrt{170 \times 130 \times 200 \times 100}} = 0.295$$

miteinander korrelieren.

Tab. 2.2.4.1: Drei in sich homogene Stichproben

Stichprobe 1

	+	-	
+	72	18	90
-	8	2	10
	80	20	100

Stichprobe 2

	+	-	
+	56	24	80
-	14	6	20
	70	30	100

Stichprobe 3

	+	-	
+	6	24	30
-	14	56	70
	20	80	100

Tab. 2.2.4.2: Die Gesamtstichprobe

	+	-	
+	134	66	200
-	36	64	100
	170	130	300

Inhomogene Items

Die Korrelation zwischen den Items kann auch dann noch erhalten bleiben, wenn die Items inhomogen sind, z. B. wenn sich in einer vierten in sich homogenen Stichprobe ergibt:

- Stichprobe 4: $p_{41} = 0.1$ und $p_{42} = 0.6$,

so dass die beiden Items unterschiedliche Rangordnungen der in den Stichproben enthaltenen Vpn definieren:

$p_{11} > p_{21} > p_{31} > p_{41}$ (Item 1)

und

$p_{12} > p_{22} > p_{42} > p_{32}$ (Item 2).

Angenommen, die vierte Stichprobe hat ebenfalls wieder den Umfang $n_4 = 100$, dann ergibt sich bei lokaler Unabhängigkeit der Items die Vierfeldertafel in Tab. 2.2.4.3

Tab. 2.2.4.3: Eine vierte, in sich homogene Teilstichprobe

	+	-	
+	6	4	10
-	54	36	90
	60	40	100

und in der erweiterten Gesamtstichprobe (Tab. 2.2.4.4)

Tab. 2.2.4.4: Die erweiterte Gesamtstichprobe

	+	-	
+	140	70	210
-	90	100	190
	230	170	400

erhalten wir eine Korrelation von

$$r_{\Phi} = \frac{140 \times 100 - 70 \times 90}{\sqrt{230 \times 170 \times 210 \times 190}} = 0.195 \,.$$

Die Items korrelieren also trotz Inhomogenität immer noch miteinander. Die Korrelation hat jedoch abgenommen.

Konsequenzen für die Testkonstruktion

Auch wenn die Annahme der lokalen Unabhängigkeit also durchaus damit verträglich ist, dass die Items eines Tests in so gut wie jeder Personenstichprobe miteinander korrelieren, resultieren daraus jedoch gewisse Einschränkungen, die man bei der Testkonstruktion zu berücksichtigen hat:

1. Man muss dafür sorgen, dass die Items logisch unabhängig sind. Z. B. darf der Test keine Items enthalten, deren Lösung bereits eine Teillösung späterer, schwierigerer Items darstellt.
2. Man muss ausschließen können, dass während der Testbearbeitung ein Lernprozess stattfindet, so dass z. B. die Erfolgschance der Vpn mit der Anzahl der zuvor gelösten Items zunimmt.

Serielle Abhängigkeit

Theoretische Ansätze, erfolgskontingentes Lernen in Latent-Trait-Modellen abzubilden und die lokale Unabhängigkeit durch eine serielle Abhängigkeit der Form

$$\text{prob}(x_v) = \text{prob}(x_{v1}, \ldots, x_{vk}) = \prod_{i=1}^{k} \text{prob}\left(x_{vi} \middle| (x_{v1}, \ldots, x_{vi-1})\right) \quad [2.2.4.2]$$

zu ersetzen (Fischer, 1972; Kempf 1974; 1977; Verhelst & Glas, 1993), haben sich als wenig zielführend erwiesen, so dass man heute davon ausgehen kann, dass die lokale Unabhängigkeit nicht nur eine technische Annahme, sondern – von einzelnen Spezialfällen abgesehen – eine unverzichtbare Voraussetzung der stochastischen Testtheorie darstellt (siehe hierzu Kap. 2.3.2, 2.3.3, 2.9.3 und 2.9.4).

Kontrollfragen

- Welche Beziehung besteht zwischen Maßzahlvariable, True-Score-Variable und Fehlervariable eines Tests t?
- Wie lauten die Axiome von Gulliksen?
- Wie lautet Novicks Modell der klassischen Testtheorie?
- Welche Konsequenzen ergeben sich daraus für das Verständnis von True-Score und Messfehler?
- Was versteht man unter
  a) der Reliabilität
  b) der Validität
  eines Tests und wie sind diese definiert?
- In welche Faktoren kann die Validität aufgespalten werden?
- Was versteht man unter
  a) parallelen
  b) essentiell parallelen
  c) τ-äquivalenten und
  d) essentiell τ-äquivalenten
  Tests bzw. Testungen?
- Was beschreibt die Validität eines Tests bezüglich eines Paralleltests?
- Worin besteht die Paralleltest-Methode der Reliabilitätsbestimmung und welche Varianten der Paralleltest-Methode lassen sich unterscheiden?
- Was beschreibt die Korrelation zweier Tests, wenn es sich bei den Tests
  a) um parallele Tests
  b) um essentiell parallele Tests
  c) um τ-äquivalente Tests
  d) um essentiell τ-äquivalente Tests
  handelt, und was beschreibt sie, wenn
  e) keine dieser Voraussetzungen zutrifft, die beiden Tests aber gleich reliabel sind, und
  f) wenn auch diese Voraussetzung nicht erfüllt ist?
- Wie groß kann die Validität eines Tests höchstens werden?
- Was beschreibt die Reliabilität eines Tests und welcher Zusammenhang besteht zwischen der Reliabilität und der Homogenität der Referenzpopulation?
- Was versteht man unter dem Standardmessfehler und wie wird er berechnet?
- Wie kann man die Reliabilität eines Tests verbessern?
- Was beschreibt die Spearman-Brown-Formel und welche Voraussetzungen müssen erfüllt sein, damit sie anwendbar ist?
- Wie verändert sich die Reliabilität mit zunehmender Testlänge und welche Voraussetzungen müssen erfüllt sein, damit man die bei n-facher Testlänge erwartete Reliabilität berechnen kann?
- Worin besteht
  a) die Testteilungemethode der Reliabilitätsbestimmung
  b) die interne Konsistenzanalyse
  und welche Varianten dieser Methoden lassen sich unterscheiden?

- Worin besteht der Vorteil der internen Konsistenzanalyse gegenüber der Testteilungsmethode?
- Warum ist von der Anwendung der Rechenformel KR-21 abzuraten?
- Welche Methoden der Reliabilitätsbestimmung gibt es, an welche Voraussetzungen sind sie gebunden und wie aussagekräftig sind ihre Ergebnisse, wenn diese Voraussetzungen verletzt sind?
- Wie kann man aus dem Vergleich verschiedener Reliabilitätsmaße Anhaltspunkte dafür gewinnen,
  a) ob die jeweiligen Anwendungsvoraussetzungen erfüllt sind, und
  b) wie groß die Reliabilität des Tests tatsächlich ist?
- Was versteht man unter einer Messung?
- Worin besteht der Unterschied zwischen
  a) quantitativen und qualitativen
  b) dichotomen und polytomen Variablen?
- Wie sind die empirischen Ordnungsrelationen definiert, auf denen die Guttman-Skala beruht?.
- Welche Voraussetzung müssen diese Ordnungsrelationen erfüllen, damit die Probanden auf einer quantitativen Dimension angeordnet werden können?
- Wodurch ist die Skaleneigenschaft einer Messung definiert?
- Welche Maßzahltransformationen sind bei
  a) Nominalskalen
  b) Ordinalskalen
  c) Intervallskalen
  d) Differenzenskalen
  e) Rationalskalen
  f) Absolutskalen
  zulässig?
- Worin besteht der Unterschied zwischen manifesten und latenten Variablen?
- Was muss gegeben sein, damit man von manifesten Indikatoren auf latente Variablen schließen kann?
- Erläutern Sie den Terminus Itemcharakteristik: Was beschreibt sie, welche Form hat sie und welche Ordnungsrelation bezüglich der zugrunde liegenden Eigenschaft der Vpn wird dadurch definiert?
- Welcher Zusammenhang besteht zwischen dem Skalentyp der manifesten Variablen und dem Skalentyp der latenten Variablen, für welche sie als Indikatoren dienen?
- Diskutieren Sie die verschiedenen Modellannahmen, die man bezüglich der Funktionsform der Itemcharakteristiken treffen kann, sowie die Bedeutung der darin vorkommenden Modellparameter!
- Diskutieren Sie die Vor- und Nachteile der o.g. Modellannahmen!
- Inwiefern stellt
  a) das Rasch-Modell
  b) die Guttman-Skala
  einen Spezialfall des Birnbaum-Modells dar?
- Inwiefern stellt das Binomialmodell einen Spezialfall
  a) linearer Itemcharakteristiken sowie
  b) des Rasch-Modells
  dar und durch welche Skalentransformation können die beiden Modelldarstellungen ineinander übergeführt werden?

- Worin besteht der Unterschied zwischen Latent-Trait-Modellen und Latent-Class-Modellen?
- Wie lautet die Modellgleichung der Latent-Class-Analyse?
- Inwiefern handelt es sich bei der Ein-Klassen-Lösung und dem saturierten Modell um Spezialfälle der Latent-Class-Analyse?
- Was versteht man unter der lokalen stochastischen Unabhängigkeit von Testtems?
- Erläutern Sie, ob bzw. inwiefern die Annahme der lokalen Unabhängigkeit in Widerspruch dazu steht, dass Testitems in der Regel miteinander korrelieren!
- Was versteht man unter der Homogenität bzw. Heterogenität
  a) eines Tests
  b) einer Personenstichprobe?
- Welche Konsequenzen ergeben sich aus der Annahme der lokalen Unabhängigkeit der Testitems für die Testkonstruktion?
- Was versteht man unter serieller Abhängigkeit von Testitems?

Übungsaufgaben

1. Ein Intelligenztest wurde einer Stichprobe von 10 Abiturienten vorgegeben. Das Testergebnis wird in IQ-Punkten ausgedrückt. Berechnen Sie die Validität des Tests bezüglich der durchschnittlichen Abiturnote der Abiturienten!

Vp-Nr.	1	2	3	4	5	6	7	8	9	10
IQ-Punkte	117	112	108	109	118	115	111	116	120	107
Abiturnote	2.9	1.9	1.3	2.5	2.3	1.6	2.3	1.5	1.2	3.1

2. Ein Test aus 20 dichotomen Items wurde einer Stichprobe von 12 Vpn nach 6 Monaten erneut vorgegeben. Die erzielten Testergebnisse sind in der nachstehenden Tabelle wiedergegeben.
   a) Berechnen Sie die Retest-Reliabilität des Tests!
   b) Welche Voraussetzungen müssen erfüllt sein, damit der errechnete Zahlenwert tatsächlich die Reliabilität des Tests abbildet?
   c) Was kann über die Reliabilität des Tests ausgesagt werden, wenn diese Voraussetzungen verletzt sind?

Vp-Nr.	1	2	3	4	5	6	7	8	9	10	11	12
Test	10	14	17	16	12	19	13	18	11	15	9	8
Testwiederholung	9	15	16	15	13	18	13	14	12	14	14	7

3. Von dem Test in Aufgabe 2 existiert eine Parallelform, die das selbe Verhaltensmerkmal gleich gut messen soll und den Vpn bei der ersten Testung ebenfalls vorgegeben wurde.
   a) Berechnen Sie die Paralleltest-Reliabilität des Tests!
   b) Was kann über die Reliabilität der beiden Tests ausgesagt werden, wenn sie tatsächlich
   i) nicht parallel, sondern nur essentiell parallel,
   ii) nicht parallel, sondern nur essentiell τ-äquivalent
   iii) nicht einmal essetiell τ-äquivalent sind?

Vp-Nr.	1	2	3	4	5	6	7	8	9	10	11	12
Paralleltest	11	13	18	17	11	20	13	19	10	16	10	9

4. Wie groß kann die Validität eines Intelligenztests höchstens ausfallen, wenn die Reliabilität des Tests 0.9 beträgt?
5. Was besagt eine Reliabilität von 0.9?
6. Die Testergebnisse eines mechanisch-technischen Verständnistests werden in IQ-Punkten ausgedrückt. Die Reliabilität des Tests beträgt 0.94. Wie groß ist der Standardmessfehler der IQ-Punkte?
7. Von einem Test bestehen zwei Parallelformen mit je k = 20 Testitems. Die Korrelation zwischen den beiden Tests beträgt 0.83.
   a) Welche Reliabilität ist zu erwarten, wenn man beide Parallelformen verwendet und daraus den Gesamtscore $X_{ot} = X_{o1} + X_{o2}$ berechnet?
   b) Welche Voraussetzungen müssen erfüllt sein, damit der errechnete Wert tatsächlich mit der Reliabilität des zusammengesetzten Tests übereinstimmt?
8. Ein Test aus 30 Items besitzt die Reliabilität 0.89.
   a) Mit welcher Reliabilität ist zu rechnen, wenn man stattdessen eine Kurzform verwendet, die nur 10 Items umfasst?
   b) Wie viele Items muss die Kurzform des Tests umfassen, wenn ihre Reliabilität nicht unter 0.8 absinken soll?
   c) Mit welcher Reliabilität ist zu rechnen, wenn man den Test auf 50 Items erweitert?
   d) Auf wie viele Items muss der Test erweitert werden, um eine Reliabilität von 0.95 zu erzielen?
9. Ein Test aus 40 dichotomen Items wurde in zwei parallele Subtests zerlegt die zu $r(X_{o1},X_{o2}) = 0.65$ miteinander korrelieren. Berechnen Sie die Reliabilität des aus den beiden Subtests bestehenden Gesamttests!
10. Ein Test wurde in drei parallele Subtests zerlegt, deren Interkorrelationen in nachstehender Tabelle dargestellt sind:

	j' = 1	j' = 2	j' = 3
j = 1	–	0.59	0.43
j = 2		–	0.51
j = 3			–

   a) Schätzen Sie die Reliabilität der Subtests und
   b) berechnen Sie die Reliabilität des Gesamttests!
11. Erwartungswert und Maßzahlvarianz eines Tests betragen $E(X_{ot}) = 10.2$ und $\sigma^2(X_{ot}) = 16.81$. Der Test wurde in zwei Subtests mit den Erwartungswerten $E(X_{o1}) = 4.9$ und $E(X_{o2}) = 5.3$ und den Maßzahlvarianzen $\sigma^2(X_{o1}) = 5.42$ und $\sigma^2(X_{o2}) = 7.82$ zerlegt. Was lässt sich über die Reliabilität des Tests aussagen?
12. Ein Test besteht aus 10 dichotomen Items mit den folgenden Lösungswahrscheinlichkeiten:

	Item-Nr.									
	1	2	3	4	5	6	7	8	9	10
p_{oi}	0.52	0.51	0.53	0.49	0.5	0.41	0.44	0.48	0.51	0.48

Die Maßzahlvarianz des Tests beträgt $\sigma^2(X_{ot})$ = 9.86. Was lässt sich über die Reliabilität des Tests aussagen?

13. Eine testtheoretische Analyse des Subtests WÜ / Form B des IST-70 ergab die folgenden Kennwerte:
    - Paralleltest-Reliabilität = 0.62
    - Split-Half-Reliabilität = 0.89
    - Koeffizient Alpha (α_2) = 0.89
    - Kuder Richardson 20 (α_k) = 0.80

    a) Welche Voraussetzungen müssen erfüllt sein, damit die jeweiligen Kennwerte gleich der Reliabilität des Tests sind?

    b) Was lässt sich aufgrund der Kennwerte darüber aussagen, ob diese Voraussetzungen tatsächlich erfüllt sind?

    c) Was lässt sich aufgrund der Kennwerte über die Reliabilität des Tests aussagen?
14. Zeigen Sie, dass die Gleichung $p_{vo} = e^{\theta_v} / (1 + e^{\theta_v})$ genau dann gilt, wenn $\theta_v = \ln(p_{vo} / (1 - p_{vo}))$!
15. Gegeben seien zwei homogene Items i = 1 und i = 2, so dass für je zwei Vpn v und w die empirische Regel $p_{v1} > p_{w1} \leftrightarrow p_{v2} > p_{w2}$ gilt und eine homogene Stichprobe von n = 150 Vpn mit p_{o1} = 0.7 und p_{o2} = 0.4. Die Items seien lokal unabhängig. Wie lautet die 4-Felder-Tafel?
16. Wie hoch korrelieren die Items in der folgenden 4-Felder-Tafel?

	+	-
+	89	13
-	12	95

17. Gegeben seien drei lokal unabhängige Testitems und eine Vp v mit den Lösungswahrscheinlichkeiten p_{v1} = 0.2, p_{v2} = 0.4 und p_{v3} = 0.6. Wie groß ist die Wahrscheinlichkeit, dass die Vp den Antwortvektor x_v = (1, 0, 1) produziert?

2.3

Klassische und stochastische Testtheorie

2.3.1 Ein allgemeines testtheoretisches Modell

Implizit ist die Annahme der lokalen Unabhängigkeit auch in Novicks (1966) Modell der klassischen Testtheorie enthalten, wie man sofort sieht, wenn man das Modell auf Tests aus k = 1 dichotomen Items anwendet und einen Blick auf die Argumentationsweise wirft, mittels derer die Geltung des dritten Axioms von Gulliksen begründet wird (vgl. Bd. 1, Kap. 2.1.3).

Beweis

Seien t und t' Tests aus je einem dichotomen Testitem i bzw. j, dann sind die Erwartungswerte der Scorevariablen gleich $\tau_{vi} = p_{vi}$ bzw. $\tau_{vj} = p_{vj}$ und die Messfehler $F_{vi} = X_{vi} - \tau_{vi}$ bzw. $F_{vj} = X_{vj} - \tau_{vj}$ unterscheiden sich von den zugehörigen Antwortvariablen X_{vi} bzw. X_{vj} nur durch die Konstanten p_{vi} bzw. p_{vj}.

Dass $E(F_{vt}) = 0$ insbesondere für all jene Vpn gilt, deren Testergebnis in Test t' mit dem Messfehler $f_{vt'} = x_{vj} - p_{vj}$ belastet ist, so dass $E(F_{vt}|f_{vt'}) = 0$, besagt daher, dass $\text{prob}(x_{vi}|x_{vj}) = \text{prob}(x_{vi})$.

Letzteres bedeutet aber nichts anderes als dass die Items lokal unabhängig sind.

Obwohl die klassische und die stochastische Testtheorie verschiedenen Forschungstraditionen entstammen, gehen sie somit von kompatiblen Grundannahmen aus, die wir in einem allgemeinen Modell der psychologischen Testtheorie formulieren können:

Das allgemeine Modell der psychologischen Testtheorie

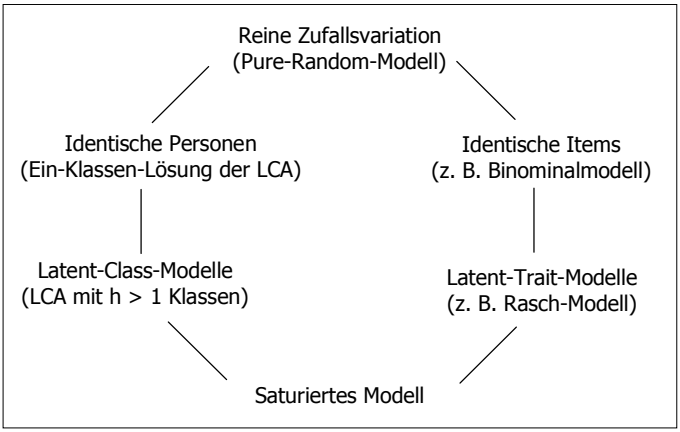
1. Jeder Vp v und jedem Item i entspricht eine diskrete Zufallsvariable X_{vi} mit dem Wertevorrat $x = 0,\ldots,m_i\text{-}1$ und der Wahrscheinlichkeitsdichte $f_{vi}(x)$.
2. Die Items sind lokal unabhängig, so dass für jeden Test aus k Items i = 1,...,k

$$\text{prob}(x_{v1},\ldots,x_{vk}) = \prod_{i=1}^{k} f_{vi}(x_{vi}). \qquad [2.3.1.1]$$

Modellspezifikationen

Bezüglich der Wahrscheinlichkeitsdichte der Antwortvariablen X_{vi} können weitere Annahmen getroffen werden, wobei zwischen den verschiedenen Spezifikationen des allgemeinen testtheoretischen Modells eine Hierarchie besteht, die in Abbildung 2.3.1.1 dargestellt ist.

Abb. 2.3.1.1: Hierarchie der Modellspezifikationen



Reine Zufallsvariation

Sind die Antwortvariablen X_{vi} personen- und itemunabhängig verteilt, so dass alle Vpn und Items die selben Antwortvariablen

$$X_{vi} \equiv X_{\bullet\bullet} \qquad [2.3.1.2]$$

mit der Wahrscheinlichkeitsdichte

$$f_{vi}(x) = f_{\bullet\bullet}(x) \qquad [2.3.1.3]$$
$$= \text{prob}(X_{vi} = x) = p_{\bullet\bullet x}$$

besitzen, dann enthält der Test weder Informationen über Unterschiede zwischen den Vpn noch Informationen über Unterschiede zwischen den Items. In diesem Fall ist er als diagnostisches Instrument unbrauchbar (Pure-Random-Modell).

Identische Personen

Keinerlei diagnostische Informationen über die Vpn enthält der Test auch dann, wenn sich zwar die Antwortvariablen verschiedener Items voneinander unterscheiden, diese jedoch personenunabhängig verteilt sind, so dass alle Vpn dieselben Antwortvariablen

$$X_{vi} \equiv X_{\bullet i} \qquad [2.3.1.4]$$

mit der Wahrscheinlichkeitsdichte

$$f_{vi}(x) = f_{\bullet i}(x) \qquad [2.3.1.5]$$
$$= \text{prob}(X_{vi} = x) = p_{\bullet ix}$$

besitzen (identische Personen). Dieser Fall entspricht der Ein-Klassen-Lösung der LCA.

Latente Personenvariablen

Diagnostische Informationen über die Vpn enthält der Test erst dann, wenn die Antwortvariablen personenabhängig verteilt sind. Ist dies der Fall, so kann die Wahrscheinlichkeitsdichte der Antwortvariablen X_{vi} als Funktion einer latenten Personenvariable Θ dargestellt werden

$$f_{vi}(x) = \text{prob}(X_{vi} = x | \theta_v) = f_i(x, \theta_v), \qquad [2.3.1.6]$$

die entweder ein- oder mehrdimensional sein kann.

Ist Θ eindimensional, so wird die Ausprägung der latenten Personenvariable durch einen Skalar θ_v dargestellt, der die Position der Vp v auf einer latenten Eigenschaftsdimension beschreibt. Ist die latente Personenvariable mehrdimensional, so wird ihre Ausprägung durch einen Vektor $\theta_v = (\theta_{v1},\ldots,\theta_{vh})$ beschrieben, dessen Elemente die Koordinaten sind, welche die Vp in einem h-dimensionalen Eigenschaftsraum verorten.

Aus Gründen der mathematischen Einfachheit betrachten wir zunächst nur den eindimensionalen Fall, wobei wir wiederum zwischen diskreten und kontinuierlichen Personenvariablen unterscheiden können.

Latent-Trait-Modelle

Ist die latente Personenvariable kontinuierlich, so wird sie auf einer metrischen Skala mit dem Wertevorrat $-\infty \leq \theta_v \leq \infty$ gemessen (Latent-Trait-Modelle). Die Wahrscheinlichkeit, dass eine zufällig herausgegriffene Vp v das Item i in der Kategorie x beantwortet, kann dann in der Form

$$\text{prob}(X_{vi} = x) = \int_{-\infty}^{\infty} f_i(x, \theta)\, g(\theta)\, d\theta \qquad [2.3.1.7]$$

dargestellt werden, worin $g(\theta)$ die Wahrscheinlichkeitsdichte der latenten Variable bezeichnet.

Identische Items

Im einfachsten Fall bestehen keine Unterschiede zwischen den verschiedenen Items, so dass alle Items dieselben Antwortvariablen

$$X_{vi} \equiv X_{v\bullet} \qquad [2.3.1.8]$$

mit der Wahrscheinlichkeitsdichte

$$\begin{aligned} f_{vi}(x) &= f_{v\bullet}(x) \\ &= \text{prob}(X_{vi} = x | \theta_v) = p_{v\bullet x} \end{aligned} \qquad [2.3.1.9]$$

besitzen (identische Items). Für dichotome Testitems entspricht dieser Fall dem Binomialmodell.

Latent-Class-Modelle

Ist die latente Personenvariable diskret, so wird sie auf einer Nominal- oder Ordinalskala mit dem Wertevorrat $\theta_g = \theta_1,\ldots,\theta_h$ gemessen, welche die Personenpopulation in h (latente) Klassen g = 1,…,h einteilt, so dass jede Vp genau einer dieser Klassen angehört

$$v \in g \Leftrightarrow \theta_v = \theta_g \qquad [2.3.1.10]$$

und die Wahrscheinlichkeitsdichte ihrer Antwortvariablen in der Form

$$\begin{aligned} f_{vi}(x) &= \text{prob}(X_{vi} = x | \theta_v) \\ &= \text{prob}(X_{vi} = x | v \in g) = p_{gix} \end{aligned} \qquad [2.3.1.11]$$

dargestellt werden kann (Latent-Class-Modelle). Die Wahrscheinlichkeit, dass eine zufällig herausgegriffene Vp v das Item i in der Kategorie x beantwortet, nimmt dann die Form

$$\text{prob}(X_{vi} = x) = \sum_{g=1}^{h} p_{gix}\, p_g \qquad [2.3.1.12]$$

an, worin

$$p_g = \text{prob}(v \in g) \qquad [2.3.1.13]$$

die (relative) Klassengröße der Klasse g bezeichnet.

Saturiertes Modell

Definiert jedes mögliche Antwortmuster eine eigene Klasse von Vpn mit $\theta_g \Leftrightarrow (x_{g1},\ldots,x_{gk})$, so ist für jede Vp v, die der Klasse g angehört

$$f_{vi}(x) = \text{prob}(X_{vi} = x|\theta_v) \qquad [2.3.1.14]$$
$$= \text{prob}(X_{vi} = x|v \in g) = \begin{cases} 1 & \text{für } x = x_{gi} \\ 0 & \text{sonst} \end{cases}.$$

Da die Antwortmuster der Vpn direkt beobachtet werden können, ist die Personenvariable Θ in diesem Fall nicht mehr latent, sondern manifest (saturiertes Modell).

Klassische Testtheorie

Da die Antwortvariablen nur endliche Werte annehmen können, sind ihre ersten beiden Momente

$$E(X_{vi}) = \sum_{x=0}^{m_i-1} x\ f_{vi}(x) \qquad [2.3.1.15]$$

und

$$E(X_{vi}^2) = \sum_{x=0}^{m_i-1} x^2\ f_{vi}(x) \qquad [2.3.1.16]$$

stets existent. Unabhängig davon, welche Modellspezifikationen getroffen werden, besitzen die Antwortvariablen daher stets einen endlichen Erwartungswert (vgl. Gleichung [2.3.1.15]) und endliche Varianz

$$\sigma^2(X_{vi}) = E(X_{vi}^2) - E(X_{vi})^2\ . \qquad [2.3.1.17]$$

Werden aus den Itemantworten Summenscores

$$x_{vt} = \sum_{i=1}^{k} x_{vi} \qquad [2.3.1.18]$$

gebildet, dann folgen die Modellannahmen von Novick und damit auch die Axiome von Gulliksen somit unmittelbar aus den Annahmen des allgemeinen testtheoretischen Modells: Jeder Vp v und jedem Test t entspricht eine Zufallsvariable X_{vt} mit endlichem Erwartungswert

$$E(X_{vt}) = \sum_{i=1}^{k} E(X_{vi}) = \sum_{i=1}^{k} \sum_{x=0}^{m_i-1} x\ f_{vi}(x) \qquad [2.3.1.19]$$

und endlicher Varianz

$$\sigma^2(X_{vt}) = \sum_{i=1}^{k} \sigma^2(X_{vi}). \qquad [2.3.1.20]$$

Den Erwartungswert der Scorevariable $E(X_{vt}) = \tau_{vt}$ bezeichnen wir in Übereinstimmung mit Novick wieder als den True-Score der Vp v in Test t.

Lediglich das saturierte Modell nimmt insofern eine Sonderstellung ein, als jeder Vp v bei gegebener Klassenzugehörigkeit $v \in g$ ein fester Score

$$x_{gt} = \sum_{i=1}^{k} x_{gi} \qquad [2.3.1.21]$$

entspricht, der keiner Zufallsvariation unterliegt, so dass die Maßzahlvarianz gleich Null ist und der True-Score mit dem tatsächlich beobachteten Score übereinstimmt.

2.3.2 Rechtfertigung der Scorebildung

Eigenschaften des True-Scores

Aufgrund der in Gleichung [2.3.1.19] dargestellten Beziehung nimmt der True-Score unter den verschiedenen Spezifikationen des allgemeinen testtheoretischen Modells die im Folgenden dargestellte Form an:

$$\tau_{vt} = k\, E(X_{\bullet\bullet}) = k \sum_{x=0}^{m_i-1} x\, p_{\bullet\bullet x} \qquad [2.3.2.1]$$

im Falle des Pure-Random-Modells (Gleichungen [2.3.1.2–2.3.1.2]) sowie

$$\tau_{vt} = \sum_{i=1}^{k} E(X_{\bullet i}) = \sum_{i=1}^{k} \sum_{x=0}^{m_i-1} x\, p_{\bullet ix} \qquad [2.3.2.2]$$

im Falle der Ein-Klassen-Lösung der LCA (identische Personen, Gleichungen [2.3.1.4–2.3.1.5]). In diesen beiden Fällen enthält also auch der Score keinerlei diagnostische Informationen über die Vpn.

Im Falle personenabhängiger, aber itemunabhängiger Antwortvariablen (identische Items, Gleichungen [2.3.1.8–2.3.1.9]) hat der True-Score die Form

$$\tau_{vt} = \sum_{i=1}^{k} E(X_{v\bullet}) = k\,E(X_{v\bullet}) = k\sum_{x=0}^{m-1} x\,p_{v\bullet x}\ , \qquad [2.3.2.3]$$

die sich für binäre Items (Binomialmodell, Gleichung [2.2.3.8]) auf

$$\tau_{vt} = k\ p_{v\bullet} \qquad [2.3.2.4]$$

reduziert.

Sind die Antwortvariablen sowohl personen- als auch itemabhängig, so hat der True-Score die Form

$$\tau_{vt} = \sum_{i=1}^{k}\sum_{x=0}^{m_i-1} x\ f_i(x, \theta_v) \qquad [2.3.2.5]$$

die sich für Latent-Trait-Modelle mit dichotomen Items auf

$$\tau_{vt} = \sum_{i=1}^{k} f_i(\theta_v) \qquad [2.3.2.6]$$

reduziert und im Falle des Rasch-Modells somit die Form

$$\tau_{vt} = \sum_{i=1}^{k} \frac{e^{\theta_v - \delta_i}}{1 + e^{\theta_v - \delta_i}} \qquad [2.3.2.7]$$

annimmt.

Im Falle von Latent-Class-Modellen hat der True-Score für alle Vpn v, die derselben Klasse g angehören, denselben Wert

$$\tau_{vt} = \tau_{gt} = \sum_{i=1}^{k}\sum_{x=0}^{m_i-1} x\ p_{gix} \qquad [2.3.2.8]$$

wobei sich Gleichung [2.3.2.8] im Falle dichotomer Items zu

$$\tau_{vt} = \tau_{gt} = \sum_{i=1}^{k} p_{gi} \qquad [2.3.2.9]$$

mit $p_{gi} = \text{prob}(X_{vi} = 1|v \in g)$ vereinfacht.

Im saturierten Modell unterliegt der Score bei gegebener Klassenzugehörigkeit keinerlei Zufallsvariation, so dass der True-Score mit dem beobachteten Score übereinstimmt

$$\tau_{vt} = \tau_{gt} = x_{gt} = \sum_{i=1}^{k} x_{gi} \; . \qquad [2.3.2.10]$$

Rechtfertigung der Scorebildung

Wie sehr es sich in der Praxis auch eingebürgert haben mag, die Testleistung eines Probanden durch seinen Summenscore auszudrücken, kann dennoch nicht *a priori* davon ausgegangen werden, dass eine Unterscheidung der Probanden aufgrund ihrer Scores auch diagnostisch sinnvoll bzw. zweckmäßig ist. Die diagnostische Brauchbarkeit des Summenscores kann aus verschiedenen Gründen eingeschränkt sein.

Personenunabhängkeit

Als gänzlich unbrauchbar erweist sich der Score u. a. dann, wenn die Antwortvariablen personenunabhängig verteilt sind, wie dies im Pure-Random-Modell und/oder bei Geltung der Ein-Klassen-Lösung der LCA (identische Person) der Fall ist. Aufgrund der Zufallsvariation der Scorevariablen werden sich die Scores der Vpn zwar auch in diesem Fall *de facto* voneinander unterscheiden. Die Scoreunterschiede sind jedoch ausschließlich auf Messfehler zurückzuführen und haben keinerlei diagnostische Aussagefähigkeit.

Saturiertes Modell

Im saturierten Modell erweist sich der Score aus geradezu entgegengesetzten Gründen als unbrauchbar. Hier hat jedes mögliche Antwortmuster $x_v = (x_{v1},\ldots,x_{vk})$ seine eigene diagnostische Bedeutung, und die Unterschiede zwischen Vpn, deren Antwortmuster den selben Score ergibt, werden durch die Scorebildung schlichtweg verwischt.

Latent-Class-Modelle

Bei den Latent-Class-Modellen hängt die Brauchbarkeit des Scores davon ab, ob sich die latenten Klassen quantitativ oder qualitativ voneinander unterscheiden. Aus Gründen der mathematischen Einfachheit diskutieren wir dies nur für dichotome Items.

Von quantitativ verschiedenen Klassen können wir in diesem Fall dann sprechen, wenn sich die latenten Klassen aufgrund ihrer klassenspezifischen Itemlösungswahrscheinlichkeiten

$$p_{gi} = \text{prob}\left(X_{vi} = 1 \middle| v \in g\right) \qquad [2.3.2.11]$$

in eine Rangreihe bringen lassen, so dass jedes Item i dieselbe Ordnungsrelation

$$\theta_g > \theta_q \Leftrightarrow p_{gi} > p_{qi} \qquad \text{für } i = 1,\ldots,k \qquad [2.3.2.12]$$

zwischen je zwei Klassen g und q definiert und die latente Personenvariable somit auf einer Ordinalskala gemessen wird.

Man kann dies graphisch darstellen, indem man die Items nach ihrer durchschnittlichen Schwierigkeit in eine Rangreihe bringt und die Itemranglätze (r_i) auf der Abszisse eines Koordinatensystems aufträgt. Je seltener ein Item gelöst wird, desto höher ist der Schwierigkeitsrangplatz des Items. Die klassenspezifischen Lösungswahrscheinlichkeiten (p_{gi}) der Items werden dann auf der Ordinate aufgetragen und miteinander verbunden, so dass wir für jede Klasse eine Profillinie erhalten.

Ist die in Gleichung [2.3.2.12] dargestellte Voraussetzung erfüllt, so sind die Profillinien überschneidungsfrei (vgl. Abb. 2.3.2.1). Die Items messen dann alle dieselbe (quantitative) Eigenschaft der Vpn. Der Test ist homogen.

Abb. 2.3.2.1: Profillinien der Latent-Class-Analyse: Quantitativ verschiedene Klassen

Überschneiden sich die Profillinien (vgl. Abb. 2.3.2.2), so ist die in Gleichung [2.3.2.12] dargestellte Voraussetzung verletzt und die latente Personenvariable wird lediglich auf einer Nominalskala gemessen. Die latenten Klassen sind somit qualitativ verschieden und die Items messen nicht alle dieselbe quantitative Eigenschaft der Vpn. Der Test ist inhomogen.

Unterscheidet man die Vpn nur aufgrund ihres Scores, so geht damit zumindest ein Teil der statistischen Information verloren, welche der Test bezüglich der latenten Personeneigenschaft lie-

fert. Im Extremfall kann dies so weit gehen, dass der Score überhaupt nicht mehr zwischen den Klassen zu differenzieren vermag.

Abb. 2.3.2.2: Profillinien der Latent-Class-Analyse: Qualitativ verschiedene Klassen

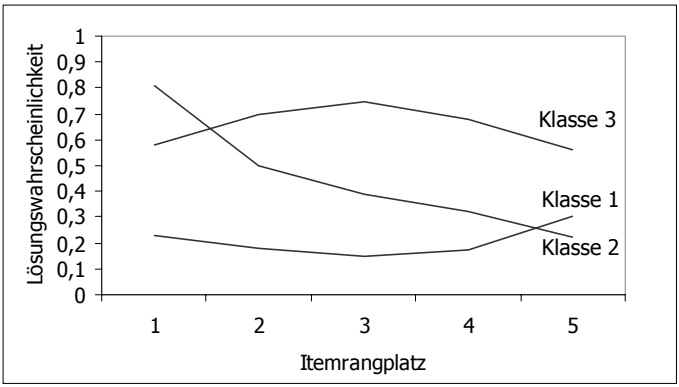
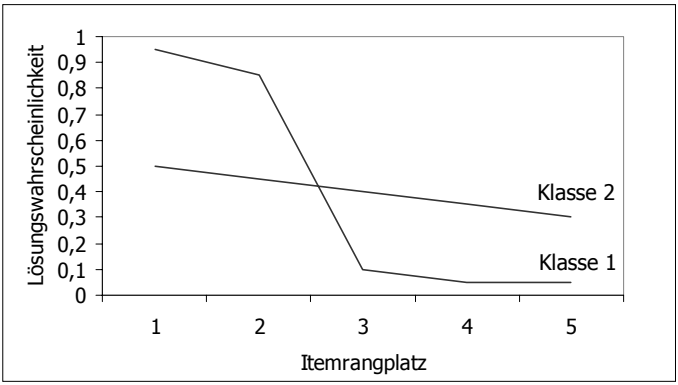


Abb. 2.3.2.3: Profillinien der Latent-Class-Analyse: Rein qualitativ verschiedene Klassen



So haben z. B. die beiden Klassen in Abb. 2.3.2.3 exakt denselben True-Score und unterscheiden sich ausschließlich dadurch, dass die Vpn der Klasse 2 bei allen Items eine mittlere Erfolgschance besitzen, während die Vpn der Klasse 1 bei den Items 3-5 nur eine sehr geringe, bei den Items 1 und 2 dagegen eine sehr hohe Lösungswahrscheinlichkeit haben. Um die beiden Klassen voneinander unterscheiden zu können, kommt es also nicht auf den Score der Vpn, sondern ausschließlich auf ihre Antwortmuster an.

Erschöpfende Statistiken

Selbst wenn alle Items dieselbe Ordnungsrelation

$$\theta_v > \theta_w \Leftrightarrow p_{vi} > p_{wi} \qquad \text{für } i = 1,\ldots,k \qquad [2.3.2.13]$$

zwischen je zwei Vpn v und w definieren, so ist damit noch nicht sichergestellt, dass der Summenscore x_{vo} die gesamte statistische Information bezüglich der latenten Personenvariable ausschöpft. Eine erschöpfende (= suffiziente) Statistik für θ_v ist der Summenscore erst dann, wenn die (bedingte) Wahrscheinlichkeit des Antwortvektors einer Vp bei gegebenem Summenscore nicht mehr von dem Personenparameter abhängt, so dass

$$\text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo}\} = \frac{\text{prob}\{(x_{v1},\ldots,x_{vk})|\theta_v\}}{\text{prob}(x_{vo}|\theta_v)} \qquad [2.3.2.14]$$

eine von θ_v unabhängige Konstante ist. Erst wenn dies der Fall ist, können wir sichergehen, dass das Antwortmuster keine über den Summenscore hinausgehende Information über die Personenvariable enthält.

Binomialmodell

Im Binomialmodell, wo sich die Testitems hinsichtlich ihrer Itemcharakteristik nicht voneinander unterscheiden, ist die Suffizienz des Summencores beinahe trivial. Jeder gegebene Proband v besitzt bei allen Items i = 1,…,k dieselbe Lösungswahrscheinlichkeit $p_{vi} = p_{v\bullet}$, die wir wegen [2.2.3.8] auch in der Form $p_{v\bullet} = \theta_v$ schreiben können.

Für die Wahrscheinlichkeit des Antwortvektors erhalten wir daher

$$\text{prob}\{(x_{v1},\ldots,x_{vk})|\theta_v\} = \prod_{i=1}^{k} p_{vi}^{x_{vi}} (1-p_{vi})^{1-x_{vi}} \qquad [2.3.2.15]$$

$$= \prod_{i=1}^{k} \theta_v^{x_{vi}} (1-\theta_v)^{1-x_{vi}}$$

$$= \theta_v^{x_{vo}} (1-\theta_v)^{k-x_{vo}}$$

und der Summenscore des Probanden folgt einer Binomialverteilung mit der Wahrscheinlichkeitsfunktion

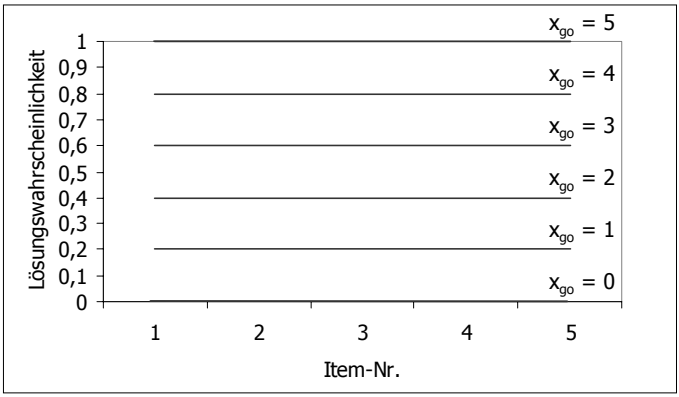
$$\text{prob}\{x_{vo}|\theta_v\} = \binom{k}{x_{vo}} \theta_v^{x_{vo}} (1-\theta_v)^{k-x_{vo}}. \qquad [2.3.2.16]$$

Einsetzen von [2.3.2.15] und [2.3.2.16] in [2.3.2.14] ergibt schließlich den von θ_v unabhängigen Ausdruck

$$\text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo} = g\} = \frac{1}{\binom{k}{g}} = \binom{k}{g}^{-1} \text{ für } g=0,\ldots,k. \qquad [2.3.2.17]$$

Das Binomialmodell teilt die Probanden somit aufgrund ihrer Summenscores in k+1 manifeste Klassen g = 0,…,k ein, die sich ausschließlich in ihrer Leistungsmenge x_{go} voneinander unterscheiden (vgl. Abb. 2.3.2.4). Ist der Summenscore eines Probanden bekannt, so liefert der Antwortvektor des Probanden keinerlei darüber hinausgehende Information über die Ausprägung der latenten Personenvariable.

Abb. 2.3.2.4: Profillinien der Latent-Class-Analyse: Binomialmodell



Rasch-Modell

Unterscheiden sich die Testitems dagegen hinsichtlich ihrer Itemcharakteristiken, so ist die Verwendung des Summenscores zumeist mit einem Informationsverlust bezüglich der latenten Personenvariablen verbunden. Eine erschöpfende Beschreibung der Testleistung eines Probanden vermag der Summenscore in diesem Falle *dann und nur dann* zu leisten (zum Beweis siehe Kap. 2.3.3), wenn die Itemcharakteristiken in Form des Rasch-Modells [2.2.3.11] darstellbar sind, das wir unter Verwendung der Schreibweise exp(x) = e^x auch durch die Gleichungen

$$
\begin{aligned}
f_{ix}(\theta_v) &= \text{prob}(X_{vi} = x) \qquad [2.3.2.18]\\
&= \frac{\lambda_{vi}^x}{1+\lambda_{vi}} \quad \text{mit } \lambda_{vi} = \xi_v \varepsilon_i\\
&= \frac{(\xi_v \varepsilon_i)^x}{1+\xi_v \varepsilon_i} \quad \text{mit } \xi_v = \exp(\theta_v) \text{ und } \varepsilon_i = \exp(-\delta_i)\\
&= \frac{\exp(x(\theta_v - \delta_i))}{1+\exp(\theta_v - \delta_i)} \text{ mit } \theta_v = \ln(\xi_v) \text{ und } \delta_i = -\ln(\varepsilon_i)
\end{aligned}
$$

darstellen können, worin x die Werte 0 und 1 annehmen kann.

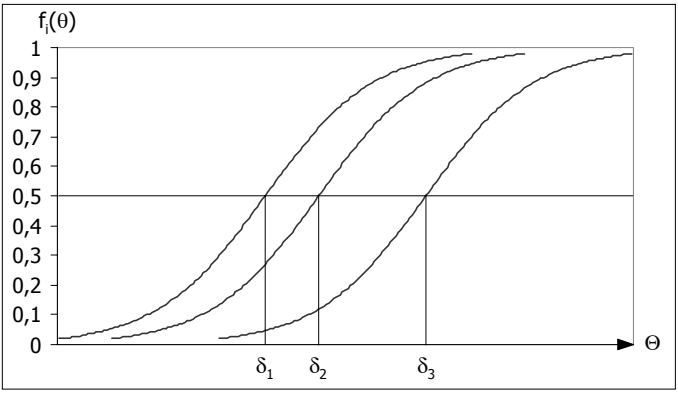


Abb. 2.3.2.5: Itemcharakteristiken des Rasch-Modells als Funktion der logarithmischen Parameter θ_v und δ_i

Abbildung 2.3.2.5 zeigt die Itemcharakteristiken des Rasch-Modells als Funktion der logarithmischen Parameter θ_v und δ_i. Da die Items alle gleich trennscharf sind, verlaufen die Itemcharakteristiken parallel zueinander. Je schwieriger ein Item ist, desto weiter ist seine Itemcharakteristikfunktion nach rechts verschoben.

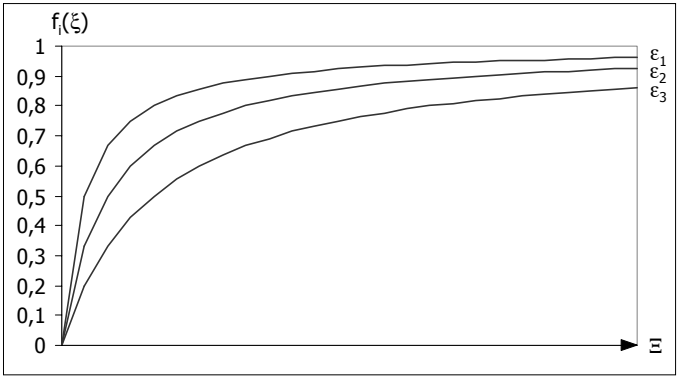


Abb. 2.3.2.6: Itemcharakteristiken des Rasch-Modells als Funktionen der delogarithmierten Parameter ξ_v und ε_i

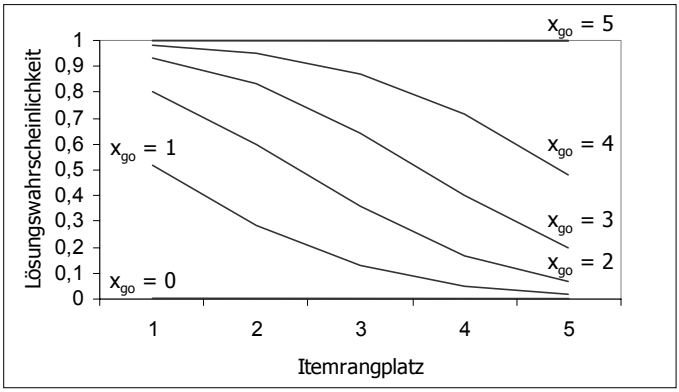
In Abbildung 2.3.2.6 sind die Itemcharakteristiken des Rasch-Modells als Funktion der delogarithmierten Parameter ξ_v und ε_i dargestellt. Der Personenparameter ξ_v kann als Fähigkeitsparameter, der Itemparameter ε_i als Leichtigkeitsparameter interpretiert werden. Je größer ε_i ist, desto rascher steigt die Itemcharakteristikfunktion an.

Ähnlich wie beim Binomialmodell teilt der Summenscore die Probanden wieder in k+1 manifeste Klassen g = 0,…,k ein, die sich

ausschließlich in ihrer Leistungsmenge x_{go} voneinander unterscheiden. Tests, die diese Voraussetzung erfüllen, so dass das Antwortmuster der Probanden keinerlei weitere Information bezüglich der latenten Personenvariable liefert, bezeichnet man als *Rasch-homogen*.

Anders als im Binomialmodell verlaufen die Profillinien hier jedoch nicht parallel zur Abszisse, sondern sind monoton fallende Funktionen (vgl. Abb. 2.3.2.7).

Abb. 2.3.2.7: Profillinien der Latent-Class-Analyse: Rasch-Modell



Gewichtete Summenscores

Über eine ähnliche Eigenschaft verfügt auch das Birnbaum-Modell (Gleichung [2.2.3.10]). Hier ist es jedoch nicht der Summenscore, sondern der gewichtete Summenscore, der die Testleistung der Probanden erschöpfend beschreibt.

Serielle Abhängigkeit

Sind die Antwortvariablen seriell abhängig von der Anzahl der zuvor gelösten Testitems, so leistet der Summenscore (unabhängig von der Präsentationsreihenfolge der Items) *genau dann* eine erschöpfende Beschreibung der Testleistung der Probanden, wenn sich die bedingten Antwortwahrscheinlichkeiten in Form des auf Kempf (1974) zurückgehenden dynamischen Testmodells darstellen lassen (zum Beweis siehe Kap. 2.3.3). Die bedingten Itemcharakteristiken des dynamischen Testmodells sind in Abbildung 2.3.2.8 dargestellt und haben die Form

$$f_{ix.r}(\theta_v) = \text{prob}\left(X_{vi} = x \middle| x_{vo}^{(i)} = r\right) \qquad [2.3.2.19]$$

$$= \frac{\lambda_{vi.r}^{x}}{1+\lambda_{vi.r}} \quad \text{mit } \lambda_{vi.r} = \xi_{v.r}\varepsilon_{i.r}$$

$$= \begin{cases} \dfrac{\xi_v + \psi_r}{\xi_v + \sigma_i} & \text{für } x = 1 \\ \dfrac{\sigma_i - \psi_r}{\xi_v + \sigma_i} & \text{für } x = 0 \end{cases} \quad \text{mit } \xi_{v.r} = \xi_v + \psi_r \text{ und } \varepsilon_{i.r} = \left(\sigma_i - \psi_r\right)^{-1},$$

worin $x_{vo}^{(i)} = \sum_{j=1}^{i-1} x_{vj}$ den partiellen Summenscore bezeichnet, welchen die Vp bei den vorangegangenen Items erzielt hat.

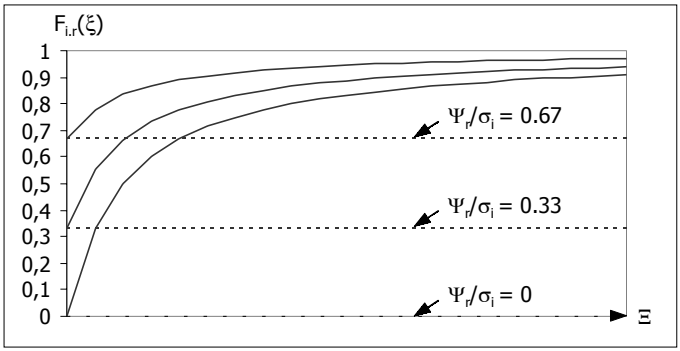


Abb. 2.3.2.8: Bedingte Itemcharakteristiken des dynamischen Testmodells

ξ_v ist ein Fähigkeitsparameter, σ_i beschreibt die Schwierigkeit des Testitems und der Transferparameter ψ_r beschreibt den Lerneffekt, der aus der Lösung von r vorangegangenen Items resultiert.

Interpretation der Modellparameter

Zu einer genaueren Interpretation der Modellparameter gelangen wir, indem wir zwei alternative Prozesse unterscheiden, die zur Aufgabenlösung führen können:

1. trait-bedingte Aufgabenlösung (= aufgrund der Fähigkeitsausstattung der Vp) und
2. transfer-bedingte Aufgabenlösung (= aufgrund des Lerneffekts, der aus der erfolgreichen Bearbeitung von r vorangegangenen Items resultiert).

Folglich gilt für die bedingte Lösungswahrscheinlichkeit

$$p_{vi.r} = p_{vi} + \left(1 - p_{vi}\right) p_{ri}\ , \qquad [2.3.2.20]$$

worin p_{vi} die Wahrscheinlichkeit der trait-bedingten und p_{ri} die Wahrscheinlichkeit der transfer-bedingten Aufgabenlösung bezeichnet.

Die Modellgleichung [2.3.2.19] folgt daraus, indem wir die Wahrscheinlichkeit der trait-bedingten Aufgabenlösung durch ein Rasch-Modell der Form

$$p_{vi} = \frac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i} \qquad [2.3.2.21]$$

beschreiben und für die Wahrscheinlichkeit der transfer-bedingten Aufgabenlösung die Beziehung

$$p_{ri} = \psi_r \varepsilon_i \qquad [2.3.2.22]$$

postulieren. Die Modellstruktur des dynamischen Testmodells [2.3.2.19] ergibt sich daraus, indem wir $\sigma_i = \varepsilon_i^{-1}$ setzen, so dass $\psi_r = p_{ri}\sigma_i$.

Beweis

Einsetzen von [2.3.2.21] und [2.3.2.22] in [2.3.2.20] ergibt

$$p_{vi.r} = \frac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i} + \left(1 - \frac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i}\right)\psi_r \varepsilon_i$$

$$= \frac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i} + \frac{\psi_r \varepsilon_i}{1 + \xi_v \varepsilon_i}$$

und nach Division von Zähler und Nenner durch ε_i erhalten wir

$$p_{vi.r} = \frac{\xi_v}{\varepsilon_i^{-1} + \xi_v} + \frac{\psi_r}{\varepsilon_i^{-1} + \xi_v}$$

$$= \frac{\xi_v + \psi_r}{\xi_v + \sigma_i}$$

Skaleneigenschaft der Modellparameter

Das dynamische Testmodell unterscheidet sich vom Rasch-Modell somit dadurch, dass neben der Aufgabenlösung aufgrund der Fähigkeitsausstattung der Vp auch die Aufgabenlösung aufgrund eines erfolgskontingenten Lernprozesses zugelassen wird, deren Wahrscheinlichkeit proportional zu dem Lerneffekt ψ_r und umgekehrt proportional zu der Aufgabenschwierigkeit σ_i ist.

Wegen $\psi_r = p_{ri}\sigma_i$ und $0 \le p_{ri} \le 1$ ist also $0 \le \psi_r \le \sigma_i$. Die Modellparameter sind bis auf Multiplikation mit einer positiven Konstanten bestimmt und werden auf einer Rationalskala gemessen.

Sind die Transferparameter alle gleich Null, so ist $p_{ri} = 0$ und das Rasch-Modell ergibt sich als Spezialfall des dynamischen Testmodells mit $\delta_i = \ln(\sigma_i)$.

Werden die Antwortvariablen auf einer polytomen Nominalskala gemessen, so entspricht jeder Vp ein m-dimensionaler Scorevektor

Polytome Antwortvariablen

$$n_{vo} = (n_{vo0}, \ldots, n_{vom-1}), \qquad [2.3.2.23]$$

worin n_{vox} die Anzahl der Testitems bezeichnet, welche die Vp v in Kategorie x beantwortet hat. Dieser Scorevektor leistet genau dann eine erschöpfende Beschreibung der Testleistung des Probanden, wenn die Wahrscheinlichkeitsdichte der Antwortvariablen für x = 0,…,m-1 in Form des auf Rasch (1961) zurückgehenden und von Andersen (1973) ausgearbeiteten polytomen Rasch-Modells

$$f_{ix}(\theta_v) = \text{prob}(X_{vi} = x) \qquad [2.3.2.24]$$

$$= \frac{\lambda_{vix}}{\sum_{y=0}^{m-1} \lambda_{viy}} \quad \text{mit } \lambda_{vix} = \xi_{vx}\varepsilon_{ix}$$

$$= \frac{\xi_{vx}\varepsilon_{ix}}{\sum_{y=0}^{m-1} \xi_{vy}\varepsilon_{iy}} \quad \text{mit } \xi_{vx} = \exp(\theta_{vx}) \text{ und } \varepsilon_{ix} = \exp(-\delta_{ix})$$

$$= \frac{\exp(\theta_{vx} - \delta_{ix})}{\sum_{y=o}^{m-1} \exp(\theta_{vy} - \delta_{iy})} \quad \text{mit } \theta_{vx} = \ln(\xi_{vx}) \text{ und } \delta_{ix} = -\ln(\varepsilon_{ix})$$

dargestellt werden kann. Dabei werden die Fähigkeit der Vp und die Schwierigkeit des Items durch m-dimensionale Vektorvariablen

$$\theta_v = (\theta_{v0}, \ldots, \theta_{vm-1}) \quad \text{und} \quad \delta_i = (\delta_{i0}, \ldots, \delta_{im-1}) \qquad [2.3.2.25]$$

bzw.

$$\xi_v = (\xi_{v0}, \ldots, \xi_{vm-1}) \quad \text{und} \quad \varepsilon_i = (\varepsilon_{i0}, \ldots, \varepsilon_{im-1}) \qquad [2.3.2.26]$$

beschrieben. Wegen

$$\sum_{x=0}^{m-1} f_{ix}(\theta_v) = 1 \qquad [2.3.2.27]$$

sind allerdings nur jeweils m-1 Komponenten von θ_v bzw. δ_i unabhängig voneinander.

Für m = 2 ergibt sich mit den Definitionen $\theta_v = \theta_{v1} - \theta_{v0}$ und $\delta_i = \delta_{i1} - \delta_{i0}$ das dichotome Rasch-Modell (Gleichung [2.3.2.18]) als Spezialfall des polytomen Modells.

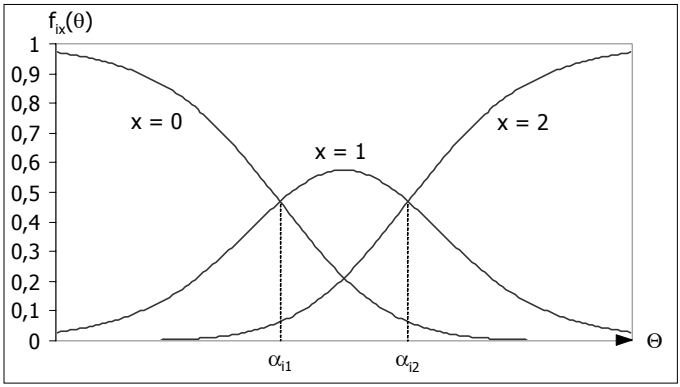
Ordinale Antwortvariablen

Werden die Antwortvariablen auf einer Ordinalskala gemessen, so können die θ_{vx} in der Form $\theta_{vx} = x\theta_v$ dargestellt werden und Gleichung [2.3.2.24] reduziert sich auf die Modellgleichung des ordinalen Rasch-Modells

$$f_{ix}(\theta_v) = \frac{\exp(x\theta_v - \delta_{ix})}{\sum_{y=0}^{m-1} \exp(y\theta_v - \delta_{iy})} \quad \text{mit mit } \delta_{i0} = 0, \qquad [2.3.2.28]$$

das auf Andrich (1978a,b) zurückgeht und in der Literatur auch als *Partial-Credit-Modell* bezeichnet wird.

Abb. 2.3.2.9: Itemfunktionen des ordinalen Rasch-Modells



Wie Andrich zeigt, resultiert das ordinale Rasch-Modell aus der Geltung des dichotomen Rasch-Modells für die sogenannten Schwellenwahrscheinlichkeiten

$$\text{prob}\{X_{vi} = x | x-1 \le X_{vi} \le x\} = \frac{\exp(\theta_v - \alpha_{ix})}{1 + \exp(\theta_v - \alpha_{ix})} \qquad [2.3.2.30]$$

für x = 1,…,m-1 ,

worin die Schwellenparameter α_{ix} die Schwierigkeit beschreiben, bei Item i die Schwelle von x-1 zu x Antwortpunkten zu überschreiten. Die Summe der Schwierigkeiten aller Schwellen, die man überschreiten muss, um in Kategorie x zu antworten, ergibt schließlich die Kategorienschwierigkeit

$$\delta_{ix} = \sum_{j=1}^{x} \alpha_{ij} \ . \qquad [2.3.2.31]$$

Beweis

Einsetzen von [2.3.2.28] in

$$\text{prob}\{X_{vi} = x | x-1 \le X_{vi} \le x\} = \frac{\text{prob}\{X_{vi} = x\}}{\text{prob}\{X_{vi} = x-1\} + \text{prob}\{X_{vi} = x\}}$$

ergibt

$$\frac{\dfrac{\exp(x\theta_v - \delta_{ix})}{\sum_{y=0}^{m-1}\exp(y\theta_v - \delta_{iy})}}{\dfrac{\exp((x-1)\theta_v - \delta_{ix-1})}{\sum_{y=0}^{m-1}\exp(y\theta_v - \delta_{iy})} + \dfrac{\exp(x\theta_v - \delta_{ix})}{\sum_{y=0}^{m-1}\exp(y\theta_v - \delta_{iy})}}$$

$$= \frac{\exp(\theta_v - \delta_{ix})}{\exp(-\delta_{ix-1}) + \exp(\theta_v - \delta_{ix})}$$

$$= \frac{\exp(\theta_v - (\delta_{ix} - \delta_{ix-1}))}{1 + \exp(\theta_v - (\delta_{ix} - \delta_{ix-1}))}$$

$$= \frac{\exp(\theta_v - \alpha_{ix})}{1 + \exp(\theta_v - \alpha_{ix})},$$

worin $\alpha_{ix} = \delta_{ix} - \delta_{ix-1}$, so dass $\delta_{ix} = \delta_{ix-1} + \alpha_{ix}$. Wegen $\delta_{i0} = 0$ folgt schließlich $\delta_{i1} = \alpha_{i1}$, $\delta_{i2} = \alpha_{i1} + \alpha_{i2}$, u.s.w.

Das Poisson-Modell

Eine weitere Modellvariante, die in der Literatur leider viel zu wenig Beachtung findet, ist das bereits von Rasch (1960) für die Analyse von Lesefehlern entwickelte Poisson-Modell, das jedoch nicht nur für diesen Anwendungsfall, sondern z. B. auch für die Analyse von Konzentrationsleistungstests geeignet ist. Allgemein gesprochen, für jede Art von Test, der aus einer Anzahl von k Items besteht, bei deren Bearbeitung eine große Anzahl von Fehlern ($m \to \infty$) unterlaufen kann, während die Wahrscheinlichkeit des einzelnen Fehlers jedoch sehr klein ist ($p \to 0$).

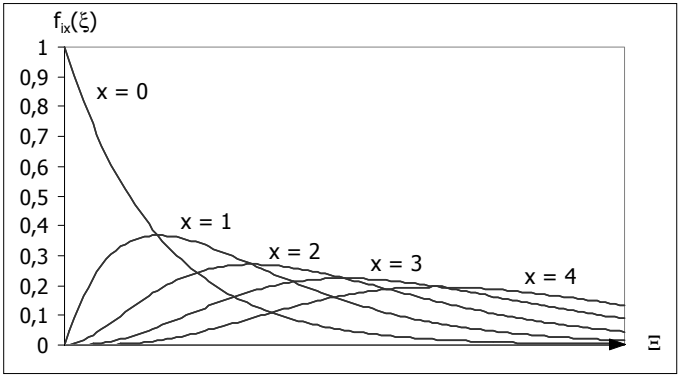
Die Verteilung der Fehleranzahl X_{vi} wird dabei durch die Wahrscheinlichkeitsfunktion einer Poisson-Verteilung mit dem Parameter $\lambda_{vi} = \xi_v \varepsilon_i$ beschrieben, so dass für x = 0,1,2,…

$$f_{ix}(\xi_v) = \text{prob}(X_{vi} = x) \qquad [2.3.2.32]$$

$$= \frac{\lambda_{vi}^x}{x!} e^{-\lambda_{vi}} \quad \text{mit } \lambda_{vi} = \xi_v \varepsilon_i$$

$$= \frac{(\xi_v \varepsilon_i)^x}{x!} e^{-\xi_v \varepsilon_i}.$$

Abb. 2.3.2.10: Itemfunktionen des Poisson-Modells



Auch im Poisson-Modell sind die Schwellenwahrscheinlichkeiten wieder durch ein dichotomes Rasch-Modell der Form

$$\text{prob}\{X_{vi} = x | x-1 \le X_{vi} \le x\} = \frac{\xi_v \alpha_{ix}}{1+\xi_v \alpha_{ix}} \qquad [2.3.2.33]$$

mit den Schwellenparametern

$$\alpha_{ix} = \varepsilon_i \, x^{-1} \qquad [2.3.2.34]$$

darstellbar. Der Beweis sei dem Leser als Übungsaufgabe überlassen.

2.3.3 Notwendige und hinreichende Bedingungen für die Suffizienz des Summenscores

Suffizienz des Summenscores bei lokal unabhängigen Antwortvariablen

Satz 2.3.3.1: Es sei I eine endliche oder abzählbar unendliche Menge von binären Testitems, aus der k Items ausgewählt und in der Reihenfolge i = 1,...k präsentiert werden. Die Antwortvariablen der Items seien lokal unabhängig verteilt. Ist dies der Fall, so leistet der Summenscore *genau dann* für jede beliebige Itemauswahl und Präsentationsreihenfolge eine suffiziente Beschreibung der Testleistung der Probanden, wenn die Antwortwahrscheinlichkeiten in Form des Rasch-Modells [2.3.2.18] darstellbar sind.

Hinreichende Bedingungen

Sind die Modellannahmen des Rasch-Modells erfüllt, so ist die bedingte Wahrscheinlichkeit des Antwortvektors bei gegebenem Summenscore $x_{vo} = g$ vom Personenparameter θ_v unabhängig und es gilt

$$\text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo}=g\}=\frac{\prod_{i=1}^{k}\varepsilon_i^{x_{vi}}}{\gamma_g(\varepsilon)} \quad \text{für } g = 0,\ldots,k. \qquad [2.3.3.1]$$

Darin bezeichnet $\gamma_g(\varepsilon)$ die symmetrische Grundfunktion g-ter Ordnung der Leichtigkeitsparameter $\varepsilon_1,\ldots,\varepsilon_k$

$$\gamma_0(\varepsilon)=1 \qquad [2.3.3.2]$$
$$\gamma_1(\varepsilon)=\varepsilon_1+\ldots+\varepsilon_k$$
$$\gamma_2(\varepsilon)=\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\ldots+\varepsilon_1\varepsilon_k+\varepsilon_2\varepsilon_3+\ldots+\varepsilon_{k-1}\varepsilon_k$$
$$\gamma_3(\varepsilon)=\varepsilon_1\varepsilon_2\varepsilon_3+\varepsilon_1\varepsilon_2\varepsilon_4+\ldots+\varepsilon_1\varepsilon_2\varepsilon_k+\varepsilon_1\varepsilon_3\varepsilon_4+\ldots+\varepsilon_{k-2}\varepsilon_{k-1}\varepsilon_k$$
$$\ldots\ldots$$
$$\gamma_k(\varepsilon)=\varepsilon_1\varepsilon_2\varepsilon_3\ldots\varepsilon_k,$$

die beginnend mit g = 0, $\gamma_o(\varepsilon)=1$ und $\gamma_o^{(i)}(\varepsilon)=1$ mittels der Rekursionsformeln

$$\gamma_{g+1}(\varepsilon)=\frac{1}{g+1}\sum_{i=1}^{k}\varepsilon_i\gamma_g^{(i)}(\varepsilon) \qquad [2.3.3.2a]$$

und

$$\gamma_{g+1}^{(i)}(\varepsilon)=\gamma_{g+1}(\varepsilon)-\varepsilon_i\gamma_g^{(i)}(\varepsilon) \qquad [2.3.3.2b]$$

berechnet werden können. $\gamma_g^{(i)}(\varepsilon)$ stellt die symmetrische Grundfunktion g-ter Ordnung der Parameter $\varepsilon_1,\ldots,\varepsilon_{i-1},\varepsilon_{i+1},\ldots,\varepsilon_k$ dar.

Beweis

Wegen

$$\text{prob}(x_{vi}|\theta_v)=\frac{(\xi_v\varepsilon_i)^{x_{vi}}}{1+\xi_v\varepsilon_i}$$

kann die Wahrscheinlichkeit des Antwortvektors $x_v = (x_{v1},\ldots,x_{vk})$ für $x_{vo} = g$ in der Form

$$\text{prob}\{(x_{v1},\ldots,x_{vk})|\theta_v\}=\prod_{i=1}^{k}\text{prob}(x_{vi}|\theta_v) \qquad [2.3.3.3]$$

$$=\prod_{i=1}^{k}\frac{(\xi_v\varepsilon_i)^{x_{vi}}}{1+\xi_v\varepsilon_i}=\frac{\xi_v^{x_{vo}}\prod_{i=1}^{k}\varepsilon_i^{x_{vi}}}{\prod_{i=1}^{k}(1+\xi_v\varepsilon_i)}$$

$$= \frac{\xi_v^g}{\prod_{i=1}^{k}(1+\xi_v\varepsilon_i)} \prod_{i=1}^{k} \varepsilon_i^{x_{vi}}$$

geschrieben werden und wir erhalten die Wahrscheinlichkeit des Summenscores $x_{vo} = g$, indem wir [2.3.3.7] über alle möglichen Antwortvektoren $y_v = (y_{v1},\ldots,y_{vk})$ aufsummieren, deren Summenscore $y_{vo} = g$ ergibt:

$$\text{prob}(x_{vo} = g|\theta_v) = \sum_{y_v:y_{vo}=g} \text{prob}\{(y_{v1},\ldots,y_{vk})|\theta_v\} \qquad [2.3.3.4]$$

$$= \sum_{y_v:y_{vo}=g} \frac{\xi_v^g}{\prod_{i=1}^{k}(1+\xi_v\varepsilon_i)} \prod_{i=1}^{k} \varepsilon_i^{y_{vi}}$$

$$= \frac{\xi_v^g}{\prod_{i=1}^{k}(1+\xi_v\varepsilon_i)} \sum_{y_v:y_{vo}=g} \prod_{i=1}^{k} \varepsilon_i^{y_{vi}}$$

$$= \frac{\xi_v^g}{\prod_{i=1}^{k}(1+\xi_v\varepsilon_i)} \gamma_g(\varepsilon)$$

Einsetzen von [2.3.3.3] und [2.3.3.4] in

$$\text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo} = g\} = \frac{\text{prob}\{(x_{v1},\ldots,x_{vk})|\theta_v\}}{\text{prob}(x_{vo} = g|\theta_v)} \qquad [2.3.3.5]$$

ergibt schließlich

$$\text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo} = g\} = \frac{\prod_{i=1}^{k} \varepsilon_i^{x_{vi}}}{\gamma_g(\varepsilon)} \qquad [2.3.3.6]$$

Ähnlich wie im Binomialmodell – das, wie wir gesehen haben, einen Spezialfall des Rasch-Modells für $\varepsilon_1 = \ldots = \varepsilon_k = 1$ darstellt (vgl. Kap. 2.2.3) – ist die bedingte Wahrscheinlichkeit des Antwortvektors daher vom Personenparameter unabhängig. Anders als im Binomialmodell ist sie jedoch nicht nur eine Funktion der Itemanzahl k und des Summenscores $x_{vo} = g$, sondern auch eine Funktion der Itemparameter.

Notwendige Bedingungen

Wie wir im Folgenden zeigen werden, ist die Suffizienz des Summenscores aber nicht nur de facto im Rasch-Modell gewährleistet, sondern das Rasch-Modell ist bei lokaler stochastischer Unabhängigkeit der Testitems das *einzig mögliche* Testmodell, in welchem der Summenscore die Testleistung der Probanden er-

schöpfend beschreibt. Bei jeder anderen Funktionsform der Itemcharakteristiken geht die Verwendung des Summenscores mit einem Informationsverlust einher.

Ist die Wahrscheinlichkeit des Antwortvektors bei gegebenem Summenscore $x_{vo} = g$ vom Personenparameter θ_v unabhängig, dann sind die Itemcharakteristiken stets in der Form

$$f_i(\theta_v) = \frac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i} \quad \text{für } i = 1,\ldots,k$$

darstellbar.

Beweis

Zum Beweis bedienen wir uns der Methode der vollständigen Induktion (vgl. Bd. 1, Kap. 1.3.4):

1. Zunächst zeigen wir, dass die Behauptung für k = 2 Items gilt.
2. Sodann zeigen wir: Gilt die Behauptung für k Items, dann gilt sie auch für k+1 Items.

Ad 1: Gegeben seien k = 2 lokal unabhängige dichotome Testitems i = 1,2 mit Lösungswahrscheinlichkeiten $0 < p_{vi} < 1$. Weiters definieren wir

$$\xi_v = \frac{p_{v1}}{1 - p_{v1}}, \qquad [2.3.3.7]$$

so dass die Itemcharakteristik des ersten Items o.B.d.A. in Form des Rasch-Modells mit $\varepsilon_1 = 1$ dargestellt werden kann:

$$f_1 = p_{v1} = \frac{\xi_v}{1 + \xi_v} \qquad [2.3.3.8]$$

Voraussetzungsgemäß sei die bedingte Wahrscheinlichkeit des Antwortvektors

$$\text{prob}\{(x_{v1}, x_{v2}) | x_{vo}\} = \frac{\text{prob}\{(x_{v1}, x_{v2}) | \theta_v\}}{\text{prob}(x_{vo} | \theta_v)} \qquad [2.3.3.9]$$

von θ_v unabhängig. Wir zeigen, dass dann auch die Itemcharakteristik des zweiten Items dem Rasch-Modell folgt und in der Form

$$f_2 = p_{v2} = \frac{\xi_v \varepsilon_2}{1 + \xi_v \varepsilon_2} \qquad [2.3.3.10]$$

dargestellt werden kann.

Für k = 2 dichotome Items gibt es genau vier mögliche Antwortvektoren

x_{v1}	x_{v2}	x_{vo}
0	0	0
0	1	1
1	0	1
1	1	2

Für $x_{vo} = 0$ und $x_{vo} = 2$ ist die bedingte Wahrscheinlichkeit des Antwortvektors gleich 1, da es in diesen beiden Fällen nur jeweils einen Antwortvektor gibt, der mit dem Summenscore verträglich ist. Mit $x_{vo} = 1$ sind dagegen zwei verschiedene Antwortvektoren verträglich. Ist die Suffizienz des Summenscores erfüllt, dann dürfen die bedingten Wahrscheinlichkeiten

$$\text{prob}\{(0,1)|x_{vo} = 1\} = \frac{\text{prob}\{(0,1)|\theta_v\}}{\text{prob}\{(1,0)|\theta_v\} + \text{prob}\{(0,1)|\theta_v\}}$$

und

$$\text{prob}\{(1,0)|x_{vo} = 1\} = \frac{\text{prob}\{(1,0)|\theta_v\}}{\text{prob}\{(1,0)|\theta_v\} + \text{prob}\{(0,1)|\theta_v\}}$$

nicht von θ_v abhängen, so dass auch der Ausdruck

$$\frac{\text{prob}\{(0,1)\}}{\text{prob}\{(1,0)\}} = \frac{(1-f_1)f_2}{f_1(1-f_2)} = c$$

eine von θ_v unabhängige Konstante ist. Einsetzen für f_1 ergibt

$$\frac{\frac{1}{1+\xi_v} f_2}{\frac{\xi_v}{1+\xi_v}(1-f_2)} = \frac{f_2}{\xi_v(1-f_2)} = c\ ,$$

so dass

$$f_2 = c\xi_v(1-f_2)$$

$$f_2 = c\xi_v - c\xi_v f_2$$

$$f_2 + c\xi_v f_2 = c\xi_v$$

$$f_2(1 + c\xi_v) = c\xi_v$$

und daher

$$f_2 = \frac{c\xi_v}{1+c\xi_v}\ .$$

Mit der Definition $\varepsilon_2 = c$ erhalten wir schließlich

$$f_2 = \frac{\xi_v \varepsilon_2}{1 + \xi_v \varepsilon_2}.$$

Ad 2: Gegeben seien k+1 lokal unabhängige dichotome Testitems i = 1,…,k+1 mit Lösungswahrscheinlichkeiten $0 < p_{vi} < 1$. Die Itemcharakteristiken der ersten k Items mögen dem Rasch-Modell folgen, so dass sie in der Form

$$f_i = \frac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i} \quad \text{für } i = 1,\ldots, k \qquad [2.3.3.11]$$

geschrieben werden können und der Summenscore

$$x_{vo} = \sum_{i=1}^{k} x_{vi} \qquad [2.3.3.12]$$

eine erschöpfende Statistik für den Personenparameter θ_v darstellt. Voraussetzungsgemäß sei auch der erweiterte Summenscore

$$x_{v*} = \sum_{i=1}^{k+1} x_{vi} = x_{vo} + x_{vk+1} \qquad [2.3.3.13]$$

eine erschöpfende Statistik für θ_v, so dass die bedingte Wahrscheinlichkeit des erweiterten Antwortvektors $(x_{v1},\ldots,x_{vk}, x_{vk+1})$

$$\text{prob}\left\{(x_{v1},\ldots,x_{vk},x_{vk+1}) \middle| x_{v*}\right\} = \frac{\text{prob}\left\{(x_{v1},\ldots,x_{vk},x_{vk+1}) \middle| \theta_v\right\}}{\text{prob}\left\{x_{v*} \middle| \theta_v\right\}} = c \qquad [2.3.3.14]$$

von θ_v unabhängig ist. Wir zeigen, dass dann auch die Itemcharakteristik des Items k+1 dem Rasch-Modell folgt und in der Form

$$f_{k+1} = p_{vk+1} = \frac{\xi_v \varepsilon_{k+1}}{1 + \xi_v \varepsilon_{k+1}} \qquad [2.3.3.15]$$

dargestellt werden kann.

O.B.d.A. betrachten wir den erweiterten Antwortvektor $(x_{v1},\ldots,x_{vk}, 1)$. Dann gilt wegen [2.3.3.3] und [2.3.3.4]

$$\text{prob}\left\{(x_{v1},\ldots,x_{vk},1) \middle| x_{v*} = g+1\right\} =$$

$$= \frac{f_{k+1}\,\text{prob}\left((x_{v1},\ldots,x_{vk}) \middle| \theta_v\right)}{f_{k+1}\,\text{prob}\left(x_{vo} = g \middle| \theta_v\right) + (1 - f_{k+1})\,\text{prob}\left(x_{vo} = g+1 \middle| \theta_v\right)}$$

$$= \frac{f_{k+1} \dfrac{\xi_v^g}{\prod_{i=1}^{k}(1 + \xi_v \varepsilon_i)} \prod_{i=1}^{k} \varepsilon_i^{x_{vi}}}{f_{k+1} \dfrac{\xi_v^g}{\prod_{i=1}^{k}(1 + \xi_v \varepsilon_i)} \gamma_g(\varepsilon) + (1 - f_{k+1}) \dfrac{\xi_v^{g+1}}{\prod_{i=1}^{k}(1 + \xi_v \varepsilon_i)} \gamma_{g+1}(\varepsilon)}$$

$$= \frac{f_{k+1}\prod_{i=1}^{k}\varepsilon_i^{x_{vi}}}{f_{k+1}\gamma_g(\varepsilon)+(1-f_{k+1})\xi_v\gamma_{g+1}(\varepsilon)} = c\,,$$

worin $c_1 = \prod_{i=1}^{k}\varepsilon_i^{x_{vi}}$, $c_2 = \gamma_g(\varepsilon)$ und $c_3 = \gamma_{g+1}(\varepsilon)$ ebenfalls von θ_v unabhängig sind, so dass

$$\frac{f_{k+1}c_1}{f_{k+1}c_2+(1-f_{k+1})\xi_v c_3} = c$$

$$= \frac{f_{k+1}c_1}{f_{k+1}c_2+\xi_v c_3 - f_{k+1}\xi_v c_3}\,.$$

Multiplikation der Gleichung mit $f_{k+1}c_2+\xi_v c_3 - f_{k+1}\xi_v c_3$ ergibt

$$f_{k+1}c_1 = f_{k+1}c_2 c + \xi_v c_3 c - f_{k+1}\xi_v c_3 c\,,$$

so dass

$$f_{k+1}(c_1 - c_2 c + \xi_v c_3 c) = \xi_v c_3 c$$

und daher

$$f_{k+1} = \frac{\xi_v c_3 c}{c_1 - c_2 c + \xi_v c_3 c}$$

$$= \frac{\xi_v \dfrac{c_3 c}{c_1 - c_2 c}}{1+\xi_v \dfrac{c_3 c}{c_1 - c_2 c}}\,.$$

Mit der Definition $\varepsilon_{k+1} = \dfrac{c_3 c}{c_1 - c_2 c}$ erhalten wir schließlich

$$f_{k+1} = \frac{\xi_v \varepsilon_{k+1}}{1+\xi_v \varepsilon_{k+1}}\,.$$

Suffizienz des Summenscores bei seriell abhängigen Antwortvariablen

Satz 2.3.3.2: Es sei I eine endliche oder abzählbar unendliche Menge von binären Testitems, aus der k Items ausgewählt und in der Reihenfolge i = 1,…,k präsentiert werden. Die Antwortvariablen der Items seien seriell abhängig von der Anzahl der zuvor gelösten Items verteilt. Dann leistet der Summenscore *genau dann* für jede beliebige Itemauswahl und Präsentationsreihenfolge eine suffiziente Beschreibung der Testleistung der Probanden, wenn die bedingten Antwortwahrscheinlichkeiten in Form des dynamischen Testmodells [2.3.2.19] darstellbar sind.

Hinreichende Bedingungen

Sind die Modellannahmen des dynamischen Testmodells erfüllt, dann ist die bedingte Wahrscheinlichkeit des Antwortvektors bei gegebenem Summenscore $x_{vo} = g$ vom Personenparameter ξ_v unabhängig und es gilt

$$\text{prob}\left\{(x_{v1},\ldots,x_{vk})\middle|x_{vo}=g\right\}=\frac{\prod_{i=1}^{k}\prod_{r=0}^{i-1}(\sigma_i-\psi_r)^{z_{vir}(1-x_{vi})}}{G(g,\psi,\sigma)}\ , \qquad [2.3.3.16]$$

worin

$$G(g,\psi,\sigma)=\sum_{y_v:y_{vo}=g}\prod_{i=1}^{k}\prod_{r=0}^{i-1}(\sigma_i-\psi_r)^{a_{vir}(1-y_{vi})} \qquad [2.3.3.17]$$

und

$$z_{vir}=\begin{cases}1 & \text{wenn } x_{vo}^{(i)}=r\\ 0 & \text{sonst}\end{cases}$$

bzw.

$$a_{vir}=\begin{cases}1 & \text{wenn } y_{vo}^{(i)}=r\\ 0 & \text{sonst}\end{cases}.$$

Beweis

Wegen [2.3.2.19] kann die Wahrscheinlichkeit des Antwortvektors $x_v = (x_{v1},\ldots,x_{vk})$ für $x_{vo} = g$ in der Form

$$\text{prob}\left\{(x_{v1},\ldots,x_{vk})\middle|\xi_v\right\}=\prod_{i=1}^{k}\text{prob}\left(x_{vi}\middle|(x_{v1},\ldots,x_{vi-1})\right) \qquad [2.3.3.18]$$

$$=\frac{\prod_{i=1}^{k}\prod_{r=0}^{i-1}(\xi_v+\psi_r)^{z_{vir}x_{vi}}(\sigma_i-\psi_r)^{z_{vir}(1-x_{vi})}}{\prod_{i=1}^{k}(\xi_v+\sigma_i)}$$

$$=\prod_{r=0}^{g-1}(\xi_v+\psi_r)\prod_{i=1}^{k}\frac{\prod_{r=0}^{i-1}(\sigma_i-\psi_r)^{z_{vir}(1-x_{vi})}}{\xi_v+\sigma_i}$$

angeschrieben werden und wir erhalten die Wahrscheinlichkeit des Summenscores $x_{vo} = g$, indem wir [2.3.3.18] über alle möglichen Antwortvektoren $y_v = (y_{v1},\ldots,y_{vk})$ aufsummieren, deren Summenscore $y_{vo} = g$ ergibt:

$$\text{prob}\left(x_{vo} = g \middle| \xi_v\right) = \sum_{y_v : y_{vo} = g} \text{prob}\left\{\left(y_{v1}, \ldots, y_{vk}\right) \middle| \xi_v\right\} \qquad [2.3.3.19]$$

$$= \prod_{r=0}^{g-1} \left(\xi_v + \psi_r\right) \sum_{y_v : y_{vo} = g} \prod_{i=1}^{k} \frac{\prod_{r=0}^{i-1} \left(\sigma_i - \psi_r\right)^{a_{vir}(1-y_{vi})}}{\xi_v + \sigma_i}$$

Einsetzen von [2.3.3.18] und [2.3.3.19] in

$$\text{prob}\left\{\left(x_{v1}, \ldots, x_{vk}\right) \middle| x_{vo} = g\right\} = \frac{\text{prob}\left\{\left(x_{v1}, \ldots, x_{vk}\right) \middle| \xi_v\right\}}{\text{prob}\left(x_{vo} = g \middle| \xi_v\right)} \qquad [2.3.3.20]$$

ergibt schließlich

$$\text{prob}\left\{\left(x_{v1}, \ldots, x_{vk}\right) \middle| x_{vo} = g\right\} = \frac{\prod_{i=1}^{k} \prod_{r=0}^{i-1} \left(\sigma_i - \psi_r\right)^{z_{vir}(1-x_{vi})}}{\sum_{y_v : y_{vo} = g} \prod_{i=1}^{k} \prod_{r=0}^{i-1} \left(\sigma_i - \psi_r\right)^{a_{vir}(1-y_{vi})}} \; . \qquad [2.3.3.21]$$

Wie Kempf & Hampapa (1977) zeigen, kann der Ausdruck im Nenner der Gleichung [2.3.3.21]

$$\sum_{y_v : y_{vo} = g} \prod_{i=1}^{k} \prod_{r=0}^{i-1} \left(\sigma_i - \psi_r\right)^{a_{vir}(1-y_{vi})} = G\left(g, \psi, \sigma\right)$$

auch in der Form

$$G\left(g, \psi, \sigma\right) = \sum_{m=0}^{k-g} \beta_m\left(g, \psi\right) \gamma_{k-g-m}\left(\sigma\right)\left(-1\right)^m \qquad [2.3.3.22]$$

dargestellt werden, worin $\gamma_{k\text{-}g\text{-}m}(\sigma)$ die symmetrische Grundfunktion (k-g-m)-ter Ordnung der Schwierigkeitsparameter $\sigma_1,\ldots,\sigma_k$ bezeichnet und $\beta_m(g,\psi)$ für m = 0,…,k-g jeweils die Summe aller möglichen Produkte von m Transferparametern aus $\psi_0,\ldots,\psi_g$ *mit* Wiederholungen darstellt:

$$\begin{aligned}
\beta_0\left(g, \psi\right) &= 1 \\
\beta_1\left(g, \psi\right) &= \psi_0 + \psi_1 + \psi_2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + \psi_g \\
\beta_2\left(g, \psi\right) &= \psi_0^2 + \psi_0\psi_1 + \psi_0\psi_2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + \psi_o\psi_g \\
&\quad + \psi_1^2 + \psi_1\psi_2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + \psi_1\psi_g \\
&\quad + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&\quad + \psi_{g-1}^2 + \psi_{g-1}\psi_g \\
&\quad + \psi_g^2
\end{aligned}$$

u.s.w.

Zur Berechnung dieser Funktion kann man sich beginnend mit g = 0 und $\beta_j(0,\psi) = \psi_0^j$ für j = 0,...,k-g der Rekursionsformel

$$\beta_m(g+1,\psi) = \sum_{j=0}^{m} \psi_{g+1}^{m-j}\,\beta_j(g,\psi) \qquad [2.3.3.23]$$

bedienen.

Notwendige Bedingungen

Ähnlich wie im Rasch-Modell – als Spezialfall des dynamischen Testmodells für $\psi_0 = \ldots = \psi_{k-1} = 0$ – ist die bedingte Wahrscheinlichkeit des Antwortvektors vom Personenparameter unabhängig. Anders als im Rasch-Modell ist sie jedoch nicht nur eine Funktion der Itemparameter, sondern auch eine Funktion der Transferparameter.

Wie wir im Folgenden zeigen werden, bleibt die Suffizienz des Summenscores im dynamischen Testmodell aber nicht nur de facto erhalten, sondern das dynamische Testmodell ist bei serieller Abhängigkeit der Antwortvariablen von der Anzahl der zuvor gelösten Items das *einzig mögliche* Testmodell, in welchem der Summenscore die Testleistung der Probanden (unabhängig von der Auswahl der Items und ihrer Präsentationsreihenfolge) erschöpfend beschreibt.

Ist die Wahrscheinlichkeit des Antwortvektors bei gegebenem Summenscore $x_{vo} = g$ vom Personenparameter ξ_v unabhängig, dann sind die bedingten Itemcharakteristiken stets in der Form

$$f_{ix.r}(\theta_v) = \frac{\xi_v + \psi_r}{\xi_v + \sigma_i} \quad \text{für } i = 1,\ldots,k$$

darstellbar.

Beweis

Zum Beweis bedienen wir uns wieder der Methode der vollständigen Induktion:

1. Zunächst zeigen wir, dass die Behauptung für k = 2 Items gilt.
2. Sodann zeigen wir: Gilt die Behauptung für k Items, dann gilt sie auch für k+1 Items.

Ad 1: Gegeben seien k = 2 seriell abhängige dichotome Testitems i = 1,2 mit bedingten Itemcharakteristiken $0 < f_{i.r} < 1$. Weiters definieren wir

$$\xi_v = \frac{f_{1.0}}{1 - f_{1.0}} \qquad [2.3.3.24]$$

und

$$\sigma_2 = \frac{f_{1.0}}{1-f_{1.0}} \frac{1-f_{2.0}}{f_{2.0}} \ , \qquad [2.3.3.25]$$

so dass die bedingten Itemcharakteristiken

$$f_{1.0} = \frac{\xi_v}{\xi_v + 1} \qquad [2.3.3.26]$$

und

$$f_{2.0} = \frac{\xi_v}{\xi_v + \sigma_2} \qquad [2.3.3.27]$$

o.B.d.A. in Form des dynamischen Testmodells mit $\psi_0 = 0$ und $\sigma_1 = 1$ darstellbar sind.[1]
Voraussetzungsgemäß sei die bedingte Wahrscheinlichkeit der Antwortvektoren

$$\text{prob}\{(1,0)|x_{vo} = 1\} = \frac{\text{prob}\{(1,0)|\xi_v\}}{\text{prob}\{(1,0)|\xi_v\} + \text{prob}\{(0,1)|\xi_v\}} \qquad [2.3.3.28]$$

und

$$\text{prob}\{(0,1)|x_{vo} = 1\} = \frac{\text{prob}\{(0,1)|\xi_v\}}{\text{prob}\{(1,0)|\xi_v\} + \text{prob}\{(0,1)|\xi_v\}} \qquad [2.3.3.29]$$

von ξ_v unabhängig, so dass

$$\frac{\text{prob}\{(1,0)|\xi_v\}}{\text{prob}\{(0,1)|\xi_v\}} = \frac{f_{1.0}(1-f_{2.1})}{(1-f_{1.0})f_{2.0}} = c \qquad [2.3.3.30]$$

eine von ξ_v unabhängige Konstante ist.
Wir zeigen, dass dann auch $f_{2.1}$ in der Form

$$f_{2.1} = \frac{\xi_v + \psi_1}{\xi_v + \sigma_2} \qquad [2.3.3.31]$$

dargestellt werden kann.
Einsetzen von [2.3.3.26] und [2.3.3.27] für $f_{1.0}$ und $f_{2.0}$ in [2.3.3.30] ergibt

$$\frac{\frac{\xi_v}{\xi_v+1}(1-f_{2.1})}{\frac{1}{\xi_v+1}\frac{\xi_v}{\xi_v+\sigma_2}} = c$$

$$(\xi_v + \sigma_2)(1-f_{2.1}) = c$$

$$\xi_v - \xi_v f_{2.1} + \sigma_2 - \sigma_2 f_{2.1} = c$$

$$\xi_v + \sigma_2 - c = \xi_v f_{2.1} + \sigma_2 f_{2.1}$$

$$\xi_v + \sigma_2 - c = f_{2.1}\left(\xi_v + \sigma_2\right)$$

$$f_{2.1} = \frac{\xi_v + \sigma_2 - c}{\xi_v + \sigma_2},$$

so dass $f_{2.1}$ mit der Definition

$$\psi_1 = \sigma_2 - c$$

die Form

$$f_{2.1} = \frac{\xi_v + \psi_1}{\xi_v + \sigma_2}$$

annimmt.

Ad 2: Gegeben seien k+1 seriell abhängige dichotome Testitems i = 1,…,k+1 mit bedingten Itemcharakteristiken $0 < f_{i.r} < 1$. Die bedingten Itemcharakteristiken der ersten k Items mögen dem dynamischen Testmodell folgen, so dass sie in der Form

$$f_{i.r} = \frac{\xi_v + \psi_r}{\xi_v + \sigma_i} \quad \text{für } i = 1,\ldots, k;\ r = 0,\ldots,k-1 \qquad [2.3.3.32]$$

geschrieben werden können und der Summenscore x_{vo} eine erschöpfende Statistik für den Personenparameter ξ_v ist.

Wegen der geforderten Unabhängigkeit von der Itemauswahl und der Reihenfolge ihrer Präsentation gelten die Modellgleichungen [2.3.3.32] auch dann, wenn man einige der Items gegen andere austauscht; insbesondere dann, wenn man das k-te Item durch das (k+1)-te ersetzt, so dass [2.3.3.32] auch für i = k+1 erfüllt ist und namentlich $f_{k+1.k-1}$ in der Form

$$f_{k+1.k-1} = \frac{\xi_v + \psi_{k-1}}{\xi_v + \sigma_{k+1}} \qquad [2.3.3.33]$$

dargestellt werden kann.

Voraussetzungsgemäß sei auch der erweiterte Summenscore

$$x_{v*} = \sum_{i=1}^{k+1} x_{vi} = x_{vo} + x_{vk+1}$$

eine erschöpfende Statistik für ξ_v, so dass die bedingte Wahrscheinlichkeit des erweiterten Antwortvektors $(x_{v1},\ldots,x_{vk}, x_{vk+1})$

$$\text{prob}\left\{\left(x_{v1},\ldots,x_{vk}, x_{vk+1}\right)\middle| x_{v*}\right\} = \frac{\text{prob}\left\{\left(x_{v1},\ldots,x_{vk}, x_{vk+1}\right)\middle|\xi_v\right\}}{\text{prob}\left\{x_{v*}\middle|\xi_v\right\}}$$

von ξ_v unabhängig ist.

Wir zeigen, dass dann auch $f_{k+1.k}$ die Form

$$f_{k+1.k} = \frac{\xi_v + \psi_k}{\xi_v + \sigma_{k+1}} \qquad [2.3.3.34]$$

annimmt. Zu diesem Zweck betrachten wir die bedingte Wahrscheinlichkeit

$$\text{prob}\{(x_{v1},\ldots,x_{vk},0)|x_{v*} = k\} = \qquad [2.3.3.35]$$

$$= \frac{\text{prob}\{(x_{v1},\ldots,x_{vk}) \wedge x_{vo} = k|\xi_v\}(1 - f_{k+1.k})}{\text{prob}(x_{vo} = k|\xi_v)(1 - f_{k+1.k}) + \text{prob}(x_{vo} = k-1|\xi_v)f_{k+1.k-1}}$$

Da es nur einen Antwortvektor $(x_{v1},\ldots,x_{vk}) = (1,1,\ldots,1)$ gibt, der mit $x_{vo} = k$ verträglich ist, ist in [2.3.3.35]

$$\text{prob}\{(x_{v1},\ldots,x_{vk}) \wedge x_{vo} = k|\xi_v\} = \text{prob}(x_{vo} = k|\xi_v)$$

so dass

$$\frac{f_{k+1.k-1}\,\text{prob}(x_{vo} = k-1|\xi_v)}{(1 - f_{k+1.k})\text{prob}(x_{vo} = k|\xi_v)} = c \qquad [2.3.3.36]$$

eine von ξ_v unabhängige Konstante ist. Darin ist

$$\text{prob}(x_{vo} = k-1|\xi_v) = \frac{\prod_{r=0}^{k-2}(\xi_v + \psi_r)}{\prod_{i=1}^{k}(\xi_v + \sigma_i)} \sum_{i=1}^{k}(\sigma_i - \psi_{i-1})$$

und

$$\text{prob}(x_{vo} = k|\xi_v) = \frac{\prod_{r=0}^{k-1}(\xi_v + \psi_r)}{\prod_{i=1}^{k}(\xi_v + \sigma_i)},$$

so dass

$$\frac{\text{prob}(x_{vo} = k-1|\xi_v)}{\text{prob}(x_{vo} = k|\xi_v)} = \frac{\sum_{i=1}^{k}(\sigma_i - \psi_{i-1})}{\xi_v + \psi_{k-1}}.$$

Einsetzen von [2.3.3.33] und [2.3.3.37] in [2.3.3.36] ergibt schließlich

$$\frac{\frac{\xi_v + \psi_{k-1}}{\xi_v + \sigma_{k+1}} \sum_{i=1}^{k}(\sigma_i - \psi_{i-1})}{(1 - f_{k+1.k})(\xi_v + \psi_{k-1})} = c$$

$$\frac{\sum_{i=1}^{k}(\sigma_i - \psi_{i-1})}{(\xi_v + \sigma_{k+1})(1 - f_{k+1.k})} = c$$

$$\sum_{i=1}^{k}(\sigma_i - \psi_{i-1}) = c(\xi_v + \sigma_{k+1})(1 - f_{k+1.k})$$

$$\sum_{i=1}^{k}(\sigma_i - \psi_{i-1}) = c\xi_v - c\xi_v f_{k+1.k} + c\sigma_{k+1} - c\sigma_{k+1} f_{k+1.k}$$

$$c\, f_{k+1.k}(\xi_v + \sigma_{k+1}) = c(\xi_v + \sigma_{k+1}) - \sum_{i=1}^{k}(\sigma_i - \psi_{i-1})$$

$$f_{k+1.k} = \frac{\xi_v + \sigma_{k+1} - \frac{1}{c}\sum_{i=1}^{k}(\sigma_i - \psi_{i-1})}{\xi_v + \sigma_{k+1}},$$

so dass $f_{k+1.k}$ mit der Definition

$$\psi_k = \sigma_{k+1} - \frac{1}{c}\sum_{i=1}^{k}(\sigma_i - \psi_{i-1})$$

die Form

$$f_{k+1.k} = \frac{\xi_v + \psi_k}{\xi_v + \sigma_{k+1}}$$

annimmt.

1. Der Beweis sei dem Leser als Übungsaufgabe überlassen.

Kontrollfragen

- Formulieren Sie ein allgemeines testtheoretisches Modell, welches sowohl die klassische Testtheorie als auch die stochastischen Testmodelle mit einschließt!
- Welche Varianten lassen sich innerhalb des allgemeinen testtheoretischen Modells unterscheiden?
- Durch welche Zusatzannahmen sind diese Varianten definiert und welche Hierarchie besteht zwischen den verschiedenen Modellspezifikationen?
- Wie errechnet sich der True-Score einer Vp
  a) im Falle des Pure-Random-Modells
  b) im Falle der Ein-Klassen-Lösung der LCA
  c) im Falle des Binomialmodells
  d) im Falle von Latent-Trait-Modellen mit dichotomen Items
  d) im Falle des (dichotomen) Rasch-Modells
  e) im Falle der Latent-Class-Analyse mit polytomen Items und
  f) im Falle der Latent-Class-Analyse mit dichotomen Items?

- Nennen Sie Gründe, warum die diagnostische Brauchbarkeit des Summenscores eingeschränkt sein kann!
- Wovon hängt die Brauchbarkeit des Summenscores in der Latent-Class-Analyse ab?
- Wann kann man in der Latent-Class-Analyse von quantitativ verschiedenen Klassen sprechen? Wann sind die Klassen qualitativ verschieden?
- Wie konstruiert man die Profillinien der latenten Klassen und wie kann mithilfe der Profillinien überprüft werden, ob sich die Klassen quantitativ oder qualitativ unterscheiden?
- Was versteht man unter der Homogenität eines Tests und wie bildet sich die Homogenität in der Latent-Class-Analyse ab?
- Auf welcher Art von Skala wird die latente Personenvariable gemessen, wenn der Test
  a) homogen
  b) inhomogen
  ist?
- Was versteht man unter der Suffizienz des Summenscores und wie ist sie definiert?
- Welche Form haben die Profillinien der LCA, wenn die Itemcharakteristiken dem Binomialmodell folgen?
- Welche Form müssen die Itemcharakteristiken
  a) bei lokal unabhängigen binären Items
  b) bei seriell abhängigen binären Items
  besitzen, damit der Summenscore eine erschöpfende Beschreibung der Testleistung der Probanden liefert?
- Welche Form haben die Itemfunktionen des Rasch-Modells für
  a) polytome
  b) ordinale
  Antwortvariablen und welche Beziehung besteht zwischen den beiden Modellen?
- Was versteht man unter den sog. Schwellenwahrscheinlichkeiten und welcher Zusammenhang besteht zwischen den Schwellenwahrscheinlichkeiten, dem binären Rasch-Modell und dem Rasch-Modell für ordinale Antwortvariablen?
- Welche Form haben die Itemfunktionen des Poisson-Modells und auf welche Art von Testaufgaben ist das Modell anwendbar?
- Welche Voraussetzungen sind
  a) bei lokal unabhängigen Itemcharakteristiken
  b) bei seriell abhängigen Itemcharakteristiken
  notwendig und hinreichend für die Suffizienz des Summenscores?

Übungsaufgaben

1. Ein Test aus vier binären Items wurde mittels Latent-Class-Analyse untersucht. Das Ergebnis ist in nachstehender Tabelle wiedergegeben:

Klasse	Lösungswahrscheinlichkeiten			
g	p_{g1}	p_{g2}	p_{g3}	p_{g4}
1	0.67	0.99	0.00	0.67
2	0.46	0.95	0.50	0.28
3	0.34	0.9	0.09	0.47
Gesamt:	0.49	0.95	0.29	0.42

Stellen Sie die Profillinien der Items dar und entscheiden Sie, ob die Klassen qualitativ verschieden sind!

2. Zeigen Sie, dass die Schwellenwahrscheinlichkeiten des Poissonmodells der Gleichung [2.3.2.33] folgen!
3. Gegeben sei ein Test aus drei lokal unabhängigen binären Items mit den Leichtigkeitsparametern $\varepsilon_1 = 0.5$, $\varepsilon_2 = 1$ und $\varepsilon_3 = 2$.
   a) Berechnen Sie die symmetrischen Grundfunktionen!
   b) Berechnen Sie die bedingte Wahrscheinlichkeit des Antwortvektors (0,1,1) bei gegebenem Score!
4. Zeigen Sie, dass

   a) $f_{1.0} = \dfrac{\xi_v}{\xi_v + 1}$ und

   b) $f_{2.0} = \dfrac{\xi_v}{\xi_v + \sigma_2}$

   erfüllt sind, wenn ξ_v und σ_2 gemäß [2.3.3.24–2.3.3.25] definiert sind!

2.4

Parameterschätzung, Modellkontrolle und Beurteilung der Testleistung der Probanden in der klassischen Testtheorie

2.4.1 Parameterschätzung

Momentenmethode

In der klassischen Testtheorie wird die zu messende Eigenschaft der Vpn durch ihren True-Score (τ_{vt}) beschrieben. Dieser ist als Erwartungswert der (individuellen) Scorevariable der Vp definiert. Zu seiner Schätzung können wir daher ganz unproblematisch die Momentenmethode anwenden und den tatsächlich beobachteten Score der Vp (x_{vt}) als Schätzwert $(\hat{\tau}_{vt})$ für den True-Score einsetzen

$$\hat{\tau}_{vt} = x_{vt}. \qquad [2.4.1.1]$$

Standardmessfehler

Wir erhalten so eine erwartungstreue Schätzfunktion für den True-Score der Vp, deren Wirksamkeit wir mittels des Standardmessfehlers

$$\sigma(\hat{\tau}_{vt}) \cong \sigma(F_{ot}) = \sigma(X_{ot})\sqrt{1-\rho^2_{XT}} \qquad [2.4.1.2]$$

(vgl. Kap. 2.2.2) abschätzen können.

2.4.2 Überprüfung der Modellannahmen

Zusatzannahmen der klassischen Testtheorie

Auch die Modellkontrolle scheint in der klassischen Testtheorie auf den ersten Blick völlig unproblematisch zu sein. Die Modellannahmen von Novick haben keinerlei empirischen Gehalt, weshalb die Rechenformeln der klassischen Testtheorie von vielen Diagnostikern als universell anwendbar betrachtet werden.

Tatsächlich müssen zu ihrer praktischen Anwendung jedoch Zusatzannahmen – wie z. B. die (essentielle) Parallelität oder (zumindest) die essentielle τ-Äquivalenz von Tests, Subtests oder

Testteilen – getroffen werden, die sehr wohl empirischen Gehalt besitzen und daher einer Überprüfung bedürfen.

Erste Anhaltspunkte für eine (mögliche) Verletzung dieser Zusatzannahmen können aus dem Vergleich der verschiedenen Reliabilitätsmaße gewonnen werden (vgl. Kap. 2.2.2). Darüber hinaus kann man die jeweils getroffenen Zusatzannahmen aber auch über die folgenden Beziehungen explizit testen.

Überprüfung der Parallelität

Seien t und t' *parallele Tests*, Subtests oder Testteile, so haben sie unabhängig von der Wahl der Referenzpopulation stets den gleichen Erwartungswert

$$E(X_{ot}) = E(X_{ot'}) \qquad [2.4.2.1]$$

die gleiche Maßzahlvarianz

$$\sigma^2(X_{ot}) = \sigma^2(X_{ot'}) \qquad [2.4.2.2]$$

und gleich hohe Korrelationen mit Drittvariablen

$$\rho(X_{ot}, Y_o) = \rho(X_{ot'}, Y_o) \qquad [2.4.2.3]$$

Überprüfung der essentiellen Parallelität

Während die Beziehungen in den Gleichungen [2.4.2.2] und [2.4.2.3] auch für *essentiell parallele Tests* gelten, schwächt sich die Beziehung [2.4.2.1] im Falle der essentiellen Parallelität zu

$$E(X_{ot} - X_{ot'}) = c_{tt'} \qquad [2.4.2.4]$$

ab, worin $c_{tt'}$ eine Konstante darstellt, die von der Wahl der Referenzpopulation unabhängig ist und also in jeder wie auch immer selegierten Subpopulation von Probanden gleich hoch ausfallen muss.

Betrachten wir z. B. die Subpopulation der Hochbegabten (g = 1) und die Subpopulation der Minderbegabten (g = 2), so muss der Erwartungswert der Scoredifferenz aus den beiden Tests in beiden Subpopulationen gleich hoch ausfallen

$$E(X_{1t} - X_{1t'}) = E(X_{2t} - X_{2t'}) \,. \qquad [2.4.2.5]$$

Überprüfung der τ-Äquivalenz und der essentiellen τ-Äquivalenz

Sind t und t' lediglich τ-äquivalent oder essentiell τ-äquivalent, so entfallen die in den Gleichungen [2.4.2.2] und [2.4.2.3] dargestellten Beziehungen.

Im Falle der *τ-Äquivalenz* gilt nur noch die Beziehung in Gleichung [2.4.2.1]. Im Falle der *essentiellen τ-Äquivalenz* tritt an ihre Stelle die in Gleichung [2.4.2.4] bzw. (gleichwertig) in Gleichung [2.4.2.5] dargestellte Beziehung.

Einen Überblick gibt Tabelle 2.4.2.1.

Tab. 2.4.2.1: Eigenschaften von parallelen, essentiell parallelen, τ-äquivalenten und essentiell τ-äquivalenten Tests

τ-Äquivalenz	Parallelität	Essentielle Parallelität	Essentielle τ-Äquivalenz
[2.4.2.1] $E(X_{ot}) = E(X_{ot'})$		[2.4.2.5] $E(X_{1t} - X_{1t'}) = E(X_{2t} - X_{2t'})$	
	[2.4.2.2] $\sigma^2(X_{ot}) = \sigma^2(X_{ot'})$ [2.4.2.3] $\rho(X_{ot}, Y_o) = \rho(X_{ot'}, Y_o)$		

Statistische Testverfahren

Zur Überprüfung der getroffenen Zusatzannahmen kann auf die gebräuchlichen statistischen Testverfahren zurückgegriffen werden, wie sie in jedem Lehrbuch der Statistik zu finden sind. Die in Tabelle 2.4.2.1 dargestellten Beziehungen entsprechen dabei der Nullhypothese. Wird die Nullhypothese verworfen, so kann darauf geschlossen werden, dass die entsprechende Zusatzannahme verletzt ist.

τ-Äquivalenz und Parallelität: Test der Hypothese in Gleichung [2.4.2.1]

Zur Überprüfung der Annahme in Gleichung [2.4.2.1] kann das Hypothesenpaar

$$H_0 : E(X_{ot} - X_{ot'}) = 0$$

$$H_1 : E(X_{ot} - X_{ot'}) \neq 0$$

mittels der Prüfgröße

$$t = \sqrt{n}\,\frac{m(X_{ot} - X_{ot'})}{\hat{\sigma}(X_{ot} - X_{ot'})} \qquad [2.4.2.6]$$

gegen die t-Verteilung mit df = n-1 Freiheitsgraden getestet werden (t-Test für abhängige Stichproben). Wird H_0 verworfen, so sind die beiden Tests nicht τ-äquivalent und daher auch nicht parallel.

Essentielle τ-Äquivalenz und essentielle Parallelität: Test der Hypothese in Gleichung [2.4.2.5]

Zur Überprüfung der Annahme in Gleichung [2.4.2.5] teilen wir die Personenstichprobe in zwei gleich große Teilstichproben g = 1 und g = 2 mit den Stichprobenumfängen $n_1 = n_2 = n/2$ und testen das Hypothesenpaar

$$H_0 : E(X_{1t} - X_{1t'}) = E(X_{2t} - X_{2t'})$$

$$H_1 : E(X_{1t} - X_{1t'}) \neq E(X_{2t} - X_{2t'})$$

mittels der Prüfgröße

$$t = \sqrt{\frac{n}{2}} \frac{m(X_{1t} - X_{1t'}) - m(X_{2t} - X_{2t'})}{\sqrt{\hat{\sigma}^2(X_{1t} - X_{1t'}) + \hat{\sigma}^2(X_{2t} - X_{2t'})}} \qquad [2.4.2.7]$$

gegen die t-Verteilung mit df = n - 2 Freiheitsgraden (t-Test für unabhängige Stichproben). Wird H_0 verworfen, so sind die beiden Tests nicht essentiell τ-äquivalent und daher auch nicht τ-äquivalent, nicht parallel und auch nicht essentiell parallel.

Parallelität und essentielle Parallelität: Test der Hypothese in Gleichung [2.4.2.2]

Da wir es – wie schon bei Gleichung [2.4.2.1] – auch hier wieder mit abhängigen Stichproben zu tun haben, kann die Überprüfung der Annahme in Gleichung [2.4.2.2] nicht mittels F-Tests erfolgen. Anstelle der Abweichungsquadrate berechnen wir daher die Abweichungsbeträge $D_{ot} = |X_{ot} - m(X_{ot})|$ und $D_{ot'} = |X_{ot'} - m(X_{ot'})|$ und prüfen die mit Gleichung [2.4.2.2] gleichwertige Nullhypothese

$$H_0 : E(D_{ot} - D_{ot'}) = 0$$

mittels t-Tests für abhängige Stichproben gegen die Alternativhypothese

$$H_1 : E(D_{ot} - D_{ot'}) \neq 0.$$

Die Prüfgröße lautet in diesem Fall

$$t = \sqrt{n} \frac{m(D_{ot} - D_{ot'})}{\hat{\sigma}(D_{ot} - D_{ot'})} \qquad [2.4.2.8]$$

und hat bei Geltung von H_0 wieder eine t-Verteilung mit df = n-1 Freiheitsgraden. Wird H_0 verworfen, so sind die Maßzahlvarianzen der beiden Tests verschieden und die Tests daher weder parallel noch essentiell parallel.

Parallelität und essentielle Parallelität: Test der Hypothese in Gleichung [2.4.2.3]

Zur Überprüfung der Annahme in Gleichung [2.4.2.3] testen wir das Hypothesenpaar

$$H_0 : \rho(X_{ot}, Y_o) = \rho(X_{ot'}, Y_o)$$

$$H_1 : \rho(X_{ot}, Y_o) \neq \rho(X_{ot'}, Y_o)$$

mittels der standardnormalverteilten Prüfgröße

$$z = \frac{\sqrt{n-3}\,(z_{tY} - z_{t'Y})}{\sqrt{2 - 2CV_1}}, \qquad [2.4.2.9]$$

worin z_{tY} und $z_{t'Y}$ die nach Fishers Z-Transformation

$$z_{ab} = \frac{1}{2}\ln\left(\frac{1 + r_{ab}}{1 - r_{ab}}\right)$$

berechneten Z-Werte der Korrelationen $r_{tY} = r(X_{ot}, Y_o)$ und $r_{t'Y} = r(X_{ot'}, Y_o)$ sind. CV_1 bezeichnet die Kovarianz der Korrelationsverteilungen von r_{tY} und $r_{t'Y}$ und berechnet sich nach

$$CV_1 = \frac{1}{(1 - r_{\#Y})^2}\left(r_{tt'}\left(1 - 2r_{\#Y}^2\right)\right) - \frac{1}{2}r_{\#Y}^2\left(1 - 2r_{\#Y}^2 - r_{tt'}^2\right)$$

mit $r_{\#Y} = (r_{tY} + r_{t'Y})/2$ und $r_{tt'} = r(X_{ot}, X_{ot'})$ (vgl. Bortz, 1993, 204ff).

Wird H_0 verworfen, so korrelieren die beiden Tests mit der Kriteriumsvariable Y unterschiedlich hoch und sind daher weder parallel noch essentiell parallel.

2.4.3 Beurteilung der Testleistungen

Konfidenzintervall für den True-Score

In der diagnostischen Praxis wird man sich mit der Punktschätzung der True-Scores nicht zufrieden geben, sondern nach einem Vertrauensbereich fragen, innerhalb dessen der True-Score eines Probanden mit einer vorgegebenen Sicherheit von γ = 95% oder 99% liegt.

Unter der Zusatzannahme, dass die individuelle Scorevariable X_{vt} näherungsweise normalverteilt ist, gibt die klassische Testtheorie hierfür das (näherungsweise) Konfidenzintervall

$$\mathrm{KONF}_{\gamma}\{x_{vt} - c\,\sigma(F_{ot}) \le \tau_{vt} \le x_{vt} + c\,\sigma(F_{ot})\} \qquad [2.4.3.1]$$

an, worin $\sigma(F_{ot})$ den nach Gleichung [2.2.2.16] berechneten Standardmessfehler darstellt und die Werte für c der Tabelle für die Standardnormalverteilung im Anhang entnommen werden können, so dass

$$\mathrm{prob}\{-c \le Z \le c\} = \gamma \qquad [2.4.3.2]$$

und je nach Konfidenzzahl c = 1.96 (γ = 95%) oder c = 2.576 (γ = 99%) betragen.

Test für den True-Score

Logisch äquivalent mit dem Konfidenzintervall [2.4.3.1] ist die approximativ nach N(0,1) verteilte Testgröße

$$z = \frac{x_{vt} - \tau_{krit}}{\sigma(F_{ot})} \qquad [2.4.3.3]$$

zur Überprüfung des Hypothesenpaares

$$H_0 : \tau_{vt} = \tau_{krit}$$

$$H_1 : \tau_{vt} \neq \tau_{krit}\ ,$$

mittels derer man überprüfen kann, ob die Testleistung einen vorgegebenen Kriteriumswert τ_{krit} signifikant über- oder unterschreitet.

Kritische Differenz der Testleistungen zweier Vpn

Aus denselben Gründen folgt auch die Testgröße

$$z = \frac{x_{vt} - x_{wt}}{\sqrt{2}\,\sigma(F_{ot})} \qquad [2.4.3.4]$$

zur Beurteilung der Leistungsunterschiede zweier Probanden v und w

$$H_0 : \tau_{vt} = \tau_{wt}$$

$$H_1 : \tau_{vt} \neq \tau_{wt}$$

näherungsweise einer Standardnormalverteilung.

Test von McNemar

Eleganter ist die Beurteilung der Leistungsunterschiede zweier Vpn mittels des parameterfreien Tests von McNemar (1949), des-

sen Anwendung allerdings daran gebunden ist, dass nicht nur die Scores, sondern auch die Antwortvektoren der Probanden genau protokolliert wurden, so dass ein itemweiser Vergleich ihrer Testergebnisse möglich ist.

Zu diesem Zweck trägt man die Itemantworten der Vpn in eine Vierfeldertafel ein, deren Zeilen die Antworten der Vp v und deren Spalten die Antworten der Vp w widerspiegeln, so dass n_{ij} die Anzahl der Items bezeichnet, welche von Vp v in Kategorie i und von Vp w in Kategorie j beantwortet wurden.

Tab. 2.4.3.1: Test von McNemar

		x_{wi}	
		1	0
x_{vi}	1	n_{11}	n_{10}
	0	n_{01}	n_{00}

Items, die von beiden Probanden gelöst wurden, und Items, die von keinem der Probanden gelöst wurden, enthalten keinerlei Information über mögliche Leistungsunterschiede zwischen den Vpn. Eine solche Information ist nur in jenen $n_{10} + n_{01}$ Items enthalten, die von genau einem der beiden Probanden gelöst wurden, so dass $x_{oi} = x_{vi} + x_{wi} = 1$.

Items mit einem Antwortmuster $(x_{vi}, x_{wi}) = (1,0)$ sprechen zugunsten der Vp v, Items mit einem Antwortmuster $(x_{vi}, x_{wi}) = (0,1)$ zugunsten von Vp w. Besteht kein Leistungsunterschied zwischen den Probanden, sollten die beiden Antwortmuster einander die Waage halten, so dass

$$H_0 : \text{prob}\{(1,0) | x_{oi} = 1\} = \text{prob}\{(0,1) | x_{oi} = 1\}.$$

Die Häufigkeit n_{10} folgt dann einer Binomialverteilung mit der Verteilungsfunktion

$$\text{prob}(n_{10} \le x) = F(x) = \sum_{y=0}^{x} \binom{n}{y} p^y (1-p)^{n-y} \qquad [2.4.3.5]$$

und den Parametern $n = n_{10} + n_{01}$ und $p = 0.5$.

Wir werden die Nullhypothese daher zugunsten der Alternativhypothese

$$H_1 : \text{prob}\{(1,0) | x_{oi} = 1\} \neq \text{prob}\{(0,1) | x_{oi} = 1\}$$

verwerfen, wenn n_{10}

- entweder kleiner/gleich jenem x ist, für welches $F(x) \leq \alpha/2$,
- oder größer/gleich jenem x ist, für welches $1 - F(x-1) \leq \alpha/2$.

Die Werte für x können der Tabelle für die Binomialverteilung im Anhang entnommen werden.

Kontrollfragen

- Wodurch wird die gemessene Personeneigenschaft in der klassischen Testtheorie beschrieben?
- Wodurch und mittels welcher Methode kann der True-Score der klassischen Testtheorie geschätzt werden?
- Wie kann die Wirksamkeit der Schätzfunktion für den True-Score abgeschätzt werden?
- Anhand welcher Beziehungen kann
  a) die Parallelität
  b) die essentielle Parallelität
  c) die τ-Äquivalenz
  d) die essentielle τ-Äquivalenz
  von Tests oder Subtests überprüft werden? Formulieren Sie Nullhypothese und Alternativhypothese und geben Sie an,
  a) mittels welcher statistischen Testgrößen und
  b) gegen welche Prüfverteilung
  die Überprüfung jeweils erfolgen kann!
- Welche Zusatzannahmen sind erforderlich, um ein Konfidenzintervall für den True-Score angeben zu können?
- Welche Zusatzannahmen werden in der klassischen Testtheorie
  a) beim Test für den True-Score
  b) bei der Beurteilung der kritischen Differenz der Testleistungen zweier Vpn
  getroffen?
- Wie lauten Nullhypothese und Alternativhypothese beim Test von McNemar?
- Gegen welche Prüfverteilung wird beim Test von McNemar getestet und wann wird die Nullhypothese verworfen?

Übungsaufgaben

1. Ein Test aus k = 20 Items besitzt die Maßzahlvarianz $\sigma^2(X_{ot}) = 20.25$ und die Reliabilität $\rho^2_{XT} = 0.91$. Schätzen Sie die Wirksamkeit der Schätzfunktion für den True-Score ab!
2. Gegeben sei eine Stichprobe von n = 30 Vpn. Von jeder Vp liegen die Testergebnisse in den beiden "Parallelformen" (A und B) eines Leis-

tungstests sowie die durchschnittliche Abiturnote (Y) vor.

Vp	A	B	Y	Vp	A	B	Y	Vp	A	B	Y
1	120	105	1.35	11	111	100	2.26	21	117	108	2.59
2	119	102	1.33	12	118	107	1.63	22	114	105	2.57
3	115	103	2.24	13	119	107	1.37	23	120	110	1.08
4	116	102	2.80	14	112	103	2.47	24	114	103	1.85
5	119	102	1.79	15	112	102	2.87	25	111	101	2.46
6	112	105	2.20	16	116	107	1.53	26	118	108	1.96
7	110	104	2.54	17	115	105	1.27	27	117	106	1.93
8	113	103	1.87	18	117	107	1.20	28	118	107	1.42
9	114	105	1.31	19	113	103	2.40	29	110	100	2.49
10	113	106	2.60	20	115	106	1.79	30	116	106	1.77

Überprüfen Sie, ob die Testformen
a) τ-äquivalent
b) essentiell τ-äquivalent
c) parallel und/oder
d) essentiell parallel sind!

3. Eine Vp v hat in einem Leistungstest 109 IQ-Punkte erzielt. Die Reliabilität des Tests beträgt $\rho^2_{XT} = 0.94$. Geben Sie einen Vertrauensbereich an, innerhalb dessen der True-Score der Vp mit einer Sicherheit von 99% liegt!
4. Eine andere Vp w hat in dem Test aus Aufgabe 3 nur 98 IQ-Punkte erzielt. Überprüfen Sie, ob die Leistungsfähigkeit der Vp vom Durchschnitt der Referenzpopulation signifikant abweicht (α = 5%)!
5. Überprüfen Sie, ob sich die Vpn aus Aufgabe 3 und 4 in ihrer Leistungsfähigkeit signifikant voneinander unterscheiden (α = 1%)!
6. Zwei Vpn v und w haben in einem Test aus k = 20 binären Items die in nachstehender Tabelle dargestellten Antwortvektoren erzielt:

Vp	Item Nr.																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
v	1	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
w	1	1	0	0	1	1	1	0	1	1	0	1	0	1	1	1	1	1	0	1

Beurteilen Sie mittels des Tests von McNemar, ob die Testleistung der Probanden signifikant verschieden ist (α = 5%)!

2.5
Parameterschätzung und Modellkontrolle im saturierten Modell

2.5.1 Parameterschätzung

Im saturierten Modell definiert jedes der

$$h_{max} = \prod_{i=1}^{k} m_i \qquad [2.5.1.1]$$

möglichen Antwortmuster $x_g = (x_{g1},\ldots,x_{gk})$ eine eigene Klasse von Vpn, so dass die Likelihood der Antwortmatrix in der Form

$$L(X) = \prod_{v=1}^{n} \text{prob}\{(x_{v1},\ldots,x_{vk})\} = \prod_{g=1}^{h_{max}} p_g^{n_g} \quad . \qquad [2.5.1.2]$$

dargestellt werden kann und ihr natürlicher Logarithmus die Form

$$\ln\{L(X)\} = \sum_{g=1}^{h_{max}} n_g \ln(p_g) \qquad [2.5.1.3]$$

annimmt, worin n_g die Anzahl der Vpn bezeichnet, deren Antwortmuster mit x_g übereinstimmt.

Als unbekannte Parameter treten hier lediglich die Klassengrößen

$$p_g = \text{prob}(v \in g) \qquad [2.5.1.4]$$
$$= \text{prob}\{(x_{v1},\ldots,x_{vk}) = (x_{g1},\ldots,x_{gk})\}$$

auf, die mittels Maximum-Likelihood-Methode durch die relative Häufigkeit der entsprechenden Antwortmuster

$$\hat{p}_g = \frac{n_g}{n} \qquad [2.5.1.5]$$

mit der Fehlerstreuung

$$\sigma(\hat{p}_g) = \sqrt{\frac{p_g(1-p_g)}{n}} \qquad [2.5.1.6]$$

geschätzt werden.

Die Anzahl der zu schätzenden Klassengrößen entspricht der Anzahl der möglichen Antwortmuster. Da die Summe der Klassengrößen jedoch Eins ergeben muss

$$\sum_{g=1}^{h_{max}} p_g = 1 \qquad [2.5.1.7]$$

sind aber nur h - 1 Klassengrößen unabhängig voneinander, so dass sich die *Anzahl der unabhängigen Modellparameter* im saturierten Modell zu

$$n(P_{sat}) = h_{max} - 1 \qquad [2.5.1.8]$$

ergibt.

2.5.2 Modellkontrolle

Optimale Beschreibung der Datenmatrix

Das saturierte Modell enthält keinerlei einschränkende Modellannahmen und hat daher auch keinen empirischen Gehalt. Aufgrund seiner Voraussetzungslosigkeit liefert es die größtmögliche Likelihood der Datenmatrix und somit – absolut gesehen – die bestmögliche Beschreibung der Daten.

Mangelnde Ökonomie

Dies wird jedoch um den Preis einer immens großen Anzahl zu schätzender Modellparameter erkauft. Selbst bei dichotomen Items wächst die Anzahl der möglichen Antwortmuster sehr rasch (vgl. Abbildung 2.5.2.1) und beträgt bei nur k=10 dichotomen Items bereits $h_{max} = 2^{10} = 1024$. Bei k = 10 dreistufigen Items ist die Anzahl der möglichen Antwortmuster $h_{max} = 3^{10} = 59049$, usw.

Man wird an dem saturierten Modell daher nur so lange festhalten, als sich kein einfacheres Modell (mit geringerer Anzahl von unabhängigen Modellparametern) finden lässt, das die Daten ähnlich gut beschreibt.

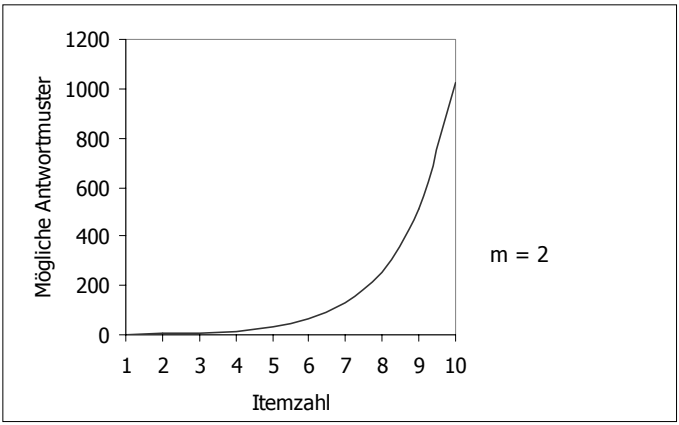


Abb. 2.5.2.1: Anzahl der möglichen Antwortmuster bei zunehmender Itemzahl

Kontrollfragen

- Welches Modell liefert die größtmögliche Likelihood der Datenmatrix und warum ist dies der Fall?
- Wie teilt das saturierte Modell die Vpn in Klassen ein?
- Was beschreiben die Modellparameter des saturierten Modells und wodurch können sie geschätzt werden?
- Wie groß ist die Anzahl der unabhängigen Modellparameter des saturierten Modells?
- Was spricht dagegen, die Daten *in jedem Fall* durch das saturierte Modell zu beschreiben, und wann wird man das saturierte Modell verwerfen?

Übungsaufgaben

1. Ein Test aus drei binären Items wurde von n = 87 Vpn bearbeitet. Nachstehende Tabelle gibt die möglichen Antwortmuster und deren Häufigkeit wieder.

g	1	2	3	4	5	6	7	8
x_g	000	100	010	001	110	101	011	111
n_g	2	7	6	9	17	15	12	19

a) Schätzen Sie die Klassengrößenparameter des saturierten Modells und
b) geben Sie an, wie viele der Parameterschätzungen voneinander unabhängig sind!
c) Berechnen Sie die Fehlerstreuung der Parameterschätzungen!
d) Berechnen Sie die logarithmierte Likelihood des saturierten Modells!

2. Wie viele verschiedene Antwortmuster sind bei einem Test aus
   a) 5 binären Items
   b) 3 fünfstufigen Items
   c) 5 binären und drei fünfstufigen Items
   möglich?

2.6

Parameterschätzung und Modellkontrolle in Modellen mit personen- und/oder itemunabhängigen Antwortvariablen

2.6.1 Parameterschätzung

Pure-Random-Modell

Die einfachste Beschreibung der Daten (mit der geringsten Anzahl von unabhängigen Modellparametern) liefert das Pure-Random-Modell, das zugleich auch die strengsten Modellannahmen trifft. Damit die Antwortvariablen X_{vi} personen- und itemunabhängig verteilt sind, müssen alle Items dieselbe Kategorienzahl

$$m_i = m \qquad [2.6.1.1]$$

und für alle Vpn dieselben Kategorienwahrscheinlichkeiten

$$\text{prob}(X_{vi} = x) = p_{\bullet\bullet x} \qquad [2.6.1.2]$$

aufweisen, so dass die Likelihood der Antwortmatrix in der Form

$$L(X) = \prod_{v=1}^{n} \text{prob}\{(x_{v1}, \ldots, x_{vk})\} \qquad [2.6.1.3]$$

$$= \prod_{v=1}^{n} \prod_{i=1}^{k} \text{prob}(x_{vi})$$

$$= \prod_{v=1}^{n} \prod_{i=1}^{k} \prod_{x=0}^{m-1} p_{\bullet\bullet x}^{n_{vix}}$$

$$= \prod_{x=0}^{m-1} p_{\bullet\bullet x}^{n_{oox}}$$

dargestellt werden kann, worin

$$n_{vix} = \begin{cases} 1 & \text{wenn } x_{vi} = x \\ O & \text{sonst} \end{cases} \qquad [2.6.1.4]$$

und

$$n_{oox} = \sum_{v=1}^{n}\sum_{i=1}^{k} n_{vix} \ . \qquad [2.6.1.5]$$

Für den natürlichen Logarithmus der Likelihood erhalten wir entsprechend

$$\ln\{L(X)\} = \sum_{x=0}^{m-1} n_{oox} \ln(p_{\bullet\bullet x}) \ . \qquad [2.6.1.6]$$

Als unbekannte Modellparameter treten darin m (personen- und itemunabhängige) Kategorienwahrscheinlichkeiten $p_{\bullet\bullet x}$ auf, die mittels der Maximum-Likelihood-Methode durch die relative Häufigkeit der entsprechenden Antwortkategorie

$$\hat{p}_{\bullet\bullet x} = \frac{n_{oox}}{nk} \qquad [2.6.1.7]$$

mit dem Schätzfehler

$$\sigma(\hat{p}_{\bullet\bullet x}) = \sqrt{\frac{p_{\bullet\bullet x}(1 - p_{\bullet\bullet x})}{nk}} \qquad [2.6.1.8]$$

geschätzt werden.

Da die Summe der Kategorienwahrscheinlichkeiten gleich Eins ist

$$\sum_{x=0}^{m-1} p_{\bullet\bullet x} = 1 \quad , \qquad [2.6.1.9]$$

sind aber nur m-1 Kategorienwahrscheinlichkeiten unabhängig voneinander, so dass sich die *Anzahl der unabhängigen Modellparameter* im Pure-Random-Modell zu

$$n(P_{PR}) = m - 1 \qquad [2.6.1.10]$$

ergibt.

Identische Personen

Auch die Ein-Klassen-Lösung der Latent-Class-Analyse postuliert personenunabhängig verteilte Antwortvariablen. Verglichen mit dem Pure Random Modell ist die Ein-Klassen-LCA jedoch insofern weniger restriktiv, als die Items unterschiedliche Kategorienzahlen m_i und itemabhängige Kategorienwahrscheinlichkeiten

$$\text{prob}(X_{vi} = x) = p_{\bullet ix} \qquad [2.6.1.11]$$

besitzen dürfen.

Die Likelihood der Antwortmatrix nimmt daher die Form

$$L(X) = \prod_{v=1}^{n} \text{prob}\{(x_{v1}, \ldots, x_{vk})\} \qquad [2.6.1.12]$$

$$= \prod_{v=1}^{n} \prod_{i=1}^{k} \text{prob}(x_{vi})$$

$$= \prod_{v=1}^{n} \prod_{i=1}^{k} \prod_{x=0}^{m_i-1} p_{\bullet ix}^{n_{vix}}$$

$$= \prod_{i=1}^{k} \prod_{x=0}^{m_i-1} p_{\bullet ix}^{n_{oix}}$$

an, worin

$$n_{oix} = \sum_{v=1}^{n} n_{vix} \ . \qquad [2.6.1.13]$$

Für den natürlichen Logarithmus der Likelihood erhalten wir entsprechend

$$\ln\{L(X)\} = \sum_{i=1}^{k} \sum_{x=0}^{m_i-1} n_{oix} \ln(p_{\bullet ix}). \qquad [2.6.1.14]$$

Als unbekannte Modellparameter treten darin pro Item m_i (personenunabhängige) Kategorienwahrscheinlichkeiten $p_{\bullet ix}$ auf, die mittels der Maximum-Likelihood-Methode durch die relative Häufigkeit geschätzt werden, mit welcher das jeweilige Item in Kategorie x beantwortet wird,

$$\hat{p}_{\bullet ix} = \frac{n_{oix}}{n} \ . \qquad [2.6.1.15]$$

Der Schätzfehler beträgt

$$\sigma(\hat{p}_{\bullet ix}) = \sqrt{\frac{p_{\bullet ix}(1 - p_{\bullet ix})}{n}} \ . \qquad [2.6.1.16]$$

Da sich die Kategorienwahrscheinlichkeiten eines Items stets zu Eins summieren

$$\sum_{x=0}^{m_i-1} p_{\bullet ix} = 1, \qquad [2.6.1.17]$$

sind nur jeweils m_i-1 Kategorienwahrscheinlichkeiten unabhängig voneinander, so dass sich die *Anzahl der unabhängigen Modellparameter* in der Ein-Klassen-LCA zu

$$n(P_{LC1}) = \sum_{i=1}^{k} (m_i - 1) \qquad [2.6.1.18]$$

ergibt. Besitzen alle Items dieselbe Kategorienzahl $m_i = m$, so vereinfacht sich Gleichung [2.6.1.18] zu

$$n(P_{LC1}) = k(m-1). \qquad [2.6.1.19]$$

Identische Items

Von einem psychometrischen Modell im engeren Sinne ist erst dann zu sprechen, wenn die Antwortvariablen personenabhängig verteilt sind. Im einfachsten Fall haben wir es dabei mit identischen Items zu tun, die dieselbe Kategorienzahl $m_i = m$ und dieselben Kategorienwahrscheinlichkeiten

$$\text{prob}(X_{vi} = x | \theta_v) = p_{v\bullet x} \qquad [2.6.1.20]$$

besitzen, so dass die latente Personenvariable als m-dimensionaler Vektor $\theta_v = (\theta_{v0}, \ldots, \theta_{vm-1})$ mit den Elementen

$$\theta_{vx} = p_{v\bullet x} \qquad [2.6.1.21]$$

dargestellt werden kann. Die Likelihood des Antwortvektors einer gegebenen Vp v hat dann die Form

$$\begin{aligned} L(x_v | \theta_v) &= \text{prob}\{(x_{v1}, \ldots, x_{vk}) | \theta_v\} \qquad [2.6.1.22] \\ &= \prod_{i=1}^{k} \text{prob}(x_{vi} | \theta_v) \\ &= \prod_{i=1}^{k} \prod_{x=0}^{m-1} p_{v\bullet x}^{n_{vix}} \\ &= \prod_{x=0}^{m-1} p_{v\bullet x}^{n_{vox}} \end{aligned}$$

mit

$$n_{vox} = \sum_{i=1}^{k} n_{vix} \quad . \qquad [2.6.1.23]$$

Für den natürlichen Logarithmus der Likelihood des Antwortvektors erhalten wir entsprechend

$$\ln\{L(x_v|\theta_v)\} = \sum_{x=0}^{m-1} n_{vox} \ln(p_{v\bullet x}) \, . \qquad [2.6.1.24]$$

Jede Vp wird somit durch m Personenparameter charakterisiert, die mittels Maximum-Likelihood-Methode durch die relative Häufigkeit

$$\hat{p}_{v\bullet x} = \frac{n_{vox}}{k} \qquad [2.6.1.25]$$

mit dem Schätzfehler

$$\sigma(\hat{p}_{v\bullet x}) = \sqrt{\frac{p_{v\bullet x}(1 - p_{v\bullet x})}{k}} \qquad [2.6.1.26]$$

geschätzt werden können. Da die Summe der Kategorienwahrscheinlichkeiten gleich Eins ist, sind nur m-1 der Kategorienwahrscheinlichkeiten unabhängig voneinander, so dass die *Anzahl der unabhängigen Personenparameter* gleich m-1 ist.

Im Falle dichotomer Testitems mit Binomialmodell

$$p_{v\bullet x} = \begin{cases} \theta_v & \text{für } x = 1 \\ 1 - \theta_v & \text{für } x = 0 \end{cases} \qquad [2.6.1.27]$$

(Binomialmodell) reduziert sich [2.6.1.23] wegen [2.3.2.15] zu

$$L(x_v|\theta_v) = \theta_v^{x_{vo}} (1 - \theta_v)^{k - x_{vo}} \qquad [2.6.1.28]$$

und der Personenparameter θ_V kann mittels Maximum-Likelihood-Methode durch die relative Häufigkeit

$$\hat{\theta}_v = \frac{x_{vo}}{k} \qquad [2.6.1.29]$$

geschätzt werden, wobei der Schätzfehler

$$\sigma\left(\hat{\theta}_v\right) = \sqrt{\frac{\theta_v\left(1-\theta_v\right)}{k}} \qquad [2.6.1.30]$$

beträgt. Für die Likelihood der Antwortmatrix ergibt sich dann wegen der Suffizienz des Summenscores (vgl. Kap. 2.3.2)

$$L(X) = \prod_{v=1}^{n} \text{prob}\left\{\left(x_{v1},\ldots,x_{vk}\right)\right\} \qquad [2.6.1.31]$$

$$= \prod_{v=1}^{n} \left\{\text{prob}\left\{\left(x_{v1},\ldots,x_{vk}\right)\middle|x_{vo}\right\}\text{prob}\left(x_{vo}\right)\right\}$$

$$= \prod_{v=1}^{n} \text{prob}\left\{\left(x_{v1},\ldots,x_{vk}\right)\middle|x_{vo}\right\}\prod_{v=1}^{n}\text{prob}\left(x_{vo}\right)$$

$$= \prod_{g=0}^{k} \text{prob}\left\{\left(x_{v1},\ldots,x_{vk}\right)\middle|x_{vo}=g\right\}^{n_g}\prod_{g=0}^{k}\text{prob}\left(x_{vo}=g\right)^{n_g}.$$

Einsetzen für $\text{prob}\left\{\left(x_{v1},\ldots,x_{vk}\right)\middle|x_{vo}=g\right\} = \binom{k}{g}^{-1}$ (vgl. Gleichung [2.3.2.17]) und $\text{prob}\left(x_{vo}=g\right) = p_g$ ergibt schließlich

$$L(X) = \prod_{g=0}^{k}\binom{k}{g}^{-n_g}\prod_{g=0}^{k}p_g^{n_g} \qquad [2.6.1.32]$$

worin n_g für die Anzahl der Vpn steht, deren Score gleich g ist.

Für den natürlichen Logarithmus der Likelihood erhalten wir daher

$$\ln\{L(X)\} = \sum_{g=0}^{k} n_g\left(\ln\left(p_g\right) - \ln\left\{\binom{k}{g}\right\}\right). \qquad [2.6.1.33]$$

Die Likelihood der Datenmatrix hängt im Binomialmodell also nur von k+1 Klassengrößeparametern ab, die mittels Maximum-Likelihood-Methode durch die relative Häufigkeit der entsprechenden Scores

$$\hat{p}_g = \frac{n_g}{n} \qquad [2.6.1.34]$$

geschätzt werden, wobei der Schätzfehler

$$\sigma\left(\hat{p}_g\right) = \sqrt{\frac{p_g\left(1-p_g\right)}{n}} \qquad [2.6.1.35]$$

beträgt. Da sich die Klassengrößen zu Eins summieren

$$\sum_{g=0}^{k} p_g = 1 \qquad [2.6.1.36]$$

sind nur k der k+1 Klassengrößeparameter unabhängig voneinander, so dass die *Anzahl der unabhängigen Modellparameter* im Binomialmodell

$$n(P_{BM}) = k \qquad [2.6.1.37]$$

beträgt.

2.6.2 Überprüfung der Modellannahmen

Pearson-χ^2

Zur Prüfung der Item- und/oder Personenunabhängigkeit der Antwortvariablen kann zunächst auf die gebräuchlichen statistischen Testverfahren zur Untersuchung der Unabhängigkeit in Kontingenztafeln (z. B. Pearson-χ^2) zurückgegriffen werden. Weitere Testverfahren werden wir in Kap. 2.11 behandeln.

	1	...	b	...	m	Σ
1	n_{11}	...	n_{1b}	...	n_{1m}	n_{1o}
:	:		:		:	:
a	n_{a1}	...	n_{ab}	...	n_{am}	n_{ao}
:	:		:		:	:
q	n_{q1}	...	n_{qb}	...	n_{qm}	n_{qo}
Σ	n_{o1}	...	n_{ob}	...	n_{om}	n_{oo}

Tab. 2.6.2.1: Kontingenztafel. Gemeinsame Verteilung (absolute Häufigkeiten n_{ab}) zweier Zufallsvariablen

Sind die Zeilen- und Spaltenvariablen einer Kontingenztafel (Tab. 2.6.2.1) unabhängig voneinander verteilt, dann müssen die Eintragungen in der Kontingenztafel proportional zu den Randsummen sein und dürfen von den erwarteten Häufigkeiten

$$e_{ab} = \frac{n_{ao} n_{ob}}{n_{oo}} \qquad [2.6.2.1]$$

nicht überzufällig stark abweichen. Wir werden die Annahme der Unabhängigkeit daher verwerfen, wenn die asymptotisch nach χ^2 mit

$$df = (q-1)(m-1) \qquad [2.6.2.2]$$

Freiheitsgraden verteilte Testgröße

$$\chi^2 = \sum_{a=1}^{q}\sum_{b=1}^{m}\frac{(n_{ab} - e_{ab})^2}{e_{ab}} \qquad [2.6.2.3]$$

den kritischen Wert der χ^2-Verteilung (siehe Anhang) übersteigt.

Kontrolle der Itemunabhängigkeit

Zur Kontrolle der Itemunabhängigkeit der Antwortvariablen können wir nun o.B.d.A. so vorgehen, dass die Zeilen der Kontingenztafel (a = 1,…,q) den verschiedenen Testitems (i = 1,…,k) und die Spalten (b = 1,…,m) den Antwortkategorien (x = 0,…,m-1) der Items entsprechen. Die Eintragungen in der Kontingenztafel sind dann die Häufigkeiten, mit welchen die Antworten auf die verschiedenen Testitems in die jeweiligen Antwortkategorien fallen

$$n_{ab} = n_{oix} \quad \text{mit} \quad n_{oix} = \sum_{v=1}^{n} n_{vix}\,, \quad i = a \text{ und } x = b\text{-}1, \qquad [2.6.2.4]$$

und die Freiheitsgrade errechnen sich zu

$$df = (k-1)(m-1)\,. \qquad [2.6.2.5]$$

Kontrolle der Personenunabhängigkeit

Zur Kontrolle der Personenunabhängigkeit unterteilen wir die Referenzpopulation in h distinkte Subpopulationen (Klassen von Vpn) g = 1,…,h und berechnen für jedes Item i eine Kontingenztafel, in deren Zeilen (a = 1,…q) die Klassenzugehörigkeit der Probanden (g = 1,…,h) steht. Die Spalten (b = 1,…,m) entsprechen wieder den Antwortkategorien des Items (x = 0,…,m_i-1).

Die Eintragungen in der Kontingenztafel geben dann die Häufigkeit wieder, mit welcher das jeweilige Testitem i von Mitgliedern der Subpopulation g in Kategorie x beantwortet wird:

$$n_{ab} = n_{gix} \quad \text{mit} \quad n_{gix} = \sum_{v \in g} n_{vix}\,, \quad g = a \text{ und } x = b\text{-}1. \qquad [2.6.2.6]$$

Auf diese Weise erhalten wir für jedes Testitem eine asymptotisch χ^2-verteilte Testgröße χ_i^2 mit

$$df_i = (h-1)(m_i - 1) \qquad [2.6.2.7]$$

Freiheitsgraden. Wegen der Additivität der χ^2-Verteilung können wir diese zu der Prüfstatistik

$$\chi_o^2 = \sum_{i=1}^{k} \chi_i^2 \qquad [2.6.2.8]$$

zusammenfassen, die wir gegen die χ^2-Verteilung mit

$$df_o = \sum_{i=1}^{k} df_i \qquad [2.6.2.9]$$

testen. Besitzen alle Items dieselbe Kategorienzahl $m_i = m$, so vereinfacht sich Gleichung [2.6.2.9] zu

$$df_o = k(h-1)(m-1) \qquad [2.6.2.10]$$

bzw. für m = 2 zu

$$df_o = k(h-1). \qquad [2.6.2.11]$$

Kontrollfragen

- Welches Modell liefert die einfachste Beschreibung der Daten?
- Welches Modell trifft die strengsten Modellannahmen und worin bestehen diese?
- Wie viele unabhängige Modellparameter sind im Pure-Random-Modell zur Beschreibung der Likelihood der Daten erforderlich?
- Inwiefern ist die Ein-Klassen-Lösung der LCA weniger restriktiv als das Pure-Random-Modell?
- Wie viele unabhängige Modellparameter sind in der Ein-Klassen-Lösung der LCA zur Beschreibung der Likelihood der Daten erforderlich?
- Inwiefern ist das Binomialmodell weniger restriktiv als das Pure-Random-Modell?
- Wie viele unabhängige Modellparameter sind im Binomialmodell zur Beschreibung der Likelihood der Daten erforderlich?
- Wie sieht die Kontingenztafel aus, mittels derer die Annahme der Itemunabhängigkeit der Antwortvariablen überprüft werden kann?
- Wie kann die Annahme der Personenunabhängigkeit mittels Kontingenztafeln überprüft werden?

Übungsaufgaben

1. Ein Test aus drei dreistufigen Items wurde von n = 9 Vpn bearbeitet. Die Antwortvektoren der Vpn sind in nachstehender Tabelle dargestellt:

v	1	2	3	4	5	6	7	8	9
(x_{vi})	012	012	101	121	010	111	112	022	212

a) Schätzen Sie die Modellparameter des Pure-Random-Modells,
b) berechnen Sie die Fehlerstreuung der Parameterschätzungen sowie
c) die logarithmierte Likelihood der Datenmatrix und
d) die Anzahl der unabhängigen Modellparameter, die zur Darstellung der Likelihood erforderlich sind!

2. Berechnen Sie aus den Daten in Aufgabe 1
a) Schätzungen der Modellparameter der Ein-Klassen-Lösung der LCA
b) die Fehlerstreuung der Parameterschätzungen sowie
c) die logarithmierte Likelihood der Datenmatrix und
d) die Anzahl der unabhängigen Modellparameter, die zur Darstellung der Likelihood erforderlich sind, und
e) vergleichen Sie das Ergebnis mit dem von Aufgabe 1!

3. Eine Vp hat in einem Test aus 20 dreistufigen Items bei 4 Items null Punkte, bei 7 Items einen Punkt und bei 9 Items zwei Punkte erzielt. Die Antwortvariablen seien itemunabhängig verteilt.
a) Schätzen Sie die Personenparameter und
b) berechnen Sie deren Fehlerstreuung sowie
c) die logarithmierte Likelihood des Antwortvektors der Vp und
d) die Anzahl der unabhängigen Modellparameter, die zur Darstellung der Likelihood erforderlich sind!

4. Gegeben sei ein Test aus 5 binären Items, deren Antwortvariablen itemunabhängig verteilt sind. In einer Stichprobe von n = 100 Vpn ergaben sich die folgenden Scorehäufigkeiten:

$x_{vo}=g$	0	1	2	3	4	5
n_g	10	17	28	24	15	6

a) Schätzen Sie die Personenparameter, welche den Scores entsprechen und berechnen Sie deren Fehlerstreuung!
b) Schätzen Sie die Klassengrößeparameter der Scoregruppen und berechnen Sie deren Fehlerstreuung!
c) Berechnen Sie die logarithmierte Likelihood der Datenmatrix, welche den o. g. Scorehäufigkeiten zugrunde liegt, und
d) die Anzahl der unabhängigen Modellparameter, die zur Darstellung der Likelihood erforderlich sind!

5. Ein Test aus k = 3 dichotomen Items wurde von n = 20 männlichen und n = 20 weiblichen Vpn bearbeitet. Die Antwortmuster der Vpn sind in nachstehender Tabelle dargestellt.

Männliche Vpn				Weibliche Vpn			
Nr.	(x_{vi})	Nr.	(x_{vi})	Nr.	(x_{vi})	Nr.	(x_{vi})
1	000	11	110	21	101	31	101
2	110	12	111	22	011	32	011
3	001	13	110	23	001	33	101
4	101	14	001	24	011	34	100
5	100	15	101	25	110	35	101
6	110	16	111	26	101	36	011
7	001	17	010	27	010	37	101
8	101	18	110	28	011	38	011
9	110	19	001	29	101	39	110
10	111	20	111	30	011	40	011

a) Berechnen Sie die Anzahl der unabhängigen Modellparameter des saturierten Modells und des Pure-Random-Modells!
b) Schätzen Sie die Klassengrößenparameter des saturierten Modells und die Kategorienwahrscheinlichkeiten des Pure-Random-Modells!
c) Überprüfen Sie die Geltung des Pure-Random-Modells hinsichtlich Itemunabhängigkeit und Personenunabhängigkeit der Antwortvariablen mittels Kontingenztafeln!

2.7

Parameterschätzung in Latent-Trait-Modellen

2.7.1 Die unbedingte Maximum-Likelihood-Methode

Unbedingte Maximum-Likelihood-Schätzung der Personenparameter

Im Binomialmodell konnten wir die Personenparameter relativ problemlos schätzen, indem wir die Maximum-Likelihood-Methode auf die Antwortvektoren der Vpn angewendet haben. Es erscheint daher nahe liegend, dieselbe Vorgehensweise auch dann beizubehalten, wenn die Itemcharakteristiken von einem oder sogar mehreren Itemparametern abhängen, so dass

$$f_i(\theta_v) = f(\theta_v, \delta_i)\,. \qquad [2.7.1.1]$$

Die Likelihood des Antwortvektors einer gegebenen Vp lautet dann

$$\begin{aligned} L(x_v|\theta_v) &= \text{prob}\{(x_{v1},\ldots,x_{vk})|\theta_v\} \qquad [2.7.1.2] \\ &= \prod_{i=1}^{k} \text{prob}(x_{vi}|\theta_v) \\ &= \prod_{i=1}^{k} f(\theta_v,\delta_i)^{x_{vi}}\,(1-f(\theta_v,\delta_i))^{1-x_{vi}} \end{aligned}$$

und nimmt z. B. im Rasch-Modell die Form

$$\begin{aligned} L(x_v|\,\theta_v) &= \prod_{i=1}^{k} \frac{(\xi_v\varepsilon_i)^{x_{vi}}}{1+\xi_v\varepsilon_i} \qquad [2.7.1.3] \\ &= \frac{\xi_v^{x_{vo}} \prod_{i=1}^{k} \varepsilon_i^{x_{vi}}}{\prod_{i=1}^{k}(1+\xi_v\varepsilon_i)} \end{aligned}$$

an. Logarithmieren und Nullsetzen der ersten Ableitung nach dem unbekannten Personenparameter θ_v

$$\frac{d\ln\{L(x_v|\theta_v)\}}{d\theta_v} = 0 \qquad [2.7.1.4]$$

liefert schließlich die sog. unbedingte (engl. *unconditional*) ML-Schätzfunktion

$$\hat{\theta}_v = u(x_v, \theta_v, \delta_1, \ldots, \delta_k) \qquad [2.7.1.5]$$

für den Personenparameter, die jedoch nicht nur von θ_v, sondern auch von den Itemparametern $\delta_1,\ldots,\delta_k$ abhängt und z. B. im Rasch-Modell die Form

$$x_{vo} = \sum_{i=1}^{k} \frac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i} \quad \text{für } v = 1,\ldots,n \qquad [2.7.1.6]$$

hat. Zur Auflösung der Schätzgleichung müssen die Itemparameter daher bereits bekannt sein, was jedoch *de facto* nicht der Fall ist.

Gemeinsame Maximum-Likelihood-Schätzung der Item- und Personenparameter

Als Ausweg aus diesem Dilemma bietet sich an, die Itemparameter gemeinsam mit den Personenparametern zu schätzen, indem man nicht nur die Likelihood des Antwortvektors einer gegebenen Person $L(x_v|\theta_v)$, sondern die Likelihood der gesamten Antwortmatrix der jeweils gegebenen Personenstichprobe

$$\begin{aligned} uL(X) &= L(X|\theta_1,\ldots,\theta_n) \qquad [2.7.1.7] \\ &= \prod_{v=1}^{n} L(x_v|\theta_v) \\ &= \prod_{v=1}^{n}\prod_{i=1}^{k} f(\theta_v,\delta_i)^{x_{vi}} \left(1 - f(\theta_v,\delta_i)\right)^{1-x_{vi}} \end{aligned}$$

maximiert. In der Literatur wird diese Likelihood, die im Rasch-Modell die Form

$$\begin{aligned} uL(X) &= \prod_{v=1}^{n}\prod_{i=1}^{k} \frac{(\xi_v \varepsilon_i)^{x_{vi}}}{1 + \xi_v \varepsilon_i} \qquad [2.7.1.8] \\ &= \frac{\prod_{v=1}^{n} \xi_v^{x_{vo}} \prod_{i=1}^{k} \varepsilon_i^{x_{oi}}}{\prod_{v=1}^{n}\prod_{i=1}^{k} (1 + \xi_v \varepsilon_i)} \end{aligned}$$

annimmt, als unbedingte oder totale Likelihood der Antwortmatrix bezeichnet. Logarithmieren und Nullsetzen der ersten partiellen Ableitungen nach den Item- und Personenparametern ergibt schließlich die unbedingten ML-Schätzgleichungen[1]

$$\frac{\partial \ln\{uL(X)\}}{\partial \theta_v} = 0 \quad \text{für } v = 1,\ldots,n \qquad [2.7.1.9]$$

und

$$\frac{\partial \ln\{uL(X)\}}{\partial \delta_i} = 0 \quad \text{für } i = 1,\ldots,k, \qquad [2.7.1.10]$$

die im Rasch-Modell die Form

$$x_{vo} = \sum_{i=1}^{k} \frac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i} \quad \text{für } v = 1,\ldots,n \qquad [2.7.1.11]$$

und

$$x_{oi} = \sum_{v=1}^{n} \frac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i} \quad \text{für } i = 1,\ldots,k \qquad [2.7.1.12]$$

haben.

Herleitung der unbedingten ML-Schätzfunktionen im Rasch-Modell

Logarithmieren der totalen Likelihood in Gleichung [2.7.1.8] ergibt wegen $\xi_v = e^{\theta_v}$ bzw. $\varepsilon_i = e^{-\delta_i}$ und entsprechend $\ln(\xi_v) = \theta_v$ bzw. $\ln(\varepsilon_i) = -\delta_i$ den Ausdruck

$$\ln\{uL(X)\} = \sum_{v=1}^{n} x_{vo}\theta_v - \sum_{i=1}^{k} x_{oi}\delta_i - \sum_{v=1}^{n}\sum_{i=1}^{k} \ln\left(1 + e^{\theta_v - \delta_i}\right) \quad . \qquad [2.7.1.13]$$

Durch Nullsetzen der ersten partiellen Ableitungen nach den unbekannten Modellparametern erhalten wir

$$\frac{\partial \ln\{uL(X)\}}{\partial \theta_v} = x_{vo} - \sum_{i=1}^{k} \frac{e^{\theta_v - \delta_i}}{1 + e^{\theta_v - \delta_i}} = 0 \qquad [2.7.1.14]$$

für die Personenparameter und

1. Für den Parameter einer bestimmten Vp v ergibt sich dabei exakt dieselbe Schätzgleichung wie oben, da die anderen Zeilen der Antwortmatrix von θ_v unabhängig sind.

$$\frac{\partial \ln\{uL(X)\}}{\partial \delta_i} = -x_{oi} + \sum_{v=1}^{n} \frac{e^{\theta_v - \delta_i}}{1 + e^{\theta_v - \delta_i}} = 0 \qquad [2.7.1.15]$$

für die Itemparameter. Diese Gleichungen sind den Gleichungen [2.7.1.11] und [2.7.1.12] äquivalent.

Das Neyman-Scott-Problem

In der Praxis erweist sich die unbedingte ML-Methode jedoch als wenig brauchbar. Wie aus einem Satz von Neyman & Scott (1948) hervorgeht, versagt die Maximum-Likelihood-Methode, wenn die Anzahl der zu schätzenden Parameter nicht gegen einen festen Wert konvergiert, während die Anzahl der Beobachtungen gegen unendlich strebt.

Gerade diese Voraussetzung wird durch die Latent-Trait-Modelle aber verletzt, sobald man die Annahme identischer Itemcharakteristiken fallen lässt. Um die unbedingte Likelihood der $n \times k$ Antworten von n Vpn auf k Testitems zu beschreiben, sind dann (mindestens) n Personen- und k Itemparameter erforderlich und es gibt keine Möglichkeit, wie die Anzahl der Beobachtungen vergrößert werden könnte ohne dass auch die Anzahl der zu schätzenden Parameter $n(P) \geq n + k$ anwächst.

Die Maximum-Likelihood-Methode liefert in diesem Falle nicht einmal konsistente Schätzfunktionen, was z. B. im Rasch-Modell zur Folge hat, dass kleine Parameter unterschätzt und große Parameter überschätzt werden.

2.7.2 Die bedingte Maximum-Likelihood-Methode

Strukturelle und inzidentelle Parameter

Methoden zur Lösung des Neyman-Scott-Problems beruhen darauf, die strukturellen und die inzidentellen Modellparameter voneinander zu trennen und getrennt voneinander zu schätzen.

- Strukturelle Parameter sind in unserem Kontext die Itemparameter. Sie sind von endlicher Anzahl und zu jedem einzelnen von ihnen können potentiell unendlich viele Beobachtungen erhoben werden.
- Die Personenparameter dagegen sind inzidentelle Parameter. Von ihnen gibt es potentiell unendlich viele und zu jedem einzelnen von ihnen liegt immer nur eine endliche Anzahl von Beobachtungen vor.

Bedingte Maximum-Likelihood-Schätzung der Itemparameter

Die eleganteste Methode zur Schätzung der strukturellen Itemparameter ist die sog. bedingte (engl. *conditional*) Maximum-Likelihood-Methode (CML-Methode), die an Stelle der unbedingten

Likelihood uL(X) die bedingte Likelihood der Antwortmatrix bei gegebenen Scores der Vpn

$$cL(X) = L\left(X \middle| x_{1o}, \ldots, x_{no}\right) \qquad [2.7.2.1]$$
$$= \prod_{v=1}^{n} \text{prob}\left\{\left(x_{v1}, \ldots, x_{vk}\right) \middle| x_{vo}\right\}$$

maximiert. Voraussetzung für die Anwendung der CML-Methode ist die Suffizienz der Scores, so dass die bedingte Likelihood von den inzidentellen Personenparametern unabhängig ist. Je nachdem, welche Art von Scores man verwendet, ist die Anwendung der CML-Methode daher entweder an die Geltung des Birnbaum-Modells (gewichtete Summenscores) oder des Rasch-Modells (ungewichtete Summenscores) gebunden.

Im Falle des Rasch-Modells kann die bedingte Likelihood der Antwortmatrix wegen [2.3.3.1] in der Form

$$cL(X) = \prod_{v=1}^{n} \text{prob}\left\{\left(x_{v1}, \ldots, x_{vk}\right) \middle| x_{vo}\right\} \qquad [2.7.2.2]$$
$$= \prod_{v=1}^{n} \frac{\prod_{i=1}^{k} \varepsilon_i^{x_{vi}}}{\gamma_{x_{vo}}(\varepsilon)}$$
$$= \frac{\prod_{i=1}^{k} \varepsilon_i^{x_{oi}}}{\prod_{g=0}^{k} \gamma_g(\varepsilon)^{n_g}}$$

angeschrieben werden. Als unbekannte Parameter treten nur noch k Itemschierigkeitsparameter $\delta_i = -\ln(\varepsilon_i)$ auf. Da die Itemschwierigkeiten im Rasch Modell auf einer Differenzenskala gemessen werden (vgl. Kap. 2.2.3) und somit o.B.d.A. zu

$$\sum_{i=1}^{k} \delta_i = 0 \qquad [2.7.2.3]$$

normiert werden können, sind nur k-1 der Parameter unabhängig voneinander, so dass sich *die Anzahl der unabhängigen Modellparameter*, die zur Darstellung der bedingten Likelihood erforderlich sind, im Rasch-Modell zu

$$n(P_{CR}) = k - 1 \qquad [2.7.2.4]$$

ergibt.

Logarithmieren und Nullsetzen der ersten partiellen Ableitungen nach den Itemparametern liefert schließlich die bedingten ML-Schätzgleichungen

$$\frac{\partial \ln\{cL(X)\}}{\partial \delta_i} = 0 \quad \text{für } i = 1,\ldots,k, \qquad [2.7.2.5]$$

die im Rasch-Modell die Form

$$x_{oi} = \varepsilon_i \sum_{g=1}^{k} n_g \frac{\gamma_{g-1}^{(i)}(\varepsilon)}{\gamma_g(\varepsilon)} \qquad [2.7.2.6]$$

haben, worin $\gamma_{g-1}^{(i)}(\varepsilon)$ wieder die symmetrische Grundfunktion (g-1)-ter Ordnung der Parameter $\varepsilon_1,\ldots,\varepsilon_{i-1},\varepsilon_{i+1},\ldots,\varepsilon_k$ bezeichnet.

Herleitung der bedingten ML-Schätzfunktionen im Rasch-Modell

Logarithmieren der bedingten Likelihood [2.7.2.2] ergibt wegen $\ln(\varepsilon_i) = -\delta_i$

$$\ln\{cL(X)\} = \sum_{i=1}^{k} x_{oi} \ln(\varepsilon_i) - \sum_{g=0}^{k} n_g \ln(\gamma_g(\varepsilon)) \qquad [2.7.2.7]$$

$$= -\sum_{i=1}^{k} x_{oi}\delta_i - \sum_{g=0}^{k} n_g \ln(\gamma_g(\varepsilon))$$

und durch Nullsetzen der ersten partiellen Ableitungen nach den Itemparametern erhalten wir wegen $\gamma_0(\varepsilon) = 1$ und $\varepsilon_i = e^{-\delta_i}$ die bedingten ML-Schätzgleichungen

$$\frac{\partial \ln\{cL(X)\}}{\partial \delta_i} = -x_{oi} - \sum_{g=1}^{k} n_g \frac{\gamma_{g-1}^{(i)}(\varepsilon)}{\gamma_g(\varepsilon)} e^{-\delta_i}(-1) \qquad [2.7.2.8]$$

$$= -x_{oi} + \varepsilon_i \sum_{g=1}^{k} n_g \frac{\gamma_{g-1}^{(i)}(\varepsilon)}{\gamma_g(\varepsilon)} = 0 \ .$$

Diese Gleichung ist der Gleichung [2.7.2.6] äquivalent.

Eigenschaften der bedingten ML-Methode

Der Vorteil der bedingten ML-Methode besteht darin, dass sie nach Andersen (1973) dieselben wünschenswerten Eigenschaften besitzt wie die klassische ML-Methode bei endlicher Parameteranzahl. CML-Schätzer sind nach Andersen (1973) (zumindest) asymptotisch normalverteilt, erwartungstreu und wirksamst.

Entsprechend ergibt sich für die Itemparameter im Rasch-Modell der asymptotische Schätzfehler

$$\sigma(\hat{\delta}_i) = \sqrt{\frac{1}{I(\delta_i)}}, \qquad [2.7.2.9]$$

worin die Informationsfunktion $I(\delta_i)$ die Form

$$I(\delta_i) = \sum_{v=1}^{n} \frac{\xi_v \varepsilon_i}{(1+\xi_v \varepsilon_i)^2} \qquad [2.7.2.10]$$

hat.

Herleitung der Informationsfunktion für die Itemparameter

Die Informationsfunktion ist definiert als

$$I(\delta_i) = E\left\{\sum_{v=1}^{n}\left(\frac{\partial \ln(f(x_{vi},\delta_i))}{\partial \delta_i}\right)^2\right\} \qquad [2.7.2.11]$$

$$= \sum_{v=1}^{n} E\left\{\left(\frac{\partial \ln(f(x_{vi},\delta_i))}{\partial \delta_i}\right)^2\right\}.$$

Einsetzen für die Wahrscheinlichkeitsdichte

$$f(x_{vi},\delta_i) = \frac{(\xi_v \varepsilon_i)^{x_{vi}}}{1+\xi_v \varepsilon_i}$$

ergibt wegen $\xi_v \varepsilon_i = e^{\theta_v - \delta_i}$ und entsprechend $\ln(\xi_v \varepsilon_i) = \theta_v - \delta_i$

$$\ln(f(x_{vi},\delta_i)) = x_{vi}(\theta_v - \delta_i) - \ln\left(1 + e^{\theta_v - \delta_i}\right),$$

so dass die erste partielle Ableitung von $\ln(f(x_{vi},\delta_i))$ nach dem Itemparameter δ_i die Form

$$\frac{\partial \ln(f(x_{vi},\delta_i))}{\partial \delta_i} = x_{vi}(-1) - \frac{e^{\theta_v - \delta_i}}{1+e^{\theta_v - \delta_i}}(-1)$$

$$= \frac{\xi_v \varepsilon_i}{1+\xi_v \varepsilon_i} - x_{vi}$$

$$= \begin{cases} \dfrac{\xi_v \varepsilon_i}{1+\xi_v \varepsilon_i} & \text{für } x_{vi} = 0 \\ \dfrac{-1}{1+\xi_v \varepsilon_i} & \text{für } x_{vi} = 1 \end{cases}$$

annimmt und der Erwartungswert der quadrierten Ableitung gleich

$$E\left\{\left(\frac{\partial \ln(f(x_{vi},\delta_i))}{\partial \delta_i}\right)^2\right\} = \sum_{x_{vi}=0}^{1}\left(\frac{\partial \ln(f(x_{vi},\delta_i))}{\partial \delta_i}\right)^2 f(x_{vi},\delta_i)$$

$$= \frac{(\xi_v\varepsilon_i)^2}{(1+\xi_v\varepsilon_i)^3} + \frac{\xi_v\varepsilon_i}{(1+\xi_v\varepsilon_i)^3}$$

$$= \frac{\xi_v\varepsilon_i(1+\xi_v\varepsilon_i)}{(1+\xi_v\varepsilon_i)^3}$$

$$= \frac{\xi_v\varepsilon_i}{(1+\xi_v\varepsilon_i)^2}$$

ist. Einsetzen in Gleichung [2.7.2.11] ergibt schließlich

$$I(\delta_i) = \sum_{v=1}^{n}\frac{\xi_v\varepsilon_i}{(1+\xi_v\varepsilon_i)^2}\,.$$

2.7.3 Die marginale Maximum-Likelihood-Methode

Die marginale Likelihood

Ein Nachteil der bedingten ML-Methode besteht allerdings darin, dass sie die Existenz suffizienter Statistiken für die inzidentellen Parameter voraussetzt, wodurch die Anwendbarkeit der CML-Methode auf das Rasch-Modell, seine Spezialfälle und seine Verallgemeinerungen beschränkt ist.

Sind die Scores keine erschöpfenden Statistiken für die inzidentellen Personenparameter, dann bleibt zur Lösung des Neyman-Scott-Problems nur noch der Ausweg, die Personenparameter θ als Realisationen einer Zufallsvariable Θ mit der Wahrscheinlichkeitsdichte $g(\theta)$ aufzufassen und an Stelle der bedingten Likelihood die sog. marginale Likelihood

$$mL(X) = L(X) \qquad [2.7.3.1]$$
$$= \prod_{v=1}^{n} \text{prob}\{(x_{v1},\ldots,x_{vk})\}$$

zu maximieren, worin

$$\text{prob}\{(x_{v1},\ldots,x_{vk})\} = \int_{-\infty}^{\infty} \text{prob}\{(x_{v1},\ldots,x_{vk})|\theta\}g(\theta)d\theta \qquad [2.7.3.2]$$

die Wahrscheinlichkeit beschreibt, dass eine zufällig herausgegriffene Vp den Antwortvektor $(x_{v1},\ldots,x_{vk})$ zeigt, so dass die marginale Likelihood gleich der Wahrscheinlichkeit ist, an einer Zufallsstichprobe von n Probanden die Antwortmatrix X zu beobachten.

Eigenschaften der marginalen ML-Methode

Der große Vorteil der marginalen Maximum-Likelihood-Methode (MML-Methode) besteht in ihrer universellen Anwendbarkeit. Dieser wird jedoch durch den entscheidenden Nachteil erkauft, dass sie Zusatzannahmen über die Verteilung der inzidentellen Parameter treffen muss, die zu einer Verzerrung der Schätzfunktionen führen können.

Eine Lösung dieses Problems ergibt sich erneut erst dann, wenn suffiziente Statistiken für die inzidentellen Parameter existieren. Ist dies der Fall, so liefert die marginale ML-Methode exakt dieselben Schätzfunktionen wie die bedingte ML-Methode.

Beweis

Sind die Scores x_{vo} erschöpfenden Statistiken für die inzidentellen Personenparameter, so kann die Wahrscheinlichkeit [2.7.3.2] in der Form

$$\text{prob}\{(x_{v1},\ldots,x_{vk})\} = \int_{-\infty}^{\infty} \text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo}\}\text{prob}(X_{vo}|\theta)g(\theta)d\theta \qquad [2.7.3.3]$$

$$= \text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo}\}\int_{-\infty}^{\infty} \text{prob}(x_{vo}|\theta)g(\theta)d\theta$$

$$= \text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo}\}\text{prob}(x_{vo})$$

dargestellt werden, worin $\text{prob}(x_{vo})$ die Wahrscheinlichkeit bezeichnet, dass eine zufällig herausgegriffene Vpn den Score x_{vo} zeigt. Die marginale Likelihood [2.7.3.1] nimmt dann die Form

$$mL(X) = \prod_{v=1}^{n} \text{prob}\{(x_{v1},\ldots,x_{vk})\} \qquad [2.7.3.4]$$

$$= \prod_{v=1}^{n} \text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo}\}\text{prob}(x_{vo})$$

$$= \prod_{v=1}^{n} \text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo}\}\prod_{v=1}^{n}\text{prob}(x_{vo})$$

an. Laut Gleichung [2.7.2.1] ist aber $\prod_{v=1}^{n} \text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo}\} = cL(X)$, so dass die marginale Likelihood in zwei Faktoren

$$mL(X) = cL(X)\prod_{g=0}^{k} prob(x_{vo} = g)^{n_g} \qquad [2.7.3.5]$$

zerfällt, deren einer die bedingte Likelihood der Antwortmatrix ist und deren anderer mit der Definition $p_g = prob(x_{vo} = g)$ unabhängig von den Itemparametern dargestellt werden kann, so dass

$$mL(X) = cL(X)\prod_{g=0}^{k} p_g^{n_g} \ . \qquad [2.7.3.6]$$

Logarithmieren ergibt

$$\ln\{mL(X)\} = \ln\{cL(X)\} + \sum_{g=0}^{k} n_g \ln(p_g), \qquad [2.7.3.7]$$

worin die Summe auf der rechten Seite von den Itemparametern unabhängig ist und daher beim Differenzieren wegfällt.
Die MML-Methode führt somit zu exakt denselben Schätzgleichungen für die Itemparameter wie die CML-Methode.[1]

Anzahl der unabhängigen Modellparameter

Während die marginale ML-Methode also im Falle des Rasch-Modells, seiner Spezialfälle und seiner Verallgemeinerungen zu den selben Schätzfunktionen führt wie die bedingte Maximum-Likelihood-Methode, ist zur Darstellung der marginalen Likelihood jedoch eine größere Anzahl von Modellparametern erforderlich. So kommen z. B. beim Rasch-Modell zu den k-1 unabhängigen Itemparametern noch k+1 Klassengrößeparameter p_g hinzu, die ebenfalls aus den Daten geschätzt werden müssen. Wegen

$$\sum_{g=0}^{k} p_g = 1 \qquad [2.7.3.8]$$

sind aber nur k Klassengrößeparameter voneinander unabhängig, so dass sich die *Anzahl der unabhängigen Modellparameter*, die zur Darstellung der marginalen Likelihood erforderlich sind, insgesamt zu

$$n(P_{MR}) = 2k - 1 \qquad [2.7.3.9]$$

ergibt.

1. Wie aus einem Theorem von Pfanzagl (1993) hervorgeht, sind die MML-Schätzer den CML-Schätzern auch dann noch zumindest *asymptotisch äquivalent*, wenn man darauf verzichtet, die marginale Likelihood gemäß Gleichung [2.7.3.4] in ihre Bestandteile zu zerlegen und sie stattdessen unter Zusatzannahmen über die Dichtefunktion der Personenparameterverteilung nach Formel [2.7.3.2] berechnet.

ML-Schätzung der Klassengrößeparameter

Die Klassengrößeparameter p_g werden mittels Maximum-Likelihood-Methode durch die relative Häufigkeit

$$\hat{p}_g = \frac{n_g}{n} \quad \text{für } g = 0,\ldots,k \qquad [2.7.3.10]$$

geschätzt, wobei der Schätzfehler

$$\sigma(\hat{p}_g) = \sqrt{\frac{p_g(1-p_g)}{n}} \qquad [2.7.3.11]$$

beträgt.

2.7.4 Die gewichtete Maximum-Likelihood-Methode

Bedingte ML-Schätzung der Personenparameter

Wegen der Symmetrie zwischen Item- und Personenparametern im Rasch-Modell ist es grundsätzlich möglich, auch die Personenparameter mittels der CML-Methode zu schätzen, indem wir die ML-Methode auf die bedingte Likelihood der Antwortmatrix bei gegebenen Item-Lösungshäufigkeiten

$$L(X|x_{o1},\ldots,x_{ok}) = \frac{\prod_{v=1}^{n} \xi_v^{x_{vo}}}{\prod_{q=0}^{n} \gamma_q(\xi)^{n_q}} \qquad [2.7.4.1]$$

anwenden, worin n_q die Anzahl der Items ist, die von $x_{oi} = q$ Probanden gelöst wurden. Analog zur CML-Schätzung der Itemparameter erhalten wir dann die Schätzgleichungen

$$x_{vo} = \xi_v \sum_{q=1}^{n} n_q \frac{\gamma_{q-1}^{(v)}(\xi)}{\gamma_q(\xi)} \quad \text{für } v = 1,\ldots,n\ , \qquad [2.7.4.2]$$

den asymptotischen Schätzfehler

$$\sigma(\hat{\theta}_v) = \sqrt{\frac{1}{I(\theta_v)}} \qquad [2.7.4.3]$$

und die Informationsfunktion

$$I(\theta_v) = \sum_{i=1}^{k} \frac{\xi_v \varepsilon_i}{(1+\xi_v \varepsilon_i)^2}\ . \qquad [2.7.4.4]$$

Der Beweis sei dem Leser als Übungsaufgabe überlassen.

Unbedingte ML-Schätzung der Personenparameter bei bekannten Itemparametern

In großen Stichproben scheitert die Berechnung der CML-Schätzfunktionen für die Personenparameter jedoch an rechentechnischen Problemen. Die Berechnung der symmetrischen Grundfunktionen $\gamma_q(\xi)$ gelingt nicht mit hinreichender Rechengenauigkeit.

In der Praxis hat sich daher eine zweistufige Vorgehensweise durchgesetzt, bei der zuerst die Itemparameter mittels CML- oder MML-Methode geschätzt werden. Die Itemschwierigkeiten als bekannt voraussetzend, werden anschließend die Personenparameter geschätzt, wofür man sich zunächst der in Kapitel 2.7.1 hergeleiteten unbedingten ML-Schätzfunktionen

$$x_{vo} = \sum_{i=1}^{k} \frac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i} \quad \text{für } v = 1,\ldots,n \qquad [2.7.4.5]$$

bedient hat. Dadurch ist das Neyman-Scott-Problem jedoch nicht behoben. Selbst wenn die Itemparameter bekannt sind, bleiben die unbedingten ML-Schätzer inkonsistent. Kleine Personenparameter werden zu klein und große zu groß geschätzt. Für Scores von $x_{vo} = 0$ bzw. $x_{vo} = k$ hat die Gleichung [2.7.4.5] überhaupt keine endliche Lösung und die Parameterschätzungen divergieren nach $-\infty$ bzw. $+\infty$.

Gewichtete ML-Schätzung der Personenparameter

Dieses Manko kann durch die auf Warm (1989) zurückgehende gewichtete (engl. *weighted*) Maximum-Likelihood-Methode (WML-Methode) behoben werden, die an Stelle der Likelihood des Antwortvektors bei gegebenem Personenparameter $L(x_v \mid \theta_v)$, die Likelihood des Personenparameters bei gegebenem Antwortvektor $L(\theta_v \mid x_v)$ maximiert und zu den Schätzgleichungen

$$x_{vo} + \frac{\sum_{i=1}^{k} p_{vi}(1 - p_{vi})(1 - 2p_{vi})}{2\sum_{i=1}^{k} p_{vi}(1 - p_{vi})} = \sum_{i=1}^{k} \frac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i} \quad \text{für } v = 1,\ldots,n \qquad [2.7.4.6]$$

führt. Diese unterscheiden sich von [2.7.4.5] durch einen Korrekturterm, der für kleine θ positive und für große θ negative Werte annimmt und somit kleine Scores nach oben und große Scores nach unten korrigiert.

Ein Nachteil der WML-Methode besteht allerdings darin, dass zur Konstruktion der Schätzfunktionen Annahmen über die Wahrscheinlichkeitsdichte der latenten Personenvariable Θ getroffen werden müssen, wofür Warm die Quadratwurzel aus der Informationsfunktion

$$g(\theta_v) = \sqrt{I(\theta_v)} \qquad [2.7.4.7]$$

einsetzt.

Herleitung der gewichteten ML-Schätzfunktionen für die Personenparameter im Rasch-Modell

Aufgrund des Bayes-Theorems

$$\text{prob}(B|A) = \text{prob}(A|B)\frac{\text{prob}(B)}{\text{prob}(A)} \qquad [2.7.4.8]$$

hat die Likelihood des Personenparameters bei gegebenem Antwortvektor die Form

$$L(\theta_v|x_v) = L(x_v|\theta_v)\frac{L(\theta_v)}{L(x_v)}, \qquad [2.7.4.9]$$

worin $L(\theta_v) = g(\theta_v)$ gleich der Wahrscheinlichkeitsdichte der latenten Personenvariable an der Stelle θ_v ist und $L(x_v) = \text{prob}(x_v)$ unabhängig von den Personenparametern dargestellt werden kann. $L(x_v|\theta_v)$ ist wegen [2.7.1.3] gleich

$$L(x_v|\theta_v) = \frac{\xi_v^{x_{vo}} \prod_{i=1}^{k} \varepsilon_i^{x_{vi}}}{\prod_{i=1}^{k}(1+\xi_v\varepsilon_i)}, \qquad [2.7.4.10]$$

so dass

$$L(\theta_v|x_v) = \frac{\xi_v^{x_{vo}} \prod_{i=1}^{k} \varepsilon_i^{x_{vi}}}{\prod_{i=1}^{k}(1+\xi_v\varepsilon_i)} \frac{g(\theta_v)}{\text{prob}(x_v)} \qquad [2.7.4.11]$$

$$= \frac{\xi_v^{x_{vo}}}{\prod_{i=1}^{k}(1+\xi_v\varepsilon_i)} g(\theta_v) \frac{\prod_{i=1}^{k} \varepsilon_i^{x_{vi}}}{\text{prob}(x_v)}.$$

Die Likelihood $L(\theta_v|x_v)$ ist daher proportional zu dem Ausdruck auf der rechten Seite von

$$L(\theta_v \mid x_v) \approx \frac{\xi_v^{x_{vo}}}{\prod_{i=1}^{k}(1+\xi_v\varepsilon_i)}\, g(\theta_v). \qquad [2.7.4.12]$$

Logarithmieren ergibt wegen $\ln(\xi_v) = \theta_v$

$$\ln\{L(\theta_v \mid x_v)\} \approx x_{vo}\theta_v - \sum_{i=1}^{k}\ln(1+\xi_v\varepsilon_i) + \ln\{g(\theta_v)\}. \qquad [2.7.4.13]$$

Durch Nullsetzen der ersten Ableitung nach dem unbekannten Personenparameter erhalten wir die Schätzgleichung

$$x_{vo} - \sum_{i=1}^{k}\frac{\xi_v\varepsilon_i}{1+\xi_v\varepsilon_i} + \frac{d\ln\{g(\theta_v)\}}{d\theta_v} = 0\,, \qquad [2.7.4.14]$$

zu deren Berechnung noch die 1. Ableitung der Wahrscheinlichkeitsdichte $g(\theta_v)$ gebildet werden muss.

Setzen wir für $g(\theta_v)$ die Wurzel aus der Informationsfunktion ein, so erhalten wir wegen [2.7.4.4]

$$g(\theta_v) = \left(\sum_{i=1}^{k}\frac{\xi_v\varepsilon_i}{(1+\xi_v\varepsilon_i)^2}\right)^{\frac{1}{2}} \qquad [2.7.4.15]$$

bzw.

$$\ln\{g(\theta_v)\} = \frac{1}{2}\ln\left(\sum_{i=1}^{k}\xi_v\varepsilon_i(1+\xi_v\varepsilon_i)^{-2}\right), \qquad [2.7.4.16]$$

so dass

$$\frac{d\ln\{g(\theta_v)\}}{d\theta_v} = \frac{1}{2}\,\frac{\sum_{i=1}^{k}\xi_v\varepsilon_i(1+\xi_v\varepsilon_i)^{-2} + \xi_v\varepsilon_i(-2)(1+\xi_v\varepsilon_i)^{-3}\xi_v\varepsilon_i}{\sum_{i=1}^{k}\xi_v\varepsilon_i(1+\xi_v\varepsilon_i)^{-2}} \qquad [2.7.4.17]$$

$$= \frac{\sum_{i=1}^{k}\frac{\xi_v\varepsilon_i}{(1+\xi_v\varepsilon_i)^2} - 2\frac{(\xi_v\varepsilon_i)^2}{(1+\xi_v\varepsilon_i)^3}}{2\sum_{i=1}^{k}\frac{\xi_v\varepsilon_i}{(1+\xi_v\varepsilon_i)^2}}$$

$$= \frac{\sum_{i=1}^{k}\frac{\xi_v\varepsilon_i}{(1+\xi_v\varepsilon_i)^2}\left(1 - 2\frac{\xi_v\varepsilon_i}{1+\xi_v\varepsilon_i}\right)}{2\sum_{i=1}^{k}\frac{\xi_v\varepsilon_i}{(1+\xi_v\varepsilon_i)^2}}$$

$$= \frac{\sum_{i=1}^{k} p_{vi}(1-p_{vi})(1-2p_{vi})}{2\sum_{i=1}^{k} p_{vi}(1-p_{vi})}.$$

Einsetzen von [2.7.4.17] in [2.7.4.14] ergibt schließlich

$$x_{vo} - \sum_{i=1}^{k} \frac{\xi_v \varepsilon_i}{1+\xi_v \varepsilon_i} + \frac{\sum_{i=1}^{k} p_{vi}(1-p_{vi})(1-2p_{vi})}{2\sum_{i=1}^{k} p_{vi}(1-p_{vi})} = 0.$$

Diese Gleichung ist der Gleichung [2.7.4.6] äquivalent.

Kontrollfragen

- Worin besteht das Neyman-Scott Problem und wie wirkt es sich auf die Schätzung der Modellparameter im Rasch-Modell aus?
- Worin besteht der Unterschied zwischen strukturellen und inzidentellen Parametern?
- Welche Methoden zur Lösung des Neyman-Scott Problems gibt es und an welche Voraussetzungen sind sie gebunden?
- Worin besteht die bedingte ML-Methode?
- Welche wünschenswerten Eigenschaften besitzt die bedingte ML-Methode?
- Worin besteht die marginale ML-Methode?
- Welche Vor- und Nachteile besitzt die bedingte ML-Methode im Vergleich zur marginalen ML-Methode?
- Warum sind die beiden Methoden im Falle des Rasch-Modells äquivalent?
- Welche Methoden zur Schätzung der Personenparameter des Rasch-Modells gibt es und welches sind ihre Vor- und Nachteile?
- Welcher grundlegende Unterschied besteht zwischen der unbedingten und der bedingten ML-Schätzung der Personenparameter einerseits und der gewichteten ML-Schätzung der Personenparameter andererseits?

Übungsaufgaben

1. Die Bearbeitung eines Tests aus k = 3 binären Items mit den Itemparametern δ_1 = -0.4, δ_2 = 0 und δ_3 = 0.4 durch n = 4 Vpn mit den Personenparametern $\theta_1 = \theta_2$ = -0.53 und $\theta_3 = \theta_4$ = 0.53 ergab die folgenden Antwortvektoren

v	1	2	3	4
(x_{vi})	010	100	110	101

Berechnen Sie die den Logarithmus der unbedingten Likelihood der entsprechenden Antwortmatrix!

2. Ein Test aus k = 3 binären Items wurde von n = 11 Vpn bearbeitet. Die Score- und Itemlösungshäufigkeiten sowie die ihnen entsprechenden Itemparameter sind in den nachstehenden Tabellen wiedergegeben.

$x_{vo = g}$	n_g	θ_g
0	1	-1,96
1	3	-0,52
2	5	0,52
3	2	1,96

i	x_{oi}	δ_i
1	7	-0.26
2	6	0,13
3	6	0,13

Berechnen Sie
a) den Logarithmus der bedingten Likelihood der Antwortmatrix,
b) die Anzahl der unabhängigen Modellparameter, die zur Darstellung der bedingten Likelihood erforderlich sind,
c) die Informationsfunktion der Itemparameter und
d) den asymptotischen Schätzfehler der Itemparameter!

3. Berechnen Sie aus den Daten in Aufgabe 2
a) den Logarithmus der marginalen Likelihood der Antwortmatrix und
b) die Anzahl der unabhängigen Modellparameter, die zur Darstellung der marginalen Likelihood erforderlich sind!
4. Zeigen Sie, dass die Gleichungen [2.7.4.1], [2.7.4.2] und [2.7.4.4] im Rasch-Modell erfüllt sind!
5. Berechnen Sie aus den Daten in Aufgabe 2
a) die Informationsfunktion der Personenparameter und
b) den asymptotischen Schätzfehler der Personenparameter!
6. Zeigen Sie, dass der Korrekturterm in [2.7.4.6] für kleine θ positive und für große θ negative Werte annimmt!

2.8

Modellkontrolle und Beurteilung der Testleistung der Probanden im Rasch-Modell

2.8.1 Eigenschaften des Rasch-Modells

Separierbarkeit der Modellparameter

Zwei Voraussetzungen sind notwendig und hinreichend dafür, dass die Itemcharakteristiken in Form des Rasch-Modells

$$f_i(\theta_v) = \frac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i} \text{ mit } \xi_v = e^{\theta_v} \text{ und } \varepsilon_i = e^{-\delta_i} \qquad [2.8.1.1]$$

dargestellt werden können: Erstens die lokale Unabhängigkeit der Testitems, wonach

$$\text{prob}\{(x_{v1}, \ldots, x_{vk})\} = \prod_{i=1}^{k} \text{prob}(x_{vi}), \qquad [2.8.1.2]$$

und zweitens die Suffizienz des Summenscores, welche die Separierung der Modellparameter ermöglicht, so dass die Zeilen der Antwortmatrix bei gegebener Zeilensumme

$$\text{prob}\{(x_{v1}, \ldots, x_{vk}) | x_{vo} = g\} = \frac{\prod_{i=1}^{k} \varepsilon_i^{x_{vi}}}{\gamma_g(\varepsilon)} \qquad [2.8.1.3]$$

unabhängig von den Personenparametern und die Spalten der Antwortmatrix bei gegebener Spaltensumme

$$\text{prob}\{(x_{1i}, \ldots, x_{ni}) | x_{oi} = q\} = \frac{\prod_{v=1}^{n} \xi_v^{x_{vi}}}{\gamma_q(\xi)} \qquad [2.8.1.4]$$

unabgängig von den Itemparametern verteilt sind.

Dies hat eine Reihe von Konsequenzen, die das Rasch-Modell in pragmatischer, statistischer und wissenschaftstheoretischer Hin-

sicht gegenüber allen anderen denkbaren Testmodellen auszeichnen:

Scorebildung

Pragmatisch: Das Rasch-Modell ist notwendig und hinreichend dafür, dass die gesamte Information über die latente Personenvariable in den Summenscores der Probanden enthalten ist. Sind die Modellannahmen des Rasch-Modells verletzt, so ist die Verwendung des Summenscores mit einem Informationsverlust verbunden. Dieser kann so weit gehen, dass die in den Antworten der Probanden enthaltene diagnostisch relevante Information gänzlich verloren geht.

Parameterschätzung

Statistisch: Das Rasch-Modell ist notwendig und hinreichend für die Lösung des Neyman-Scott-Problems mittels der bedingten Maximum-Likelihood-Methode. Folgen die Itemcharakteristiken einer anderen Funktionsform, so ist die CML-Methode nicht anwendbar. Zur Schätzung der Modellparameter stehen dann nur die unbedingte und/oder die marginale Maximum-Likelihood-Methode zur Verfügung. Die UML-Methode liefert jedoch keine konsistenten Schätzfunktionen und die MML-Methode macht empirisch gehaltvolle Zusatzannahmen bezüglich der Verteilung der latenten Personenvariable erforderlich.[1] Sind diese verletzt, so führt dies erneut zum Versagen der Schätzmethode.[2]

Spezifische Objektivität

Wissenschaftstheoretisch: Neben diesen statistischen und pragmatischen Konsequenzen hat die Separierbarkeit der Parameter auch eine grundlegende wissenschaftstheoretische Bedeutung, wofür Rasch (1968) den Terminus spezifische Objektivität geprägt hat. Das Rasch-Modell ist notwendig und hinreichend für die wechselseitige Unabhängigkeit der Vergleiche zwischen Messobjekten (Vpn) und Messinstrumenten (Items): Die Aussagen, welche über die Relationen zwischen n = 1,2,3... Vpn gewonnen werden können, sind davon unabhängig, welche Items ausgewählt und dem Vergleich zugrundegelegt werden.[3] Umgekehrt sind die Aussagen, welche über die Relation zwischen k = 1,2,3... Items gewonnen werden können, unabhängig davon, anhand welcher Personenstichprobe sie gewonnen werden.

1. Nur im Rasch-Modell kann die MML-Methode ohne Zusatzannahmen über die Verteilung der Personenparameter angewendet werden. Sie liefert dann exakt dieselben Schätzfunktionen wie die CML-Methode.
2. Im Rasch-Modell ist die MML-Methode auch in diesem Falle der CML-Methode immerhin noch asymptotisch äquivalent.
3. Vorausgesetzt, die Items messen alle dieselbe latente Eigenschaft Θ.

2.8.2 Kontrolle der spezifischen Objektivität

Stichprobenunabhängigkeit der Itemparameterschätzungen

Diese grundlegende Eigenschaft des Rasch-Modells können wir nun benutzen, um die empirische Geltung der Modellannahmen zu überprüfen.

Zu diesem Zwecke zerlegen wir die Personenstichprobe in h einander nicht überlappende Teilstichproben: z. B. eine Stichprobe von Minderbegabten, eine Stichprobe von Normalbegabten und eine Stichprobe von Hochbegabten. Sind die Modellvoraussetzungen des Rasch-Modells erfüllt und damit spezifisch objektive – d. h. von der Wahl der Personenstichprobe unabhängige – Messungen der Itemschwierigkeiten gewährleistet, dann müssen die aus den Teilstichproben gewonnenen CML-Schätzer der Itemparameter $\delta_i^{(1)}$, $\delta_i^{(2)}$, ..., $\delta_i^{(h)}$ untereinander und mit den aus der Gesamtstichprobe gewonnenen $\delta_i^{(G)}$ übereinstimmen, so dass

$$\delta_i^{(1)} = \delta_i^{(2)} = \ldots = \delta_i^{(h)} = \delta_i^{(G)}. \qquad [2.8.2.1]$$

Graphische Modellkontrolle

Trägt man die Parameterschätzungen in ein zweidimensionales Streuungsdiagramm ein, müssen die Items daher im Falle der Modellgeltung alle auf einer Geraden liegen, die mit einem Anstieg von 45° durch den Ursprung des Koordinatensystems führt (vgl. Abb. 2.8.2.1).

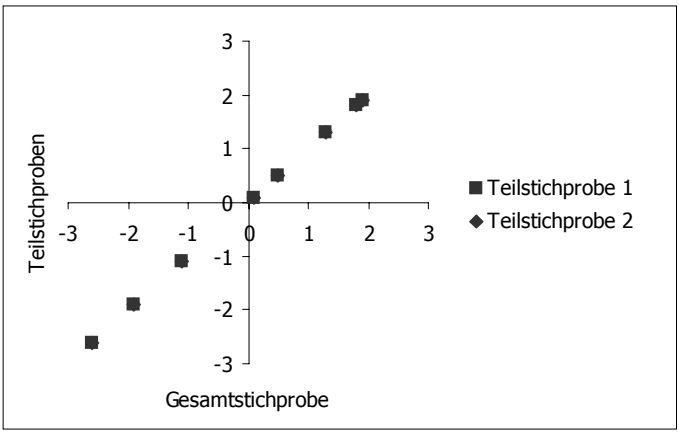


Abb. 2.8.2.1: Graphische Modellkontrolle des Rasch-Modells. Ideale Modellgeltung

In der Praxis werden wir – je nach Stichprobengröße – jedoch mehr oder minder starke Abweichungen von der 45°-Geraden finden (vgl. Abb. 2.8.2.2), so dass sich die Frage ergibt, ob es sich dabei lediglich um Zufallsfehler handelt, oder ob die Abweichun-

gen statistisch signifikant sind und das Rasch-Modell verworfen werden muss.

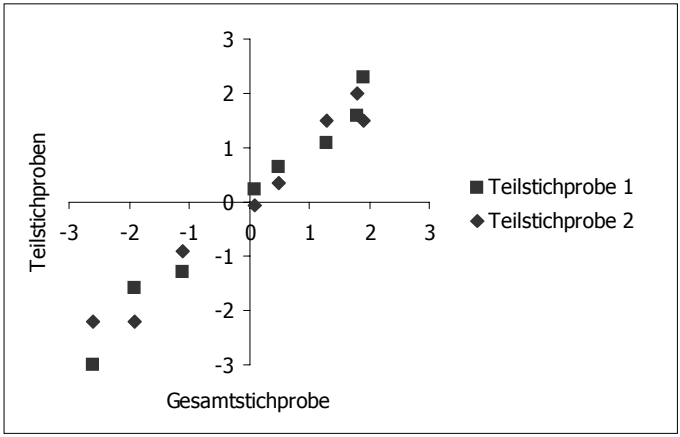


Abb. 2.8.2.2: Graphische Modellkontrolle des Rasch-Modells. Geringfügige Zufallsfehler

Bedingter Likelihood-Quotienten-Test

Um dies zu prüfen, kann man sich eines Likelihood-Quotienten-Tests (Andersen, 1973) bedienen, der die bedingte Likelihood der Datenmatrix unter Annahme der Modellgeltung (H_0: gleiche Itemparameter in allen Subpopulationen von Probanden)

$$L_0 = cL\left(X \middle| \delta_1^{(G)}, \ldots, \delta_k^{(G)}\right) \qquad [2.8.2.2]$$

und die bedingte Likelihood der Datenmatrix unter Annahme der Alternativhypothese (H_1: verschiedene Itemparameter in den verschiedenen Subpopulationen)

$$L_1 = \prod_{g=1}^{h} cL\left(X^{(g)} \middle| \delta_1^{(g)}, \ldots, \delta_k^{(g)}\right) \qquad [2.8.2.3]$$

einander gegenüberstellt, worin $X^{(g)}$ die Antwortmatrix der g-ten Teilstichprobe bezeichnet. Wegen $L_0 \leq L_1$ nimmt der Likelihood-Quotient

$$\lambda = \frac{L_0}{L_1} \qquad [2.8.2.4]$$

stets Werte zwischen Null und Eins an, wobei λ genau dann gleich 1 ist, wenn die Parameterschätzungen aus den Teilstichproben untereinander und mit den Parameterschätzungen aus

der Gesamtstichprobe exakt übereinstimmen. Die asymptotisch nach χ^2 mit

$$df = (h-1)(k-1) \qquad [2.8.2.5]$$

Freiheitsgraden verteilte Testgröße

$$-2\ln(\lambda) = 2\left|\ln(L_0) - \ln(L_1)\right| \qquad [2.8.2.6]$$

nimmt daher umso größere Werte an, je stärker die Modellannahmen des Rasch-Modells verletzt sind. Man wird die Annahme der Modellgeltung daher verwerfen, wenn die Prüfgröße $-2\ln(\lambda)$ auf einem Signifikanzniveau von $\alpha = 5\%$ den kritischen Wert der χ^2-Verteilung mit df = (h-1)(k-1) Freiheitsgraden übersteigt.

Konstruktion der Teilstichproben

Die Konstruktion der Teilstichproben kann entweder aufgrund eines externen Kriteriums – z. B. Schulbildung, Alter, Geschlecht – oder aber aufgrund des Summenscores der Probanden erfolgen. Vorausgesetzt ist lediglich, dass jede Vp in genau einer der Teilstichproben enthalten ist, so dass deren Schnittmengen leer sind und ihre Vereinigungsmenge die Gesamtstichprobe ergibt.

Scoregruppen

Werden die Teilstichproben aufgrund der Scores der Vpn gebildet, so ergeben sich zunächst k+1 Scoregruppen mit $x_{vo} = 0,\ldots,k$, von denen jedoch nur k-1 zur bedingten Likelihood der Datenmatrix beitragen. Mit Scores von $x_{vo} = 0$ und $x_{vo} = k$ ist jeweils nur ein möglicher Antwortvektor verträglich, so dass die bedingte Likelihood des Antwortvektors bei gegebenem Score in diesen Fällen gleich Eins ist.

Zur Berechnung des Likelihood-Quotienten-Tests stützt man sich daher auf eine reduzierte Gesamtstichprobe, in welcher die Vpn mit Scores von 0 oder k nicht enthalten sind. Diese wird dann in h = k-1 Teilstichproben mit Scores von $x_{vo} = 1,\ldots,k-1$ zerlegt, so dass sich die Freiheitsgrade der asymptotischen Prüfverteilung zu df = (k-2)(k-1) ergeben.

Mediansplit

Damit die Prüfverteilung durch die χ^2-Verteilung hinreichend gut approximiert wird, sind allerdings sehr große Teilstichproben erforderlich. Um dies zu gewährleisten, kann man z. B. so vorgehen, dass man auf die Zerlegung der (reduzierten) Gesamtstichprobe in ihre einzelnen Scoregruppen verzichtet und sie stattdessen lediglich in h = 2 Teilstichproben mit

$$g = \begin{cases} 1 \text{ für } x_{vo} \leq Md \ (\text{"Low} - \text{Score} - \text{Gruppe"}) \\ 2 \text{ für } x_{vo} > Md \ (\text{"High} - \text{Score} - \text{Gruppe"}) \end{cases} \qquad [2.8.2.7]$$

aufspaltet, worin Md den Median der Scoreverteilung bezeichnet.

Reproduzierbarkeit der Itemlösungshäufigkeiten

Auf der Zerlegung der (reduzierten) Gesamtstichprobe in die einzelnen Scoregruppen beruht auch der folgende Test, der die Itemlösungshäufigkeiten innerhalb der Scoregruppen g = 1,…,k-1

$$x_{gi} = \sum_{v: x_{vo} = g} x_{vi} \qquad [2.8.2.8]$$

mit den bei Geltung des Rasch-Modells erwarteten Lösungshäufigkeiten

$$e_{gi} = n_g \, \text{prob}\left(x_{vi} = 1 \middle| x_{vo} = g\right) \qquad [2.8.2.9]$$

vergleicht, worin

$$\text{prob}\left(x_{vi} = 1 \middle| x_{vo} = g\right) = \frac{\varepsilon_i \, \gamma_{g-1}^{(i)}(\varepsilon)}{\gamma_g(\varepsilon)} . \qquad [2.8.2.10]$$

Die asymptotisch nach χ^2 mit

$$df = (k-2)(k-1) \qquad [2.8.2.11]$$

verteilte Prüfgröße

$$\chi_o^2 = \sum_{g=1}^{k-1} \sum_{i=1}^{k} \frac{\left(x_{gi} - e_{gi}\right)^2}{e_{gi}} \qquad [2.8.2.12]$$

nimmt umso größere Werte an, je stärker die empirischen Itemlösungshäufigkeiten von den bei Geltung des Rasch-Modells erwarteten Häufigkeiten abweichen.[1]

Parameterfreie Modellkontrolle

Zerlegt man die reduzierte Antwortmatrix anhand eines externen Kriteriums in h Teilstichproben, so erlaubt die in den Gleichungen [2.8.2.9] und [2.8.2.10] dargestellte Beziehung darüber hinaus die Konstruktion eines parameterfreien Modelltests, der erstmals bereits von Scheiblechner (1970) erwähnt wurde.

1. Die Herleitung der Gleichung [2.8.2.10] sei dem Leser als Übungsaufgabe überlassen.

Zu diesem Zweck erstellen wir für jede der Scoregruppen g = 1,…,k-1 getrennt eine Matrix der Itemlösungshäufigkeiten $x_{gi}^{(q)}$ innerhalb der Teilstichproben q = 1,…,h, wie sie in Tabelle 2.8.2.1 dargestellt ist.

Tab. 2.8.2.1: Zerlegung der Scoregruppen anhand eines externen Kriteriums

	1	…	i	…	k	
1	$x_{g1}^{(1)}$	…	$x_{gi}^{(1)}$	…	$x_{gk}^{(1)}$	$n_g^{(1)}$
:	:		:		:	:
q	$x_{g1}^{(q)}$	…	$x_{gi}^{(q)}$	…	$x_{gk}^{(q)}$	$n_g^{(q)}$
:	:		:		:	:
h	$x_{g1}^{(h)}$	…	$x_{gi}^{(h)}$	…	$x_{gk}^{(h)}$	$n_g^{(h)}$
Σ	x_{g1}	…	x_{gi}	…	x_{gk}	n_g

Sind die Modellannahmen des Rasch-Modells erfüllt, dann müssen die bedingten Lösungshäufigkeiten $x_{gi}^{(q)}$ wegen [2.8.2.10] proportional zu den unbedingten Lösungshäufigkeiten x_{gi} sein, so dass sich die erwarteten Häufigkeiten

$$e_{gi}^{(q)} = x_{gi} \frac{n_g^{(q)}}{n_g} \qquad [2.8.2.13]$$

ergeben, worin $n_g^{(q)}$ die Anzahl der Probanden bezeichnet, die der q-ten Teilstichprobe angehören und deren Score gleich g ist.

Auf diese Weise kann man für jede Scoregruppe g = 1,…,k-1 eine asymptotisch nach χ^2 mit

$$df_g = (h-1)(k-1) \qquad [2.8.2.14]$$

Freiheitsgraden verteilte Statistik

$$\chi_g^2 = \sum_{q=1}^{h} \sum_{i=1}^{k} \frac{\left(x_{gi}^{(q)} - e_{gi}^{(q)}\right)^2}{e_{gi}^{(q)}} \qquad [2.8.2.15]$$

berechnen. Wegen der Additivität der χ^2-Verteilung können die χ_g^2-Statistiken schließlich zu der Prüfgröße

$$\chi_o^2 = \sum_{g=1}^{k-1} \chi_g^2 \qquad [2.8.2.16]$$

zusammengefasst und gegen die χ^2-Verteilung mit

$$df_o = \sum_{g=1}^{k-1} df_g = (h-1)(k-1)^2 \qquad [2.8.2.17]$$

Freiheitsgraden getestet werden.

Anwendungsvoraussetzungen

Voraussetzung für die Anwendung des Pearson-χ^2-Verfahrens sind hinreichend große Personenstichproben. Pfanzagl (1966) gibt an, dass die Approximation der Prüfverteilung durch die χ^2-Verteilung erst dann hinreichend gut ist, wenn keiner der Erwartungswerte kleiner als Eins und höchstens ein Fünftel der Erwartungswerte kleiner als Fünf ist.

Ist diese Voraussetzung verletzt, so gibt es zwei Auswege:

Datenaggregation

Der erste Ausweg besteht darin, die Prüfgröße χ_o^2 aus aggregierten Daten zu berechnen, z. B. aus den zusammengefassten Häufigkeiten aller Scoregruppen, was im Falle der parameterfreien Modellkontrolle zu der asymptotisch nach χ^2 mit

$$df = (h-1)(k-1) \qquad [2.8.2.18]$$

Freiheitsgraden verteilten Prüfstatistik

$$\chi_o^2 = \sum_{q=1}^{h} \sum_{i=1}^{k} \frac{\left(x_{oi}^{(q)} - e_{oi}^{(q)}\right)^2}{e_{oi}^{(q)}} \qquad [2.8.2.19]$$

führt, worin

$$x_{oi}^{(q)} = \sum_{g=1}^{k-1} x_{gi}^{(q)} \qquad [2.8.2.20]$$

und

$$e_{oi}^{(q)} = \sum_{g=1}^{k-1} e_{gi}^{(q)} \ . \qquad [2.8.2.21]$$

Die Datenaggregation ist jedoch mit einem nicht unerheblichen Verlust an Freiheitsgraden verbunden. Um die Anzahl der Freiheitsgrade möglichst groß zu halten, empfiehlt es sich daher, in [2.8.2.20] und [2.8.2.21] nicht über alle Scoregruppen zu summieren, sondern nur so lange, bis die erwarteten Häufigkeiten hinreichend groß sind, und die endgültige Testgröße analog zu [2.8.2.16] durch Addition der derart erhaltenen Statistiken zu bilden.

Bootstrap-Verfahren

Der zweite Ausweg geht auf Efron & Tishirami (1993) zurück und wird in der Literatur als Bootstrap-Verfahren[1] bezeichnet. Er besteht darin, die tatsächliche Prüfverteilung mittels Datensimulation *empirisch* zu ermitteln. Dazu gehen wir folgendermaßen vor:

1. Wir simulieren mittels Zufallsgenerator einige Hunderte oder Tausende von Datenmatrizen, für welche das Rasch-Modell gilt und die dieselben Itemlösungshäufigkeiten und dieselbe Scoreverteilung besitzen wie unsere empirischen Daten.
   Im Falle der Prüfgröße [2.8.2.12] werden also die x_{oi} (für i = 1,..,k) und die n_g (für g = 0,…,k) konstant gehalten; im Falle der Prüfgröße [2.8.2.16] die x_{gi} (für g = 1,…,k-1 und i = 1,..,k) und die $n_g^{(q)}$ (für g = 1,…,k-1 und q = 1,…,h).
2. Aus den simulierten Datenmatrizen berechnen wir die zugehörigen Werte der Prüfgröße und schätzen die Wahrscheinlichkeit $\text{prob}\left(X_o^2 \geq \chi_o^2\right)$ durch die relative Häufigkeit, mit welcher die simulierten Werte den empirisch ermittelten Wert der Prüfgröße erreichen oder übersteigen.
3. Schließlich verwerfen wir das Rasch-Modell, wenn $\text{prob}\left(X_o^2 \geq \chi_o^2\right) \leq \alpha$.

2.8.3 Kontrolle der lokalen Unabhängigkeit

Hatten die bisher dargestellten Modellgeltungstests die lokale Unabhängigkeit der Testitems vorausgesetzt und die Suffizienz der Summenscores überprüft, so werden wir in diesem Kapitel einen Modelltest kennen lernen, der die Suffizienz der Summenscores voraussetzt und die lokale Unabhängigkeit der Items überprüft.

Serielle Abhängigkeit

Wie wir in Kapitel 2.3.3 gezeigt haben, bleibt die Suffizienz des Summenscores bei Verletzung der lokalen Unabhängigkeit genau dann erhalten, wenn die bedingten Lösungswahrscheinlichkeiten in der Form des dynamischen Testmodells [2.3.2.19] darstellbar sind.

Die bedingte Likelihood der Antwortmatrix

Ist dies der Fall, dann hängt die bedingte Likelihood der Antwortmatrix

1. Rost (1996, 338) erwähnt, dass sich die Bezeichnung Bootstrapping von den Schlaufen an Cowboystiefeln herleitet. "Bootstrapping" steht für den Versuch, sich an den eigenen Stiefelschlaufen hochzuheben (ähnlich wie Baron Münchhausen, der sich am eigenen Schopf aus dem Sumpf gezogen haben soll). Für die damit bezeichneten Verfahren der Modellgeltungskontrolle ist die Metapher nicht ganz unpassend: Man will wissen, ob ein Modell auf die Daten passt und beantwortet die Frage mittels Daten, die man mit diesem Modell erzeugt hat.

$$cL(X) = \frac{\prod_{i=1}^{k}\prod_{r=0}^{i-1}(\sigma_i - \psi_r)^{n_{ri}}}{\prod_{g=1}^{k-1} G(g,\psi,\sigma)^{n_g}} \qquad [2.8.3.1]$$

nur über die Scorehäufigkeiten n_g und die so genannten bedingten Antworthäufigkeiten

$$n_{ri} = \sum_{v=1}^{n} z_{vir}(1 - x_{vi}) \qquad [2.8.3.2]$$

von den Daten ab.

Die Matrix der bedingten Antworthäufigkeiten

Die bedingten Antworthäufigkeiten n_{ri} beschreiben die Anzahl der Vpn, die Item i *nicht* lösen, nachdem sie r vorangegangene Items gelöst haben. Sie bilden eine obere Dreiecksmatrix, deren Spaltensummen $n_{oi} = n - x_{oi}$ die Anzahl der Vpn wiedergeben, welche das entsprechende Item *nicht* gelöst haben. Die Zeilensummen n_{ro} geben an, wie häufig insgesamt ein Item *nicht* gelöst wurde, nachdem r vorangegangene Items gelöst worden waren.

Tab. 2.8.3.1: Die bedingte Antwortmatrix

r	1	2	..	i	..	k	Σ
0	n_{01}	n_{02}	..	n_{0i}	..	n_{0k}	n_{0o}
1		n_{12}	..	n_{1i}	..	n_{1k}	n_{1o}
.			..	..	..	..	.
.				..	..	..	.
.					..	..	.
k − 1						$n_{k-1,k}$	$n_{k-1,o}$
Σ	n_{o1}	n_{o2}	..	n_{oi}	..	n_{ok}	

Die Matrix der erwarteten Häufigkeiten

Sind die Items lokal unabhängig voneinander, dann sind die Transferparameter ψ_r alle gleich Null und das dynamische Testmodell reduziert sich auf das Rasch-Modell.

Für die Wahrscheinlichkeit, dass eine Vp mit dem Summenscore $x_{vo} = g$ Item i *nicht* gelöst hat, nachdem sie $x_{vo}^{(i)} = r$ der vorangegangenen Items gelöst hatte, ergibt sich dann die Gleichung

$$\text{prob}\left(x_{vi} = 0 \middle| x_{vo} = g, x_{vo}^{(i)} = r\right) = \frac{\gamma_r^{(j\geq i)}(\varepsilon)\gamma_{g-r}^{(j\leq i)}(\varepsilon)}{\gamma_g(\varepsilon)}, \qquad [2.8.3.3]$$

worin $\gamma_r^{(j\geq i)}$ unter Weglassung der Itemparameter $\varepsilon_i,\ldots,\varepsilon_k$ und $\gamma_{g-r}^{(j\leq i)}$ unter Weglassung der Itemparameter $\varepsilon_1,\ldots,\varepsilon_i$ berechnet wird.

Daraus errechnen sich die bei Geltung des Rasch-Modells erwarteten Häufigkeiten

$$e_{ri} = \sum_{g=r}^{k-i+r} n_g \frac{\gamma_r^{(j \geq i)}(\varepsilon)\gamma_{g-r}^{(j \leq i)}(\varepsilon)}{\gamma_g(\varepsilon)}, \qquad [2.8.3.4]$$

deren Spaltensummen

$$e_{oi} = \sum_{r=0}^{i-1} e_{ri} = n - x_{oi} \qquad [2.8.3.5]$$

wieder die Anzahl der Vpn wiedergeben, welche das entsprechende Item *nicht* gelöst haben, so dass $e_{oi} = n_{oi}$. Die Zeilensummen der e_{ri}-Matrix ergeben die erwarteten Zeilensummen der bedingten Antwortmatrix

$$e_{ro} = \sum_{i=r+1}^{k} e_{ri} \,. \qquad [2.8.3.6]$$

Sie geben an, wie häufig bei Geltung des Rasch-Modells insgesamt zu erwarten ist, dass ein Item *nicht* gelöst wird, nachdem r vorangegangene Items gelöst wurden.

Die Konstruktion der Prüfstatistiken

Indem wir die erwarteten Häufigkeiten in Gleichung [2.8.3.5] und [2.8.3.6] mit den entsprechenden empirischen Häufigkeiten vergleichen, erhalten wir schließlich die beiden Prüfstatistiken

$$\omega_1^2 = \sum_{i=1}^{k} \sum_{r=0}^{i-1} \frac{(n_{ri} - e_{ri})^2}{e_{ri}} \qquad [2.8.3.7]$$

und

$$\omega_2^2 = \sum_{r=0}^{k-1} \frac{(n_{ro} - e_{ro})^2}{e_{ro}}, \qquad [2.8.3.8]$$

deren asymptotische Verteilung allerdings nicht bekannt ist, da einige der Voraussetzungen des Pearson-χ^2 verletzt sind.

Bootstrapping

Um die lokale Unabhängigkeit der Items mittels der Teststatistiken [2.8.3.7] und/oder [2.8.3.8] zu überprüfen, können wir uns analog zu oben jedoch wieder eines Bootstrap-Verfahrens bedienen und das Rasch-Modell verwerfen, wenn die mittels Datensimulation empirisch ermittelte Wahrscheinlichkeit $\text{prob}(\Omega^2 \geq \omega^2) \leq \alpha$.

2.8.4 Beurteilung der Testleistungen

Konfidenzintervall für den Personenparameter

Ähnlich wie für die True-Scores der klassischen Testtheorie (vgl. Kap. 2.4.3) kann man auch für die Personenparameter des Rasch-Modells einen Vertrauensbereich angeben, in dem sie mit einer vorgegebenen Sicherheit von γ liegen.

Unter Verweis auf die asymptotischen Eigenschaften von ML-Schätzern gibt Fischer (1974, 296f) hierfür das (asymptotische) Konfidenzintervall

$$\text{KONF}_{\gamma}\left\{\hat{\theta}_v - c\,\sigma(\hat{\theta}_v) \le \theta_v \le \hat{\theta}_v + c\,\sigma(\hat{\theta}_v)\right\} \qquad [2.8.4.1]$$

an, worin $\sigma(\hat{\theta}_v)$ den asymptotischen Schätzfehler bezeichnet und sich als Quadratwurzel aus dem Kehrwert der Informationsfunktion (Gleichung [2.7.4.4]) berechnet. Die Werte für c können wieder der Tabelle für die Standardnormalverteilung im Anhang entnommen werden.

Test für den Personenparameter

Logisch äquivalent mit dem Konfidenzintervall [2.8.4.1] ist die asymptotisch nach N(0,1) verteilte Testgröße

$$z = \frac{\hat{\theta}_v - \theta_{krit}}{\sigma(\hat{\theta}_v)} \qquad [2.8.4.2]$$

zur Überprüfung des Hypothesenpaares

$$H_0: \theta_v = \theta_{krit}$$

$$H_1: \theta_v \neq \theta_{krit}\,,$$

mittels derer man überprüfen kann, ob der Personenparameter einen vorgegebenen Kriteriumswert θ_{krit} signifikant über- oder unterschreitet.

Kritische Differenz der Testleistungen zweier Vpn

Sind lediglich die Scores der Vpn, nicht jedoch ihre Antwortmuster bekannt, so kann man sich zur Beurteilung der Leistungsunterschiede zweier Probanden v und w

$$H_0: \theta_v = \theta_w$$

$$H_1: \theta_v \neq \theta_w$$

in Analogie zu [2.4.3.4] der asymptotisch nach N(0,1) verteilten Testgröße

$$z = \frac{\hat{\theta}_v - \hat{\theta}_w}{\sqrt{\sigma^2(\hat{\theta}_v) + \sigma^2(\hat{\theta}_w)}} \qquad [2.8.4.3]$$

bedienen.

Da man es in der Regel mit relativ kleinen Itemzahlen zu tun hat, sind die oben dargestellten asymptotischen Ergebnisse allerdings nur bedingt brauchbar.

Test von McNemar

Zumindest für den Vergleich der Testleistungen zweier Vpn bietet der bereits in Kapitel 2.4.3 vorgestellte Test von McNemar nicht nur einen Ausweg aus diesem Dilemma, sondern er ist gleichsam der Königsweg, auf welchem das Hypothesenpaar

$$H_0 : \theta_v = \theta_w$$

$$H_1 : \theta_v \neq \theta_w$$

getestet werden kann. Bei Geltung des Rasch-Modells nimmt der Wahrscheinlichkeitsparameter p in Gleichung [2.4.3.5] nämlich die von den Itemparametern unabhängige Form

$$p = \text{prob}\{(1,0) | x_{oi} = 1\} = \frac{\xi_v}{\xi_v + \xi_w} \qquad [2.8.4.4]$$

an[1], so dass die Verteilung der Testgröße des Tests von McNemar (n_{10}) ausschließlich von den Personenparametern abhängt und also einen spezifisch objektiven Vergleich der Probanden gewährleistet.

Kontrollfragen

- Welche beiden Voraussetzungen sind notwendig und hinreichend für das Rasch-Modell?
- Welche Konsequenzen, die das Rasch-Modell gegenüber allen anderen denkbaren Testmodellen auszeichnen, ergeben sich daraus?
- Was versteht man unter spezifischer Objektivität?
- Welche Eigenschaft des Rasch-Modells wird bei der graphischen Modellkontrolle überprüft?
- Worauf muss man bei der Konstruktion der Teilstichproben achten?
- Welche Möglichkeiten gibt es, die graphische Modellkontrolle in einen

1. Die Herleitung der Gleichung [2.8.4.4] sei dem Leser als Übungsaufgabe überlassen.

statistischen Test zu übersetzen?

- Welches sind die Gemeinsamkeiten und Unterschiede des bedingten Likelihood-Quotienten-Tests nach Andersen, der Überprüfung der Reproduzierbarkeit der Itemlösungshäufigkeiten mittels Pearson-χ^2 und der parameterfreien Modellkontrolle nach Scheiblechner?
- An welche Voraussetzungen ist die Anwendung des Pearson-χ^2-Verfahrens gebunden, und welche Auswege gibt es, wenn diese Voraussetzungen verletzt sind?
- Was versteht man unter einem Bootstrap-Verfahren?
- Wie kann man die lokale Unabhängigkeit der Items überprüfen?
- Wie sind die bedingten Antworthäufigkeiten n_{ri} definiert und was beschreiben Sie?
- Welche wünschenswerte Eigenschaft besitzt der Test von McNemar bei Geltung des Rasch-Modells?

Übungsaufgaben

1. Zeigen Sie, dass die Lösungswahrscheinlichkeit eines Items bei gegebenem Score im Rasch-Modell die Form [2.8.2.10] besitzt!
2. Ein Fragebogen aus k = 4 binären Items zur Geschlechtsrollensozialisation wurde von 594 Vpn bearbeitet und einer Rasch-Analyse unterzogen. Zwecks Modellkontrolle wurde die Personenstichprobe aufgrund der Summenscores der Vpn per Mediansplit in zwei distinkte Teilstichproben mit $x_{vo} \leq 2$ ("Low-Score-Gruppe") und $x_{vo} > 2$ ("High-Score-Gruppe") zerlegt. Nachstehende Tabellen geben die CML-Schätzungen der Itemparameter und die zugehörigen Likelihoodfunktionen wieder

	δ_1	δ_2	δ_3	δ_4	ln{cL(X)}
Low-Score-Gruppe	0.3955	-0.8635	1.4030	-0.9350	-279.8295
High-Score-Gruppe	0.8378	-1.0095	1.5823	-1.4105	-115.9738
Gesamtstichprobe	0.4966	-0.8987	1.3922	0.9900	-397.8519

a) Führen Sie eine graphische Modellkontrolle durch und
b) entscheiden Sie mittel bedingtem Likelihood-Quotienten-Test, ob die Daten mit dem Rasch-Modell verträglich sind.

3. Die Zerlegung der (reduzierten) Gesamtstichprobe in die einzelnen Scoregruppen ergab die in den nachstehenden Tabellen wiedergegebenen empirischen und erwarteten Itemlösungshäufigkeiten:

x_{vo} = g	Empirische Häufigkeiten			
	i = 1	i = 2	i = 3	i = 4
1	10	40	1	49
2	44	101	20	97
3	99	122	48	124

$x_{vo} = g$	Erwartete Häufigkeiten			
	i = 1	i = 2	i = 3	i = 4
1	10.14	40.91	4.14	44.82
2	38.5	102.18	16.77	104.54
3	95.37	119.92	48.09	120.62

Überprüfen Sie die Geltung des Rasch-Modells mittels Pearson-χ^2-Test!

4. Die Teilung der Stichprobe aufgrund des Geschlechts der Vpn ergab die in den nachstehenden Tabellen wiedergegebenen Itemlösungshäufigkeiten:

Scoregruppe 1: $x_{vo} = 1$						
	q	i = 1	i = 2	i = 3	i = 4	$n_1^{(q)}$
männl.	1	7	29	4	10	50
weibl.	2	3	11	6	30	50
Summe		10	40	10	40	100

Scoregruppe 2: $x_{vo} = 2$						
	q	i = 1	i = 2	i = 3	i = 4	$n_2^{(q)}$
männl.	1	24	43	3	43	60
weibl.	2	20	58	8	63	71
Summe		44	101	11	106	131

Scoregruppe 3: $x_{vo} = 3$						
	q	i = 1	i = 2	i = 3	i = 4	$n_3^{(q)}$
männl.	1	30	41	17	41	43
weibl.	2	60	81	31	83	85
Summe		90	122	48	124	128

Überprüfen Sie die Geltung des Rasch-Modells mittels parameterfreier Modellkontrolle!

5. Die nachstehenden Tabellen geben die Matrix der empirischen und der erwarteten bedingten Antworthäufigkeiten wieder. Berechnen Sie die Prüfgrößen ω_1^2 und ω_2^2 zur Überprüfung der lokalen Unabhängigkeit der Testitems!

r	Empirische Häufigkeiten			
	i = 1	i = 2	i = 3	i = 4
0	215	68	49	0
1		28	129	51
2			112	34
3				4

Erwartete Häufigkeiten			
i = 1	i = 2	i = 3	i = 4
214.99	56.8	44.82	0
	39.19	147.75	55.18
		97.43	26.46
			7.38

6. Gegeben sei eine Vp v, deren Testergebnis in einem Leistungstest ein geschätzter Personenparameter von $\hat{\theta}_v = 4.32$ entspricht. Der asymptotische Schätzfehler des Personenparameters beträgt $\sigma(\hat{\theta}_v) = 0.921$.
   a) Geben Sie einen Vertrauensbereich an, innerhalb dessen der Parameter der Person mit einer Sicherheit von γ = 95% liegt!
   b) Überprüfen Sie auf einem Signifikanzniveau von α = 1%, ob die latente Leistungsfähigkeit der Vpn signifikant größer ist als die durchschnittliche Leistungsfähigkeit von $\overline{\Theta} = 0.89$!
7. Der Testleistung einer anderen Vp w entspricht ein geschätzter Personenparameter von $\hat{\theta}_w = 1.67$, dessen Schätzfehler $\sigma(\hat{\theta}_w) = 0.342$ beträgt. Untersuchen Sie, ob sich die Leistungsfähigkeit der beiden Vpn signifikant unterscheidet (α = 1%)!
8. Zeigen Sie, dass im Rasch-Modell die Gleichung [2.8.4.4] erfüllt ist!

2.9
Verallgemeinerungen des Rasch-Modells

2.9.1 Mehrfaktorielle Rasch-Modelle

Mehrfaktorielles Poisson-Modell

Bisher haben wir nur solche Rasch-Modelle betrachtet, die die Wahrscheinlichkeitsdichte der Antwortvariablen auf zwei Faktoren zurückführen: die Fähigkeit der Vp und die Schwierigkeit des Items.

Der Gedanke, dass man das Modell um weitere Faktoren erweitern könnte, mittels derer sich z. B. der Einfluss messen ließe, welchen die experimentelle Variation der Versuchsbedingungen auf das Antwortverhalten der Probanden ausübt, wurde erstmals bereits von Rasch (1965) verfolgt und für das Poisson-Modell

$$\text{prob}(X_{vij} = x) = \frac{\xi_v \varepsilon_i \beta_j}{x!} e^{-\xi_v \varepsilon_i \beta_j} \quad \text{mit } x = 0,1,2\ldots \qquad [2.9.1.1]$$

detailliert ausgearbeitet, worin der Parameter β_j den Effekt der Versuchsbedingung j beschreibt.

Mehrfaktorielle logistische Modelle

Micko (1970) hat die mehrfaktorielle Erweiterung auf das dichotome Rasch-Modell

$$\text{prob}(X_{vij} = x) = \frac{(\xi_v \varepsilon_i \beta_j)^x}{1 + \xi_v \varepsilon_i \beta_j} \qquad [2.9.1.2]$$

und Kempf (1972a) auf das polytome Rasch-Modell

$$\text{prob}(X_{vij} = x) = \frac{\xi_{vx} \varepsilon_{ix} \beta_{jx}}{\sum_{y=0}^{m-1} \xi_{vy} \varepsilon_{iy} \beta_{jy}} \quad \text{mit} = 0,\ldots,m-1 \qquad [2.9.1.3]$$

übertragen, das den Effekt der Versuchsbedingung in Form einer m-dimensionalen Vektorvariable

$$\beta_j = (\beta_{j0},\ldots,\beta_{jm-1}) \qquad [2.9.1.4]$$

darstellt, deren Komponenten den Einfluss der Versuchsbedingung auf die verschiedenen Antwortkategorien beschreiben. In der englischsprachigen Literatur (z. B. Linacre, 1989; Linacre & Wright, 2002) werden die mehrfaktoriellen logistischen Modelle auch als Many-Facet-Rasch-Modelle bezeichnet. Eckes (2006) nennt sie Multifacetten-Rasch-Modelle.

Multifaktorielle Versuchspläne

Werden mehrere experimentelle Faktoren a = 1,…,z mit jeweils c_a Faktorstufen b = 0,…,c_a-1 gleichzeitig variiert, dann ist Versuchsbedingung j durch die jeweils realisierten Faktorstufen definiert und kann durch die Gewichtszahlen

$$g_{ab}^{(j)} = \begin{cases} 1 & \text{falls Stufe b des Faktors a realisiert ist} \\ 0 & \text{sonst} \end{cases} \qquad [2.9.1.5]$$

beschrieben werden. Ähnlich wie bei der Varianzanalyse kann der Effekt der Versuchsbedingung dann in die Effekte der einzelnen Faktoren aufgespaltet werden, so dass in [2.9.1.1] und [2.9.1.2]

$$\ln(\beta_j) = \sum_{a=1}^{z} \sum_{b=0}^{c_a-1} g_{ab}^{(j)} \, \omega_b^{(a)} \;, \qquad [2.9.1.6]$$

bzw. in [2.9.1.3]

$$\ln(\beta_{jx}) = \sum_{a=1}^{z} \sum_{b=0}^{c_a-1} g_{ab}^{(j)} \, \omega_{bx}^{(a)} . \qquad [2.9.1.7]$$

Darin beschreibt der (logarithmische) Effektparameter $\omega_b^{(a)}$ den Effekt der Stufe b des Faktors a. Im Falle des polytomen Modells stellt $\omega_b^{(a)} = \left(\omega_{b0}^{(a)},\ldots,\omega_{bm-1}^{(a)}\right)$ eine Vektorvariable dar. In den anderen beiden Modellen ist $\omega_b^{(a)}$ ein Skalar.

Normierung der Modellparameter

Ähnlich wie die Itemschwierigkeiten δ_i = -ln(ε_i) werden auch die Effekte der experimentellen Faktoren auf Differenzenskalen gemessen, so dass man bezüglich jedes der experimentellen Faktoren a = 1,2,… eine der Faktorstufen (o.B.d.A. sei dies die Stufe b = 0) als Standardbedingung definieren und den zugehörigen Parameter $\omega_0^{(a)}$ gleich Null setzen kann. Pro Faktor sind somit c_a-1 unabhängige Effektparameter zu schätzen.

Versuchspläne ohne Wechselwirkungen

Betrachten wir z. B. einen zweifaktoriellen Versuchsplan (z = 2), worin der erste Faktor (a = 1) ein dreistufiger (c_1 = 3) und der zweite Faktor (a = 2) ein zweistufiger (c_2 = 2) ist. Dann ergeben

sich $c_1 \times c_2 = 6$ Versuchsbedingungen, denen die Gewichtszahlen und Effektparameter in Tabelle 2.9.1.1 entsprechen.

Tab. 2.9.1.1: Zweifaktorieller Versuchsplan ohne Wechselwirkungen

Versuchsplan			Gewichtszahlen $g_{ab}^{(j)}$					$\ln(\beta_j)$
j	F1	F2	a = 1			a = 2		
			b=0	b=1	b=2	b=0	b=1	
1	0	0	1	0	0	1	0	$\omega_0^{(1)} + \omega_0^{(2)}$
2	0	1	1	0	0	0	1	$\omega_0^{(1)} + \omega_1^{(2)}$
3	1	0	0	1	0	1	0	$\omega_1^{(1)} + \omega_0^{(2)}$
4	1	1	0	1	0	0	1	$\omega_1^{(1)} + \omega_1^{(2)}$
5	2	0	0	0	1	1	0	$\omega_2^{(1)} + \omega_0^{(2)}$
6	2	1	0	0	1	0	1	$\omega_2^{(1)} + \omega_1^{(2)}$

Eigenschaften der mehrfaktoriellen Rasch-Modelle

Zu den hervorragenden Eigenschaften der oben dargestellten Modelle gehört die Existenz erschöpfender Statistiken nicht nur für die Item- und Personenparameter, sondern auch für die Effektparameter $\omega_b^{(a)}$. Dadurch wird – bei geeigneter Versuchsplanung – die vollständige Separierbarkeit aller Modellparameter voneinander gewährleistet. Vergleiche zwischen den Vpn können unabhängig von der Auswahl der Items und unabhängig von der Auswahl der Versuchsbedingungen erfolgen; Vergleiche zwischen den Items unabhängig von der Auswahl der Vpn und unabhängig von der Auswahl der Versuchsbedingungen; und die verschiedenen Stufen eines experimentellen Faktors können unabhängig von der Auswahl der Vpn, unabhängig von der Auswahl der Items und unabhängig von den anderen Faktoren und deren Faktorenstufen miteinander verglichen werden.

Vollständige Wechselwirkungen

Eine Konsequenz der vollständigen spezifischen Objektivität ist es allerdings, dass – anders als bei der Varianzanalyse – keine vollständigen Wechselwirkungen zwischen den experimentellen Faktoren vorgesehen sind.

Interagieren zwei Faktoren a und a' mit den Faktorstufen b = 0,…,c-1 und b' = 0,…,c'-1 vollständig miteinander, so kann man sie jedoch zu einem (Meta-)Faktor a* = (a,a') mit den Faktorstufen b* = (b,b') zusammenfassen. Für den Versuchsplan oben erhält man dann z. B. die Gewichtszahlen und Effektparameter in Tabelle 2.9.1.2.

Tab.2.9.1.2: Zweifaktorieller Versuchsplan mit vollständigen Wechselwirkungen

Versuchsplan			Gewichtszahlen $g^{(j)}_{a^*b^*}$						$\ln(\beta_j)$
j	Faktoren		b* = (b,b')						
	F1=a	F2=a'	(0,0)	(0,1)	(1,0)	(1,1)	(2,0)	(2,1)	
1	0	0	1	0	0	0	0	0	$\omega^{(a,a')}_{0,0}$
2	0	1	0	1	0	0	0	0	$\omega^{(a,a')}_{0,1}$
3	1	0	0	0	1	0	0	0	$\omega^{(a,a')}_{1,0}$
4	1	1	0	0	0	1	0	0	$\omega^{(a,a')}_{1,1}$
5	2	0	0	0	0	0	1	0	$\omega^{(a,a')}_{2,0}$
6	2	1	0	0	0	0	0	1	$\omega^{(a,a')}_{2,1}$

Aus der Matrix der Effektparameter des Meta-Faktors a* in Tabelle 2.9.1.3 können die Haupteffekte der beiden Faktoren anschließend mittels der Methode der kleinsten Quadrate gewonnen werden, indem man die Zeilen- und Spaltenmittelwerte

$$\overline{\omega}^{(a)}_{b} = \frac{1}{c'}\sum_{b'=0}^{c'-1}\omega^{(a,a')}_{b,b'} \quad \text{für } b = 0,\ldots,c\text{-}1 \qquad [2.9.1.8]$$

und

$$\overline{\omega}^{(a')}_{b'} = \frac{1}{c}\sum_{b=0}^{c-1}\omega^{(a,a')}_{b,b'} \quad \text{für } b' = 0,\ldots,c'\text{-}1 \qquad [2.9.1.9]$$

bildet. Die Interaktionseffekte $\kappa^{(a^*)}_{bb'}$ berechnen sich daraus durch

$$\kappa^{(a,a')}_{b,b'} = \omega^{(a,a')}_{b,b'} - \left(\overline{\omega}^{(a)}_{b} + \overline{\omega}^{(a')}_{b'}\right), \qquad [2.9.1.10]$$

so dass

$$\omega^{(a,a')}_{b,b'} = \overline{\omega}^{(a)}_{b} + \overline{\omega}^{(a')}_{b'} + \kappa^{(a,a')}_{b,b'}. \qquad [2.9.1.11]$$

Die so berechneten Haupteffekte der interagierenden Faktoren a und a' können jedoch nicht spezifisch objektiv erfasst werden. Ähnlich wie in der Varianzanalyse auch, beschreiben sie lediglich die *durchschnittlichen* Effekte der Faktoren und sind davon abhängig, welche Stufen des jeweils anderen Faktors im Experiment realisiert wurden.

Tab. 2.9.1.3: Matrix der Effektparameter des Meta-Faktors a*

	0	..	b'		c'−1
0	$\omega_{0,0}^{(a,a')}$	..	$\omega_{0,b'}^{(a,a')}$	..	$\omega_{0,c'-1}^{(a,a')}$
.	.	..	.	..	.
.	.	..	.	..	.
b	$\omega_{b,0}^{(a,a')}$	..	$\omega_{b,b'}^{(a,a')}$	..	$\omega_{b,c'-1}^{(a,a')}$
.	.	..	.	..	.
.	.	..	.	..	.
c − 1	$\omega_{c-1,0}^{(a,a')}$	..	$\omega_{c-1,b'}^{(a,a')}$	..	$\omega_{c-1,c'-1}^{(a,a')}$

Partielle Wechselwirkungen

Eine weitaus günstigere Situation ergibt sich, wenn die experimentellen Faktoren nur partiell interagieren, so dass ein Faktor nur mit bestimmten Stufen eines (oder mehrerer) anderer Faktoren interagiert.

Ist dies der Fall, so können wir die Effekte der interaktionsfreien Versuchsbedingungen als Summe der Haupteffekte der entsprechenden Faktorstufen darstellen, die gemeinsamen Effekte der interagierenden Faktorstufen durch einen zusätzlichen (virtuellen) Faktor beschreiben und sowohl die Haupteffekte als auch die gemeinsamen Effekte spezifisch objektiv messen.

Haupteffekte und gemeinsame Effekte

Anders als in der Varianzanalyse wollen wir Haupteffekte jedoch zunächst nur für die interaktionsfreien Versuchsbedingungen postulieren.

Z. B. möge der zweite Faktor in unserem Versuchsplan (F2) nur mit der dritten Faktorstufe (b_1=2) des ersten Faktors (F1) interagieren. Dann können wir die Effekte der interaktionsfreien Versuchsbedingungen j = 1 bis 4 in der Form

$$\ln(\beta_j) = \omega_{b_1}^{(1)} + \omega_{b_2}^{(2)} \quad \text{mit } b_1 = 0{,}1 \text{ und } b_2 = 0{,}1$$

durch die Haupteffekte der entsprechenden Faktorstufen darstellen, wobei wir diesen Versuchsbedingungen *keine Faktorstufe des virtuellen Faktors* (F3) zuordnen. Den interaktionshaltigen Versuchsbedingungen j = 5 und 6 wiederum ordnen wir *keine Haupteffekte* zu und stellen die gemeinsamen Effekte der interagierenden Faktorstufen mittels

$$\ln(\beta_j) = \omega_{b_3}^{(3)} = \begin{cases} \omega_{2,0}^{(1,2)} & \text{für } j = 5 \text{ (mit } b_3 = 1) \\ \omega_{2,1}^{(1,2)} & \text{für } j = 6 \text{ (mit } b_3 = 2) \end{cases}$$

als Effekte des virtuellen Faktors dar.[1]

Haupteffekte können dabei nur jenen Faktorstufen zugeordnet werden, die an mindestens einer interaktionsfreien Versuchsbedingung beteiligt sind. Da dies für die dritte Stufe des 1. Faktors ($b_1 = 2$) in unserem Beispiel nicht der Fall ist, ist eine entsprechende Spalte in der Gewichtszahlenmatrix in Tabelle 2.9.1.4 auch nicht enthalten.

Versuchsplan				Gewichtszahlen $g_{ab}^{(j)}$						$\ln(\beta_j)$
				a=1		a=2		a=3		
j	F1	F2	F3	b_1=		b_2=		b_3=		
				0	1	0	1	1	2	
1	0	0	–	1	0	1	0	0	0	$\omega_0^{(1)} + \omega_0^{(2)}$
2	0	1	–	1	0	0	1	0	0	$\omega_0^{(1)} + \omega_1^{(2)}$
3	1	0	–	0	1	1	0	0	0	$\omega_1^{(1)} + \omega_0^{(2)}$
4	1	1	–	0	1	0	1	0	0	$\omega_1^{(1)} + \omega_1^{(2)}$
5	(2,0)		1	0	0	0	0	1	0	$\omega_1^{(3)} = \omega_{2,0}^{(1,2)}$
6	(2,1)		2	0	0	0	0	0	1	$\omega_2^{(3)} = \omega_{2,1}^{(1,2)}$

Tab. 2.9.1.4: Zweifaktorieller Versuchsplan mit Wechselwirkungen zwischen der Faktorstufe $b_1 = 2$ und dem Faktor F2

Haupteffekte und Wechselwirkungen

Um die gemeinsamen Effekte interagierender Faktorstufen gemäß

$$\omega_{b,b'}^{(a,a')} = \omega_b^{(a)} + \omega_{b'}^{(a')} + \kappa_{b,b'}^{(a,a')} \qquad [2.9.1.12]$$

in Haupteffekte und Wechselwirkungen zerlegen zu können, sind die Voraussetzungen sogar noch strenger. Die Rede von einem (additiven) Haupteffekt einer Faktorstufe macht nämlich nur dann einen Sinn, wenn diese an *mindestens zwei* interaktionsfreien Versuchsbedingungen beteiligt ist. Dafür ist mindestens ein 3x3-Versuchsplan erforderlich, wie er in Tabelle 2.9.1.5 dargestellt ist.

Anders als oben beschreibt der virtuelle Faktor F3 darin nicht mehr die gemeinsamen Effekte $\omega_{b,b'}^{(a,a')}$ der interagierenden Faktorstufen, sondern die durch

1. Man beachte, dass die Nummerierung der Faktorstufen des virtuellen Faktors erst ab $b_3 = 1$ beginnt.

$$\kappa_{b,b'}^{(a,a')} = \omega_{b,b'}^{(a,a')} - \left(\omega_b^{(a)} + \omega_{b'}^{(a')}\right) \qquad [2.9.1.13]$$

definierten Interaktionseffekte. Um dies zu gewährleisten, ordnen wir nunmehr sowohl den interaktionsfreien als auch den interaktionshaltigen Versuchsbedingungen die Haupteffekte jener Faktorstufen zu, die an der jeweiligen Versuchsbedingung beteiligt sind. Den interaktionsfreien Versuchsbedingungen dagegen wird wie bisher *keine Faktorstufe des virtuellen Faktors* zugeordnet.

Tab. 2.9.1.5: Zweifaktorieller Versuchsplan mit Wechselwirkungen zwischen den Faktorstufen $b_1 = 2$ und $b_2 = 1$

Versuchsplan				Gewichtszahlen $g_{ab}^{(j)}$							$\ln(\beta_j)$
				a=1			a=2			a=3	
j	F1	F2	F3	b_1=			b_2=			b_3=	
				0	1	2	0	1	2	1	
1	0	0	–	1	0	0	1	0	0	0	$\omega_0^{(1)} + \omega_0^{(2)}$
2	0	1	–	1	0	0	0	1	0	0	$\omega_0^{(1)} + \omega_1^{(2)}$
3	0	2	–	1	0	0	0	0	1	0	$\omega_0^{(1)} + \omega_2^{(2)}$
4	1	0	–	0	1	0	1	0	0	0	$\omega_1^{(1)} + \omega_0^{(2)}$
5	1	1	–	0	1	0	0	1	0	0	$\omega_1^{(1)} + \omega_1^{(2)}$
6	1	2	–	0	1	0	0	0	1	0	$\omega_1^{(1)} + \omega_2^{(2)}$
7	2	0	–	0	0	1	1	0	0	0	$\omega_2^{(1)} + \omega_0^{(2)}$
8	(2,1)		1	0	0	1	0	1	0	1	$\omega_2^{(1)} + \omega_1^{(2)} + \omega_1^{(3)} = \omega_2^{(1)} + \omega_1^{(2)} + \kappa_{2,1}^{(1,2)}$
9	2	2	–	0	0	1	0	0	1	0	$\omega_2^{(1)} + \omega_2^{(2)}$

Auf die formalen Bedingungen, welche die Gewichtszahlenmatrix erfüllen muss, damit die Modellparameter identifizierbar sind, werden wir im nächsten Kapitel eingehen.

2.9.2 Das lineare logistische Testmodell

Die Modellstruktur

Formal äquivalent mit dem mehrfaktoriellen logistischen Modell ist das sogenannte lineare logistische Testmodell (LLTM), das zunächst von Fischer (1973, 1983) für dichotome Daten entwickelt und später von Fischer & Parzer (1991a, b) und Fischer & Pococny (1994) für ordinale Daten verallgemeinert wurde.

Der Grundgedanke des LLTM besteht darin, die Aufgabenparameter in eine gewichtete Summe von $m \leq k$ sogenannten *Basisparametern* η_j zu zerlegen, wofür im dichotomen Fall die Modellstruktur

$$\ln(\varepsilon_i) = \sum_{j=1}^{m} q_{ij}\eta_j + c \quad \text{für } i = 1,\ldots,k \qquad [2.9.2.1]$$

postuliert wird. Darin ist c eine Normierungskonstante, die dem Umstand geschuldet ist, dass die $\ln(\varepsilon_i)$ auf einer Differenzenskala gemessen werden. Die Gewichtszahlen q_{ij} bilden die Designmatrix $Q = ((q_{ij}))$ und stellen Hypothesen bezüglich der Aufgabenstruktur dar, die mittels des Modells getestet werden sollen.

Je größer η_j ist, desto mehr trägt die entsprechende Basiskomponente zur Leichtigkeit der Aufgabe bei. Die Basisparameter sind daher ebenfalls als Leichtigkeitsparameter zu interpretieren.

Hypothesen über die Aufgabenstruktur

Im einfachsten Fall nehmen die Gewichtszahlen nur die Werte Null oder Eins an. Dies ist z. B. dann der Fall, wenn die Aufgabenschwierigkeiten $\delta_i = -\ln(\varepsilon_i)$ auf die Denkoperationen zurückgeführt werden sollen, die zur Lösung der Aufgaben erforderlich sind ($q_{ij} = 1$) oder nicht ($q_{ij} = 0$).

So könnte man z. B. versuchen, die Schwierigkeit einfacher Rechenaufgaben dadurch zu erklären, welche Rechenarten zur Lösung der Aufgaben erforderlich sind. Als Beispiel hierfür, zeigt Tabelle 2.9.2.1 (a) die Designmatrix eines Rechentests aus drei Items

Aufgabe 1: 2 + 5 = ___

Aufgabe 2: 3 x 9 = ___

Aufgabe 3: 2x3 + 7 = ___ ,

zu deren Lösung einstellige Zahlen addiert (j = 1) und/oder multipliziert (j = 2) werden müssen.

Die Gewichtszahlen müssen jedoch nicht auf Null und Eins beschränkt bleiben, sondern können jeden beliebigen ganzzahligen Wert annehmen. Wenn zur Lösung mancher Aufgaben unseres Rechentests z. B. dieselben Recharten mehrmals angewendet werden müssen, so kann man dies durch eine Designmatrix abbilden, in der die Gewichtszahlen q_{ij} die Häufigkeit angeben, mit

der die entsprechende Rechenart ausgeführt werden muss. So zeigt die Tabelle 2.9.2.1 (b) die Designmatrix eines Rechentests mit den Items

Aufgabe 1: 7 + 2 + 6 = ___

Aufgabe 2: 4 x 6 = ___

Aufgabe 3: 5 + 3x2 = ___ .

Tab. 2.9.2.1: Designmatrizen zur Darstellung von Hypothesen über die Aufgabenstruktur

i	η_1	η_2	$\ln(\varepsilon_i)$
1	1	0	$\eta_1 + c$
2	0	1	$\eta_2 + c$
3	1	1	$\eta_1 + \eta_2 + c$

a) Anwendung der Rechenarten

i	η_1	η_2	$\ln(\varepsilon_i)$
1	2	0	$2\eta_1 + c$
2	0	1	$\eta_2 + c$
3	1	1	$\eta_1 + \eta_2 + c$

b) Anwendungshäufigkeit der Rechenarten

Voraussetzungen bezüglich der Designmatrix

Auch die Einschränkung auf ganzzahlige Gewichtszahlen ist nicht zwingend. Nach Rost (1996, 247) kann es in bestimmten Anwendungskontexten (z. B. bei der Analyse von Lernprozessen) mitunter sinnvoll sein, Rationalzahlen als Gewichtszahlen zu verwenden.

Voraussetzung ist allerdings in jedem Fall, dass die Designmatrix maximalen Rang

$$R(Q) = m \qquad [2.9.2.2]$$

besitzen muss, da andernfalls das Gleichungssystem [2.9.2.1] keine eindeutige Lösung hat.

Binomialmodell

Will man die Hypothese testen, dass die Aufgaben alle gleich schwierig sind (Binomialmodell), so reduziert sich die Anzahl der Basisparameter auf m = 1 und die Designmatrix hat die in Tabelle 2.9.2.2 (a) dargestellte Form eines Spaltenvektors, der lauter Einsen enthält.

Rasch-Modell

Setzt man – wie in Tabelle 2.9.2.2 (b) – die Einheitsmatrix als Designmatrix ein, so ist m = k und das LLTM reduziert sich auf das Rasch-Modell.

i	η_1	$\ln(\varepsilon_i)$
1	1	$\eta_1 + c$
2	1	$\eta_1 + c$
3	1	$\eta_1 + c$

a) Binomialmodell

i	η_1	η_2	η_3	$\ln(\varepsilon_i)$
1	1	0	0	$\eta_1 + c$
2	0	1	0	$\eta_2 + c$
3	0	0	1	$\eta_3 + c$

b) Rasch-Modell

Tab. 2.9.2.2: Designmatrizen zur Darstellung des Binomialmodells und des Rasch-Modells

Parameterschätzung

Die Modellparameter des LLTM können mittels der CML-Methode geschätzt werden, indem man die ersten partiellen Ableitungen der bedingten Likelihood [2.7.2.1] nach den Basisparametern gleich Null setzt. Die CML-Schätzer erhält man dann durch Auflösung des Gleichungssystems

$$\sum_{i=1}^{k} x_{oi} q_{ij} = \sum_{i=1}^{k} \sum_{g=1}^{k} n_g \frac{\gamma_{g-1}^{(i)}(\varepsilon)\varepsilon_i}{\gamma_g(\varepsilon)} q_{ij} \quad \text{für } j = 1,\ldots,m \qquad [2.9.2.3]$$

nach den η_j. Die Herleitung des Gleichungssystems sei dem Leser als Übungsaufgabe überlassen.

Hypothesentestung

Zur Testung der in der Designmatrix niedergelegten Hypothesen bezüglich der Aufgabenstruktur schlägt Fischer (1974) einen bedingten Likelihood-Quotienten-Test vor, der die bedingte Likelihood der Datenmatrix unter Annahme des LLTM (H_0: Designmatrix = Q)

$$L_0 = cL(X|Q) \qquad [2.9.2.4]$$

und die bedingte Likelihood der Datenmatrix unter Annahme des Rasch-Modells (H_1: Designmatrix = E)

$$L_1 = cL(X|E) \qquad [2.9.2.5]$$

einander gegenüberstellt und die Designmatrix Q verwirft, wenn die asymptotisch nach χ^2 mit

$$df = k - R(Q_*) \qquad [2.9.2.6]$$

Freiheitsgraden verteilte Testgröße

$$-2\ln(\lambda) = 2\left|\ln(L_0) - \ln(L_1)\right| \qquad [2.9.2.7]$$

den kritischen Wert der χ^2-Verteilung übersteigt. Darin bezeichnet $R(Q_*)$ den Rang der erweiterten Designmatrix, die man erhält, wenn man zu Q eine weitere Spalte hinzufügt, die lauter Einsen enthält.

Ist der Rang der erweiterten Designmatrix ebenfalls maximal, so dass

$$R(Q_*) = m + 1, \qquad [2.9.2.8]$$

dann werden die Basisparameter auf einer Absolutskala gemessen und die Anzahl der Freiheitsgrade ist entsprechend gleich

$$df = k - m - 1. \qquad [2.9.2.9]$$

Andernfalls ist

$$R(Q_*) = R(Q) = m \qquad [2.9.2.10]$$

und die Anzahl der Freiheitsgrade nimmt den Wert

$$df = k - m \qquad [2.9.2.11]$$

an.

Voraussetzungen des Likelihood-Quotienten-Tests

Voraussetzung für die Berechnung von Likelihood-Quotienten-Tests ist in jedem Fall die Geltung des unter der Alternativhypothese postulierten Modells. Eine Prüfung von Hypothesen über die Aufgabenparameter mittels des LLTM ist daher erst dann zulässig, wenn zuvor die Geltung des Rasch-Modells sichergestellt wurde, wofür man sich der in Kap. 2.8 dargestellten Modellgeltungstests bedienen kann.

2.9.3 Versuchsplanung

Unterscheidung zwischen Items und Aufgaben

Gibt man in Abweichung von unserem bisherigen Sprachgebrauch die Synonymität von Items und Aufgaben auf und versteht unter den Aufgaben stattdessen das Zusammentreffen der Items mit den jeweiligen Versuchsbedingungen, unter welchen sie bearbeitet werden, so kann auch das mehrfaktorielle Rasch-Modell in Form eines LLTM dargestellt werden.

Ein-Gruppen-Versuchspläne

Tabelle 2.9.3.1 enthält als Beispiel hierfür einen zweifaktoriellen Versuchsplan ohne Wechselwirkungen, bei dem ein Test aus drei Items mit den Schwierigkeitsparametern δ_1, δ_2 und δ_3 unter vollständiger Variation zweier experimenteller Faktoren bearbeitet wird, von denen der eine (F1) ein dreistufiger Faktor mit den Effektparametern $\omega_0^{(1)}$, $\omega_1^{(1)}$ und $\omega_2^{(1)}$ und der andere (F2) ein zweistufiger Faktor mit den Effektparametern $\omega_0^{(2)}$ und $\omega_1^{(2)}$ ist.

A	Versuchsplan			Designmatrix					
	I	F1	F2	$-\delta_1$	$-\delta_2$	$-\delta_3$	$\omega_1^{(1)}$	$\omega_2^{(1)}$	$\omega_1^{(2)}$
				η_1	η_2	η_3	η_4	η_5	η_6
1	1	0	0	1	0	0	0	0	0
2	2	0	0	0	1	0	0	0	0
3	3	0	0	0	0	1	0	0	0
4	1	1	0	1	0	0	1	0	0
5	2	1	0	0	1	0	1	0	0
6	3	1	0	0	0	1	1	0	0
7	1	2	0	1	0	0	0	1	0
8	2	2	0	0	1	0	0	1	0
9	3	2	0	0	0	1	0	1	0
10	1	0	1	1	0	0	0	0	1
11	2	0	1	0	1	0	0	0	1
12	3	0	1	0	0	1	0	0	1
13	1	1	1	1	0	0	1	0	1
14	2	1	1	0	1	0	1	0	1
15	3	1	1	0	0	1	1	0	1
16	1	2	1	1	0	0	0	1	1
17	2	2	1	0	1	0	0	1	1
18	3	2	1	0	0	1	0	1	1

Tab. 2.9.3.1: Designmatrix eines zweifaktoriellen Ein-Gruppen-Versuchsplans ohne Wechselwirkungen
A = Aufgabennummer
I = Itemnummer
F1 = Stufe des 1. Faktors
F2 = Stufe des 2. Faktors

Jede Vp bearbeitet jedes der Testitems unter jeder der möglichen Versuchsbedingungen, so dass sich insgesamt 3x2x3 Aufgaben ergeben, deren Leichtigkeit durch sechs Basisparameter erklärt wird. Da die Effektparameter der Standardbedingungen $\omega_0^{(a)}$ zu Null normiert sind, wird ihnen auch kein Basisparameter zugeordnet.

Minimale Itemanzahl

Die geringste Anzahl von Testitems, die zur Realisierung eines solchen Versuchsplanes erforderlich ist, ergibt sich, wenn jede Vp unter jeder der Versuchsbedingungen mit denselben Items getestet wird. In diesem Falle genügt ein einziges Item. Als Beispiel hierfür gibt Tabelle 2.9.3.2 die Designmatrix eines Versuchsplans mit vier Versuchsbedingungen wieder.

Tab. 2.9.3.2: Designmatrix eines Ein-Gruppen-Versuchsplans mit vier Versuchsbedingungen
A = Aufgabennummer
I = Itemnummer
VB = Versuchsbedingung

	Versuchsplan		Designmatrix			
			$-\delta_1$	$\omega_1^{(1)}$	$\omega_2^{(1)}$	$\omega_3^{(1)}$
A	I	VB	η_1	η_2	η_3	η_4
1	1	0	1	0	0	0
2	1	1	1	1	0	0
3	1	2	1	0	1	0
4	1	3	1	0	0	1

Versuchspläne mit mehreren Versuchsgruppen

Ein-Gruppen-Versuchspläne mit vollständigen Messwiederholungen sind aber kaum praktikabel und man wird bestrebt sein, die Anzahl der Messwiederholungen so gering wie möglich zu halten.

Tab. 2.9.3.3: Designmatrix eines Versuchsplans mit sechs Versuchsgruppen
T = Teilstichprobe
A = Aufgabennummer
I = Itemnummer
VB = Versuchsbedingung

Versuchsplan				Designmatrix			
				$-\delta_1$	$\omega_1^{(1)}$	$\omega_2^{(1)}$	$\omega_3^{(1)}$
T	A	I	VB	η_1	η_2	η_3	η_4
VG1	1	1	0	1	0	0	0
	2	1	1	1	1	0	0
VG2	1	1	0	1	0	0	0
	3	1	2	1	0	1	0
VG3	1	1	0	1	0	0	0
	4	1	3	1	0	0	1
VG4	2	1	1	1	1	0	0
	3	1	2	1	0	1	0
VG5	2	1	1	1	1	0	0
	4	1	3	1	0	0	1
VG6	3	1	2	1	0	1	0
	4	1	3	1	0	0	1

Da das LLTM grundsätzlich mit Missing Data umgehen kann und nicht voraussetzt, dass alle Vpn dieselbe Auswahl von Aufgaben bearbeiten müssen, kann dies erreicht werden, indem man die

Personenstichprobe in Teilstichproben (= Versuchsgruppen) zerlegt und jede Versuchsgruppe (VG) unter zwei verschiedenen Versuchsbedingungen testet, so dass sich die Anzahl der Testungen pro Vp auf zwei reduziert und jedem möglichen Paar von Versuchsbedingungen eine Versuchsgruppe entspricht. Im Falle eines Versuchsplans mit vier Versuchsbedingungen ergibt sich z. B. die in Tabelle 2.9.3.3 dargestellte Designmatrix mit sechs Versuchsgruppen.

Versuchsplan				Designmatrix				
T	A	I	VB	$-\delta_1$	$-\delta_2$	$\omega_1^{(1)}$	$\omega_2^{(1)}$	$\omega_3^{(1)}$
				η_1	η_2	η_3	η_4	η_5
VG1	1	1	0	1	0	0	0	0
	2	2	1	0	1	1	0	0
VG2	1	1	0	1	0	0	0	0
	3	2	2	0	1	0	1	0
VG3	1	1	0	1	0	0	0	0
	4	2	3	0	1	0	0	1
VG4	5	1	1	1	0	1	0	0
	3	2	2	0	1	0	1	0
VG5	5	1	1	1	0	1	0	0
	4	2	3	0	1	0	0	1
VG6	6	1	2	1	0	0	1	0
	7	2	3	0	1	0	0	1
KG1	1	1	0	1	0	0	0	0
	8	2	0	0	1	0	0	0
KG2	5	1	1	1	0	1	0	0
	2	2	1	0	1	1	0	0
KG3	6	1	2	1	0	0	1	0
	3	2	2	0	1	0	1	0
KG4	9	1	3	1	0	0	0	1
	4	2	3	0	1	0	0	1

Tab. 2.9.3.4: Designmatrix eines Versuchsplans mit sechs Versuchsgruppen und vier Kontrollgruppen
T = Teilstichprobe
A = Aufgabennummer
I = Itemnummer
VB = Versuchsbedingung

Versuchspläne mit Versuchsgruppen und Kontrollgruppen

Um die mehrmalige Bearbeitung derselben Testitems durch dieselben Vpn zu vermeiden, wird man bestrebt sein, bei der ersten und zweiten Testung der Vpn verschiedene Items zu verwenden. Damit die Effektparameter von den Itemschwierigkeitsparametern getrennt werden können, muss man den Versuchsplan dazu um (mindestens) eine Kontrollgruppen (KG) erweitern, die unter konstanten Versuchsbedingungen getestet wird. Besser ist es,

ebenso viele Kontrollgruppen wie Versuchsbedingungen vorzusehen. Im Falle eines Versuchsplans mit vier Versuchsbedingungen ergibt sich dann z. B. die in Tabelle 2.9.3.4 dargestellte Designmatrix mit sechs Versuchsgruppen und vier Kontrollgruppen.

Versuchspläne ohne Messwiederholungen

Ganz ohne Messwiederholungen kann der Effekt der Versuchsbedingungen jedoch nicht spezifisch objektiv gemessen werden. In Versuchsplänen mit unabhängigen Stichproben, in denen jede Vp nur unter einer Versuchsbedingung getestet wird, können die Effektparameter nicht von den Personenparametern getrennt werden.

Für die Messung und Testung des Effekts der Versuchsbedingungen bleibt uns nur der Weg der traditionellen Statistik offen, indem wir die unter den jeweiligen Versuchsbedingungen gemessenen Personenparameter $\theta_{vj} = \theta_v + \ln(\beta_j)$ als Daten in eine Varianzanalyse einsetzen. Erfolgt die Zuordnung der Vpn zu den Versuchsgruppen per Zufall, so erhält man auf diese Weise zumindest noch erwartungstreue Schätzungen der experimentellen Effekte.

2.9.4 Hypothesentestung

Test für die Modellstruktur

Ob die in der Designmatrix niedergelegte Modellstruktur geeignet ist, die experimentellen Effekte zu beschreiben, kann getestet werden, indem man die bedingte Likelihood der Datenmatrix unter Annahme des LLTM (H_0: Designmatrix = Q) und die bedingte Likelihood der Datenmatrix unter Annahme des Rasch-Modells (H_1: Designmatrix = E) in Form eines Likelihood-Quotienten-Tests miteinander vergleicht (vgl. Kap. 2.9.2).

Test für die experimentellen Effekte

Darauf aufbauend kann man in einem nächsten Schritt die statistische Signifikanz der experimentellen Effekte testen. Dazu erstellen wir die Designmatrix Q_0, welche sich ergibt, wenn die zu testenden Effekte gleich Null gesetzt werden, und berechnen einen bedingten Likelihood-Quotienten-Test, der die bedingte Likelihood der Datenmatrix unter Annahme der reduzierten Designmatrix (H_0: Designmatrix = Q_0)

$$L_0 = cL(X|Q_0) \qquad [2.9.4.1]$$

und die bedingte Likelihood der Datenmatrix unter Annahme der vollständigen Designmatrix (H_1: Designmatrix = Q)

$$L_1 = cL(X|Q) \qquad [2.9.4.2]$$

einander gegenüberstellt. Die Nullhypothese (= kein experimenteller Effekt) wird verworfen, wenn die asymptotisch nach χ^2 mit

$$df = m - m_0 \qquad [2.9.4.3]$$

verteilte Testgröße

$$-2\ln(\lambda) = 2|\ln(L_0) - \ln(L_1)| \qquad [2.9.4.4]$$

den kritischen Wert der χ^2-Verteilung übersteigt.

Die reduzierte Designmatrix

Die reduzierte Designmatrix Q_0 erhalten wir, indem wir all jene Spalten von Q löschen, deren zugehörige Effektparameter auf Signifikanz getestet werden sollen; m_0 bezeichnet die danach verbleibende Spaltenanzahl.

Auf diese Weise kann man sowohl gezielt einzelne Effekte, als auch mehrere Effekte gleichzeitig oder sämtliche Effekte simultan testen.

2.9.5 Degenerierte Modellvarianten

Einschränkungen des Rasch-Modells

Den pragmatischen, statistischen und wissenschaftstheoretischen Vorzügen des Rasch-Modells (Rechtfertigung der Scorebildung, CML-Schätzung der Strukturparameter und spezifische Objektivität der Vergleiche zwischen Vpn, Items und Versuchsbedingungen) steht der Nachteil gegenüber, dass es maximal drei Arten von Parametern kennt, welche die Wahrscheinlichkeitsdichte der Antwortvariablen bestimmen: Personenparameter, die eine latente Fähigkeit der Vpn messen; Itemparameter, welche die Schwierigkeit (bzw. Leichtigkeit) der Items beschreiben, und Effektparameter, die den Einfluss der Versuchsbedingungen abbilden.

Degenerierte Modellvarianten

Aus psychologischer Sicht mag es deshalb als wünschenswert erscheinen, das Modell um zusätzliche Parameter zu erweitern, die z. B. Ratewahrscheinlichkeiten, erfolgskontingente Lernprozesse oder Unterschiede bezüglich der Trennschärfe der Items abzubilden erlauben. Da das Rasch-Modell die genannten Eigenschaften aber nicht nur de facto aufweist, sondern das *einzig mögliche* Modell ist, das diese Vorzüge besitzt (vgl. Satz 2.3.3.1), sind sol-

che Versuche jedoch von vorneherein zum Scheitern verurteilt. Man kann zwar die Modellparameter dann immer noch mittels der marginalen ML-Methode schätzen, aber die spezifischen Eigenschaften des Rasch-Modells gehen verloren, weshalb man solche Modellerweiterungen auch als *degenerierte Modellvarianten* bezeichnen kann.

Ratewahrscheinlichkeiten

So schlägt z. B. Keats (1974) zur Darstellung der Itemcharakteristik von Testaufgaben, deren richtige Antwort von den Probanden erraten werden kann, die Modellgleichung

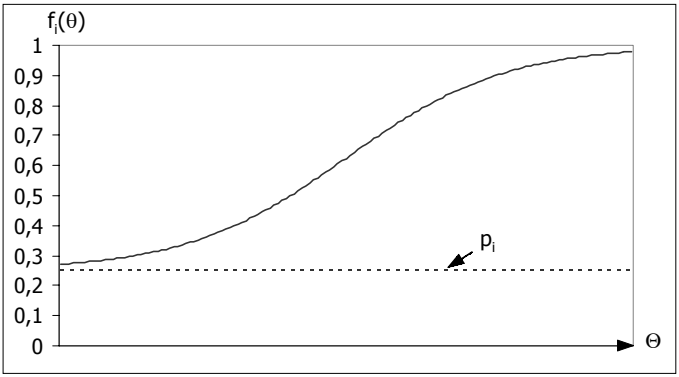
$$f_i(\theta_v) = p_i + (1 - p_i)\frac{\exp(\theta_v - \delta_i)}{1 + \exp(\theta_v - \delta_i)} \qquad [2.9.5.1]$$

vor, worin p_i die Wahrscheinlichkeit bezeichnet, dass die richtige Antwort erraten wird. Mittels der logistischen Funktion

$$\frac{\exp(\theta_v - \delta_i)}{1 + \exp(\theta_v - \delta_i)}$$

wird in diesem Modell also nicht die Wahrscheinlichkeit einer richtigen Antwort beschrieben, sondern die Wahrscheinlichkeit, mit welcher ein Proband die Aufgabenlösung findet, wenn er nicht bloß rät.

Abb. 2.9.5.1: Itemcharakteristik des um Ratewahrscheinlichkeiten erweiterten Rasch-Modells



Der Rateparameter p_i stellt somit die untere Asymptote der Itemcharakteristik dar, die bei keiner noch so niedrigen Ausprägung der Fähigkeit θ unterschritten werden kann (vgl. Abbildung 2.9.5.1). Er kann entweder eine empirisch zu bestimmende Grö-

ße sein oder eine theoretische Größe, die aufgrund des Antwortformats der Items a priori festzusetzen ist. So kann er z. B. mit $p_i = 0{,}25$ angenommen werden, wenn das Antwortformat vier Antwortmöglichkeiten vorsieht, von denen genau eine richtig ist.

Wie Colonius (1977) gezeigt hat, erlaubt das Modell jedoch keine spezifisch objektiven Messungen. Dass die Suffizienz des Summenscores in dem Modell *nicht* gewährleistet sein *kann*, folgt unmittelbar aus Satz 2.3.3.1. Zu zeigen, dass sie de facto nicht gegeben ist, sei dem Leser als Übungsaufgabe überlassen.

Erfolgskontingente Lernprozesse

Als degenerierte Modellvariante erweist sich auch die von Verhelst & Glas (1993) vorgeschlagene dynamische Erweiterung des Rasch-Modells, welche die serielle Abhängigkeit der Antwortvariablen durch die bedingten Itemcharakteristiken

$$f_{i.r}(\theta_v) = \frac{\exp(\theta_v - \delta_i + \beta_{ir})}{1 + \exp(\theta_v - \delta_i + \beta_{ir})} \qquad [2.9.5.2]$$

darstellt, worin der Lerneffekt, der aus der richtigen Beantwortung von r vorangegangenen Items resultiert, durch den Parameter β_{ir} ausgedrückt wird.

Dass die Suffizienz des Summenscores in diesem Modell nicht gewährleistet sein *kann*, folgt unmittelbar aus Satz 2.3.3.2. Zu zeigen, dass sie de facto nicht gegeben ist, sei dem Leser als Übungsaufgabe überlassen.

Virtuelle Items

Dass Verhelst & Glas den Verlust suffizienter Statistiken für die Personenparameter nicht bemerken, ist einem Trick geschuldet, der bereits auf Fischer (1972) zurückgeht und der "überraschend einfach" (Rost, 1996, 282) erscheint, leider jedoch auf einem Denkfehler beruht: Verhelst & Glas meinen nämlich das Konzept der bedingten Itemcharakteristiken mittels der Konstruktion von *virtuellen Items* umgehen und die Modellparameter im Rahmen des linearen logistischen Testmodells schätzen zu können.

Das Konstruktionsprinzip

Um zu verdeutlichen, wie diese virtuellen Items konstruiert werden, betrachten wir einen Test aus k = 3 Items, denen 8 verschiedene Antwortmuster entsprechen (vgl. Tab. 2.9.5.1). Mit i = 1,…,k bezeichnen wir die realen Items, mit r = 0,…,i-1 bezeichnen wir die Anzahl der zuvor gelösten Items und mit j = [i.r] bezeichnen wir das virtuelle Item, das der Bearbeitung des realen Items i entspricht, wenn zuvor r andere Items gelöst wurden.

Tab. 2.9.5.1: Reale und virtuelle Antwortmatrix

Reale Items			Virtuelle Items					
			1	2	3	4	5	6
1	2	3	[1.0]	[2.0]	[2.1]	[3.0]	[3.1]	[3.2]
1	1	1	1	*	1	*	*	1
1	1	0	1	*	1	*	*	0
1	0	1	1	*	0	*	1	*
1	0	0	1	*	0	*	0	*
0	1	1	0	1	*	*	1	*
0	1	0	0	1	*	*	0	*
0	0	1	0	0	*	1	*	*
0	0	0	0	0	*	0	*	*

Da die Vpn vor Bearbeitung des ersten Items noch gar kein anderes Item bearbeitet haben, ist das erste virtuelle Item [1.0] mit dem ersten realen Item identisch und die erste Spalte der virtuellen Antwortmatrix stimmt mit der ersten Spalte der realen Antwortmatrix überein.

Dem zweiten realen Item entsprechen zwei virtuelle Items [2.0] und [2.1]. Welches der beiden virtuellen Items eine Vp bearbeitet, hängt davon ab, ob sie das erste Item gelöst hat oder nicht. Virtuelle Items, welche eine Vp nicht bearbeitet, sind in der Tabelle mit einem Sternchen markiert und werden von Verhelst & Glas wie *Missing Data* behandelt.

Dem dritten realen Item entsprechen drei virtuelle Items [3.0], [3.1] und [3.2] und welches der drei virtuellen Items eine Vp beantwortet, hängt davon ab, welchen partiellen Score sie bei den ersten beiden Items erzielt hat, u.s.w.

Äquivalenz der beiden Darstellungsweisen

Wie man sofort erkennt, ist die Konstruktion virtueller Items lediglich eine andere Darstellungsform für bedingte Itemcharakteristiken, so dass

$$f_{1.r}(\theta_v) = \text{prob}\left(X_{vi} = 1 \middle| x_{vo}^{(i)} = r\right) = \text{prob}\left(x_{vi} = 1 \middle| j = [i.r]\right). \quad [2.9.5.3]$$

Der verhängnisvolle Denkfehler

Da das lineare logistische Testmodell nicht voraussetzt, dass alle Vpn die selbe Auswahl von Items bearbeitet haben, erscheint es auf den ersten Blick plausibel, das Modell [2.9.5.2] als einen Spezialfall des LLTM aufzufassen, wobei die β_{i0} o.B.d.A. zu Null nor-

miert werden können, so dass ihnen in der Designmatrix (vgl. Tabelle 2.9.5.2) keine eigene Spalte entspricht.

Tab. 2.9.5.2: Designmatrix der virtuellen Items

Item-Nr.		$-\delta_1$	$-\delta_2$	$-\delta_3$	β_{21}	β_{31}	β_{32}
		η_1	η_2	η_3	η_4	η_5	η_6
[1.0]	1	1	0	0	0	0	0
[2.0]	2	0	1	0	0	0	0
[2.1]	3	0	1	0	1	0	0
[3.0]	4	0	0	1	0	0	0
[3.1]	5	0	0	1	0	1	0
[3.2]	6	0	0	1	0	0	1

Um die Modellparameter mittels des LLTM schätzen zu können, muss man aber wissen, *welche* virtuellen Items die Vpn jeweils bearbeitet haben. D. h. wir benötigen für jede Vp einen Zuordnungsvektor $g_v = (g_{v1},\ldots,g_{vk'})$, worin k' die Anzahl der virtuellen Items bezeichnet und

$$g_{vj} = \begin{cases} 1 & \text{wenn Vp v das virtuelle Item j bearbeitet hat} \\ 0 & \text{sonst} \end{cases} .$$

Zwischen der Zuordnung der Vpn zu den virtuellen Items und ihren Antworten auf die ersten k-1 Items besteht jedoch eine eindeutige Entsprechung, so dass

$$\text{prob}\{(x_{v1},\ldots,x_{vk-1})|(g_{v1},\ldots,g_{vk'})\} = 1 .$$

Ist die Zuordnung der Vpn zu den virtuellen Items bekannt, dann stehen ihre Antworten auf die ersten k-1 Items fest, so dass

$$\text{prob}\{(x_{v1}\ldots,x_{vk})|\theta_v,(g_{v1},\ldots,g_{vk'})\} = \text{prob}(x_{vk}|\theta_v,g_{vk'})$$

und auch der Summenscore nur noch mit der Antwort auf das letzte Item variiert. Folglich ist

$$\text{prob}\{x_{vo}|\theta_v,(g_{v1},\ldots,g_{vk'})\} = \text{prob}(x_{vk}|\theta_v,g_{vk'})$$

und die bedingte Likelihood des Antwortvektors

$$\text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo},(g_{v1},\ldots,g_{vk'})\}=1$$

hängt von den zu schätzenden Parametern gar nicht mehr ab.

2.9.6 Grenzfälle des Rasch -Modells

Das dynamische Testmodell

Anders als in dem von Verhelst & Glas vorgeschlagenen Modell [2.9.5.2] bleibt im dynamischen Testmodell [2.3.2.19] die Suffizienz des Summenscores zwar erhalten (vgl. Kap. 2.3.2) und damit auch die Separierbarkeit der Item- und Personenparameter gewährleistet. Die Lösung der CML-Schätzgleichungen wirft im dynamischen Testmodell jedoch numerische Probleme auf, die zum einen den linearen Nebenbedingungen $\psi_r \leq \sigma_i$ und zum anderen der Tatsache geschuldet sind, dass die Berechnung der Funktion $G(g,\psi,\sigma)$ (siehe Gleichung [2.3.3.22]) und ihrer partiellen Ableitungen bei großem k mit erheblicher Rechenungenauigkeit belastet ist.

Mangelnde Verallgemeinerbarkeit

Dass das dynamische Testmodell nur als ein Grenzfall des Rasch-Modells einzuschätzen ist, beruht jedoch auf grundsätzlicheren Überlegungen und ist der mangelnden Verallgemeinerbarkeit des Modells geschuldet. So hat Kempf (1977, 48-55) zwar die Übertragbarkeit der Modellstruktur auf andere Formen der seriellen Abhängigkeit aufgezeigt (z. B. Markov-Prozesse), eine Verallgemeinerung des Modells für polytome oder ordinale Antwortvariablen ist jedoch nicht möglich.

Die einzig denkbare Modellstruktur (siehe Satz 2.3.3.2), die bei seriell abhängigen binären Items die Suffizienz des Summenscores zu gewährleisten vermag, scheint somit eher ein methodisches Artefakt des binären Antwortformats zu sein als ein allgemeines Konstruktionsprinzip für spezifisch objektive Testmodelle mit seriell abhängigen Antwortvariablen.

Das Birnbaum-Modell

Ebenfalls einen Grenzfall des Rasch-Modells stellt die von Birnbaum entwickelte Erweiterung des Rasch-Modells um einen Trennschärfeparameter α_i dar:

$$f_i(\theta_v)=\frac{\exp(\alpha_i(\theta_v-\delta_i))}{1+\exp(\alpha_i(\theta_v-\delta_i))} \qquad [2.9.6.1]$$

Wie Birnbaum (1968) zeigt, ist das Modell [2.9.6.1] notwendig und hinreichend dafür, dass der *gewichtete* Summenscore

$$x_{v\bullet} = \sum_{i=1}^{k} \alpha_i x_{vi}$$

eine suffiziente Statistik für den Personenparameter θ_v ist, so dass die bedingte Wahrscheinlichkeit des Antwortvektors einer Vp

$$\text{prob}\{(x_{v1},\ldots,x_{vk})|x_{v\bullet} = g\} = \frac{\text{prob}\{(x_{v1},\ldots,x_{vk})|\theta_v\}}{\sum_{y_v:y_{v\bullet}=g} \text{prob}\{(y_{v1},\ldots,y_{vk})|\theta_v\}} \quad [2.9.6.2]$$

$$= \frac{\prod_{i=1}^{k} \varepsilon_i^{\alpha_i x_{vi}}}{\sum_{y_v:y_{v\bullet}=g} \prod_{i=1}^{k} \varepsilon_i^{\alpha_i y_{vi}}}$$

nur noch von den Gewichtszahlen α_i und den Itemparametern $\varepsilon_i = \exp(-\delta_i)$ abhängt.

Separierbarkeit der Modellparameter

Sind die Trennschärfeparameter α_i bekannt, dann hat das Birnbaum-Modell somit dieselben wünschenswerten Eigenschaften wie das Rasch-Modell. Die Separierbarkeit der Modellparameter ist gewährleistet und die Itemschwierigkeiten δ_i können unabhängig von den Personenpareametern θ_v geschätzt werden, indem man die Maximum-Likelihood-Methode auf die bedingte Likelihood der Antwortmatrix

$$cL(X) = \prod_{v=1}^{n} \text{prob}\{(x_{v1},\ldots,x_{vk})|x_{v\bullet}\} \quad [2.9.6.3]$$

$$= \prod_{v=1}^{n} \frac{\prod_{i=1}^{k} \varepsilon_i^{\alpha_i x_{vi}}}{\sum_{y_v:y_{v\bullet}=x_{v\bullet}} \prod_{i=1}^{k} \varepsilon_i^{\alpha_i y_{vi}}}$$

$$= \frac{\prod_{i=1}^{k} \varepsilon_i^{\alpha_i x_{oi}}}{\prod_g \left(\sum_{y_v:y_{v\bullet}=g} \prod_{i=1}^{k} \varepsilon_i^{\alpha_i y_{vi}} \right)^{n_g}}$$

anwendet, worin sich das Produktzeichen im Nenner über alle möglichen Scores erstreckt und n_g die Anzahl der Vpn bezeichnet, deren Score $x_{v\bullet} = g$ ist. x_{oi} steht für die Anzahl der Vpn, welche Item i gelöst haben.

Grenzen der Anwendbarkeit

Damit dies möglich ist, müssen jedoch zwei Voraussetzungen erfüllt sein:

1. Die Gewichtszahlen a_i müssen *a priori* bekannt sein, da man andernfalls die gewichteten Summenscores $x_{v\bullet}$ nicht berechnen kann.
2. Die Gewichtszahlen müssen so beschaffen sein, dass die Antwortvektoren der Vp durch die zugehörigen gewichteten Summenscores nicht eindeutig determiniert sind.

Ganzzahlige Trennschärfeparameter

Als Gewichtszahlen kommen daher nur ganzzahlige Trennschärfeparameter in Betracht, wodurch der Test in m Subtests t = 1,…,m zerfällt, innerhalb derer alle Items die selbe Trennschärfe $\alpha_i = t$ besitzen. Ist dies der Fall, so ist innerhalb jedes der m Subtests das Rasch-Modell erfüllt und daher jeder der (ungewichteten) Subtestscores

$$x_{vo}^{(t)} = \sum_{i \in t} x_{vi}$$

eine suffiziente Statistik für den Personenparameter θ_v.

Identifizierbarkeit der Trennschärfeparameter

Zumindest theoretisch könnte man diese Eigenschaft der Subtests verwenden, um mit Hilfe des Rasch-Modells zu identifizieren, *welche Items* jeweils einen Subtest bilden. Darauf aufbauend könnte die Zuordnung der (ganzzahligen) Trennschärfeparameter zu den Subtests schließlich so vorgenommen werden, dass man sich für jene Zuordnung entscheidet, bei welcher die (marginale) Likelihood der Datenmatrix den größten Wert annimmt.

Praktisch erweist sich aber auch dies nicht als handhabbar, weil die Anzahl der Möglichkeiten, wie aus k Items m Subtests gebildet werden können, zu rasch ansteigt. Die Separierbarkeit der Modellparameter bleibt im Birnbaum-Modell somit eine bloß theoretische Modelleigenschaft, deren praktische Umsetzung an mangelnder Berechenbarkeit scheitert.

2.9.7 Mischverteilungs-Rasch-Modelle

Die Restriktivität des Rasch-Modells

Trotz seiner stochastischen Modellstruktur ist das Rasch-Modell relativ restriktiv. Die Anforderung, dass alle Items eines Tests dieselbe latente Dimension messen müssen, wird von den meisten gängigen Tests nicht erfüllt und die Messdimensionen, die sich mit Hilfe des Rasch-Modells operationalisieren lassen, sind in der Regel sehr eng definiert. Der Summenscore vermag die Testleistung der Probanden meist nur dann erschöpfend zu beschreiben, wenn der Test aus sehr ähnlichen Items aufgebaut ist, die

sich inhaltlich kaum voneinander unterscheiden. Von der Vorstellung, so umfassende latente Eigenschaftsdimensionen wie z. B. "die Intelligenz schlechthin" messen zu können, wird man sich in der Psychologie wohl verabschieden müssen – zumindest dann, wenn man an den Begriff der Messung so strenge Kriterien anlegt wie in anderen Naturwissenschaften und sich nicht schon mit *ad hoc*-Quantifizierungen zufrieden gibt.

Unterschiedliche Lösungsstrategien

Selbst bei Aufgaben, die inhaltlich sehr ähnlich sind, kann es sein, dass die Probanden an die Aufgabenlösung mit unterschiedlichen Strategien herangehen. So kann man z. B. die Würfelaufgaben des IST-70, die so etwas wie räumliches Vorstellungsvermögen erfassen sollen, nicht nur auf dem Wege der Anschauung, sondern auch analytisch zu lösen versuchen.

Je nachdem, mit welcher Strategie man die Aufgabe zu lösen versucht, sind somit unterschiedliche Fähigkeiten gefordert und Aufgaben, die mit der einen Strategie relativ leicht zu lösen sind, können sich bei Anwendung der anderen Strategie als schwierig erweisen.

Kurz und gut: Es ist damit zu rechnen, dass die Würfelaufgaben keinen homogenen Subtest bilden, da sie je nach bevorzugter Lösungsstrategie verschiedene Fähigkeiten messen, so dass man eigentlich nur jene Vpn aufgrund ihrer Leistungsmenge miteinander vergleichen kann, die sich der selben Strategie bedienen. Und es ist damit zu rechnen, dass die selbe Aufgabe nicht für alle Vpn auch die gleiche Schwierigkeit hat.

Es kann daher weder gefordert werden, dass die Vergleiche zwischen Vpn von der Auswahl der Items unabhängig sind, noch kann gefordert werden, dass die Vergleiche zwischen Items von der Auswahl der Probanden unabhängig sind.

Die Modellannahmen

Solche Überlegungen, die das Prinzip der spezifischen Objektivität für obsolet zu erklären scheinen, bilden die Grundlage der auf Rost (1990), Mislevy & Verhelst (1990) und Kelderman & Macready (1990) zurückgehenden Mischverteilungs-Rasch-Modelle, die wir aus Gründen der mathematischen Einfachheit hier nur in ihrer einfachsten Form diskutieren wollen: das zweifaktorielle Modell für binäre Items

$$f_{ig}(\theta_v) = \text{prob}\{X_{vi} = 1 | v \in g\} = \frac{\exp(\theta_v - \delta_{ig})}{1 + \exp(\theta_v - \delta_{ig})}, \qquad [2.9.7.1]$$

das die Vpn in h latente Klassen g = 1,…,h einteilt und die Itemschwierigkeiten δ_{ig} als Funktion der Klassenzugehörigkeit auffasst. Für die Wahrscheinlichkeit, mit der eine zufällig herausgegriffene Vp das Item i löst, erhält man dann den Ausdruck

$$\text{prob}(X_{vi} = 1) = \sum_{g=1}^{h} p_g \frac{\exp(\theta_v - \delta_{ig})}{1 + \exp(\theta_v - \delta_{ig})}, \qquad [2.9.7.2]$$

worin die

$$p_g = \text{prob}(v \in g) \qquad [2.9.7.3]$$

die Klassengrößen bezeichnen.

Ein zweistufiger diagnostischer Prozess

Tatsächlich hat man damit aber das Prinzip der spezifischen Objektivität nicht aufgegeben, sondern gerade um spezifische Objektivität gewährleisten zu können, einen zweistufigen diagnostischen Prozess eingeleitet:

- Zuerst ist zu entscheiden, welcher latenten Klasse die Vpn angehören.
- Erst dann können die Mitglieder der selben Klasse hinsichtlich ihrer Leistungsmenge spezifisch objektiv miteinander verglichen werden.

Um auf unser obiges Beispiel zurückzukommen: Zuerst ist zu identifizieren, welche Lösungsstrategie ein Proband bevorzugt, erst dann können wir messen, wie gut er diese Strategie beherrscht.

Kontrollfragen

- Worin besteht die mehrfaktorielle Erweiterung des Rasch-Modells?
- Welche Skaleneigenschaft haben die Effektparameter im mehrfaktoriellen Rasch-Modell und welche Normierungsbedingungen ergeben sich daraus?
- Erläutern Sie die methodische Vorgehensweise, wie man in Versuchsplänen mit vollständigen Wechselwirkungen (durchschnittliche) Haupteffekte der Faktorstufen berechnen kann!
- Welche Möglichkeiten gibt es, in Versuchsplänen mit partiellen Wechselwirkungen die Effekte der Versuchsbedingungen darzustellen, und welche Voraussetzungen müssen erfüllt sein, damit man die gemeinsamen Effekte interagierender Faktorstufen in Haupteffekte und Wechselwirkungen zerlegen kann?
- Worin besteht der Grundgedanke des linearen logistischen Testmodells?

- Welche Skaleneigenschaft besitzen die Basisparameter im linearen logistischen Testmodell?
- Wodurch werden im linearen logistischen Testmodell die Hypothesen über die Aufgabenstruktur dargestellt?
- Welche Voraussetzung muss im linearen logistischen Testmodell die Designmatrix erfüllen?
- Bei welcher Designmatrix ergibt sich
  a) das Binomialmodell
  b) das Rasch-Modell
  als Spezialfall des linearen logistischen Testmodells?
- Mittels welcher Methode können die Parameter des linearen logistischen Testmodells geschätzt werden?
- Welche Voraussetzung muss erfüllt sein, damit man die Hypothesen bezüglich der Aufgabenstruktur mittels eines bedingten Likelihood-Quotienten-Tests prüfen kann?
- Wie wird zwischen Aufgaben und Items unterschieden, wenn man das mehrkategorielle Rasch-Modell in Form des linearen logistischen Testmodells darstellt?
- Welche Voraussetzung(en) muss ein Versuchsplan erfüllen, damit man die Effekte der Versuchsbedingungen unabhängig von der Auswahl der Personenstichprobe(n) messen kann?
- Wie kann man die Signifikanz der experimentellen Effekte überprüfen?
- Was versteht man unter degenerierten Modellvarianten des Rasch-Modells?
- Nennen Sie Beispiele für degenerierte Modellvarianten!
- Woran scheitert die Anwendung des LLTM zur Parameterschätzung im Modell von Glas & Verhelst?
- Inwiefern hat das dynamische Testmodell die selben wünschenswerten Eigenschaften wie das Rasch-Modell und warum ist es dennoch nur als Grenzfall des Rasch-Modells einzuschätzen?
- Inwiefern hat das Birnbaum-Modell die selben wünschenswerten Eigenschaften wie das Rasch-Modell und warum ist es dennoch nur als Grenzfall des Rasch-Modells einzuschätzen?
- Diskutieren Sie, inwiefern die Mischverteilungs-Rasch-Modelle dem Prinzip der spezifischen Objektivität folgen (oder auch nicht)!

Übungsaufgaben

1. Gegeben sei ein 3x3-Versuchsplan ohne Wechselwirkungen.
   a) Erstellen Sie die Gewichtszahlenmatrix und stellen Sie die logarithmischen Effektparameter $\ln(\beta_j)$ durch die Effekte der Faktorstufen dar!
   b) Bestimmen Sie die Anzahl der unabhängigen Effektparameter!
2. Gegeben sei ein 2x2-Versuchsplan mit vollständigen Wechselwirkungen.
   a) Erstellen Sie die Gewichtszahlenmatrix und stellen Sie die logarithmischen Effektparameter $\ln(\beta_j)$ durch die gemeinsamen Effekte der interagierenden Faktorstufen dar!
   b) Bestimmen Sie die Anzahl der unabhängigen Effektparameter!
3. In einem Versuchsplan mit vollständigen Wechselwirkungen haben

sich die in nachstehenden Tabelle dargestellten Effektparameter $\omega_{b,b'}^{(a,a')}$ ergeben:

	b' = 0	b' = 1	b' = 2	b' = 3
b = 0	0.0	-2.8	-2.0	-0.8
b = 1	2.3	0.4	0.7	0.9
b = 2	1.1	-0.3	0.1	0.3
b = 3	0.5	-2.4	-1.3	2.9

a) Berechnen Sie (durchschnittliche) Haupteffekte der Faktoren sowie b) die zugehörigen Interaktionseffekte!

4. Gegeben sei ein 3x3-Versuchsplan mit partiellen Wechselwirkungen zwischen den Faktorstufen $b_1 = 1$ und $b_2 = 2$ sowie zwischen $b_1 = 2$ und $b_2 = 1$.
   a) Erstellen Sie die Gewichtszahlenmatrix zur Beschreibung des Effekts der Versuchsbedingungen durch Haupteffekte und gemeinsame Effekte!
   b) Erstellen Sie die Gewichtszahlenmatrix zur Beschreibung des Effekts der Versuchsbedingungen durch Haupteffekte und Wechselwirkungen!
   c) Bestimmen Sie die Anzahl der unabhängigen Effektparameter der beiden Modellvarianten!
5. Zeigen Sie, dass die Anwendung der bedingten Maximum-Likelihood Methode auf das lineare logistische Testmodell zu den Schätzgleichungen [2.9.2.3] führt!
6. Gegeben sei ein Test aus vier binären Items, für welchen das Rasch-Modell erfüllt ist. Die bedingte Likelihood der Daten unter Annahme des Rasch-Modells betrage $\ln\{cL(X|\varepsilon_1,\ldots,\varepsilon_4)\} = -407.3737$. Ihre Hypothese lautet, dass die Aufgabenlösung durch zwei Denkoperationen erklärt werden kann, von denen die erste nur zur Lösung der Aufgaben i = 1 und i = 4, die zweite nur zur Lösung der Aufgaben i = 2 und i = 3 erforderlich ist.
   a) Erstellen Sie die Designmatrix, die diese Hypothese im linearen logistischen Testmodell abbildet, und
   b) testen Sie, ob die Hypothese beibehalten werden kann, wenn die bedingte Likelihood des linearen logistischen Testmodells $\ln\{cL(X|\eta_1,\eta_2)\} = -428.4119$ beträgt!
7. Zeigen Sie am Beispiel eines Tests mit k = 2 Items, dass der Summenscore in dem um Ratewahrscheinlichkeiten erweiterten Rasch-Modell [2.9.5.1] keine erschöpfende Beschreibung der Testleistung eines Probanden erlaubt!
8. Zeigen Sie am Beispiel eines Tests mit k = 2 Items, dass der Summenscore in dem um einen erfolgskontingenten Lernparameter erweiterten Rasch-Modell [2.9.5.2] keine erschöpfende Beschreibung der Testleistung eines Probanden erlaubt!

2.10

Identifikation der Klassenanzahl, Parameterschätzung und Beurteilung der Testleistung der Probanden in der Latent-Class-Analyse

2.10.1 Beurteilung der Testleistung

Marginale Likelihood

In der Latent-Class-Analyse hängt die (marginale) Likelihood der Datenmatrix von zwei Arten von Parametern ab: von den Klassengrößen $p_g = \text{prob}(v \in g)$ und von den (klassenspezifischen) Kategorienwahrscheinlichkeiten $p_{gix} = \text{prob}(X_{vi}=x|v \in g)$.

Unter Verwendung der Schreibweise

$$z_{vix} = \begin{cases} 1 & \text{wenn Vp v Item i in Kategorie x beantwortet} \\ 0 & \text{sonst} \end{cases} \qquad [2.10.1.1]$$

kann die Modellgleichung [2.2.3.14] daher auch in der Form

$$mL(X) = \prod_{v=1}^{n} \text{prob}\{(x_{v1},\ldots,x_{vk})\} \qquad [2.10.1.2]$$

$$= \prod_{v=1}^{n} \left(\sum_{g=1}^{h} \text{prob}(v \in g)\, \text{prob}\{(x_{v1},\ldots,x_{vk})|v \in g\} \right)$$

$$= \prod_{v=1}^{n} \left(\sum_{g=1}^{h} p_g \prod_{i=1}^{k} \prod_{x=0}^{m_i-1} p_{gix}^{z_{vix}} \right)$$

dargestellt werden.

Membership-Wahrscheinlichkeiten

Sind die Modellparameter bekannt, kann die Beurteilung der Testleistung eines Probanden relativ einfach erfolgen, indem man die Vp jener Klasse zuordnet, der sie aufgrund ihres Antwortmusters mit größter Wahrscheinlichkeit angehört.

Dazu berechnet man für jede der Klassen g = 1,…,h die sog. *Membership-Wahrscheinlichkeit* der Vp

$$p_{g.x_v} = \text{prob}\{v \in g | (x_{v1},\ldots,x_{vk})\} \qquad [2.10.1.3]$$

$$= \frac{\text{prob}(v \in g)\text{prob}\{(x_{v1},\ldots,x_{vk}) | v \in g\}}{\sum_{d=1}^{h}\text{prob}(v \in d)\text{prob}\{(x_{v1},\ldots,x_{vk}) | v \in d\}}$$

$$= \frac{p_g \prod_{i=1}^{k}\prod_{x=0}^{m_i-1} p_{gix}^{z_{vix}}}{\sum_{d=1}^{h} p_d \prod_{i=1}^{k}\prod_{x=0}^{m_i-1} p_{dix}^{z_{vix}}}$$

und ordnet die Vp schließlich jener Klasse zu, deren Membership-Wahrscheinlichkeit am größten ist.

2.10.2 Parameterschätzung

Umgehung des Neyman-Scott-Problems

Um die Membership-Wahrscheinlichkeiten berechnen zu können, müssen zuerst die Modellparameter p_g und p_{gix} geschätzt werden. Auf den ersten Blick erscheint dies relativ unproblematisch. Da wir es mit einer endlichen Anzahl von latenten Klassen zu tun haben und die Antwortvariablen aller Vpn, die der selben Klasse g angehören, die selben Kategorienwahrscheinlichkeiten p_{gix} besitzen, kann die Personenstichprobe beliebig vergrößert werden, ohne dass die Anzahl der zu schätzenden Modellparameter anwächst.

Während das für Latent-Trait-Modelle so verhängnisvolle Neyman-Scott-Problem in der LCA somit elegant umgangen wird, stellt sich hier jedoch ein Schätzproblem ganz anderer Art, bis zu dessen Lösung es fast ein Vierteljahrhundert gedauert hat.

Manifeste Klassen

Wäre die Klassenzugehörigkeit der Probanden a priori bekannt, so könnten die Modellparameter p_g und p_{gix} mittels Maximum-Likelihood-Methode geschätzt werden, indem man die relativen Häufigkeiten

$$\hat{p}_g = \frac{n_g}{n} \qquad [2.10.2.1]$$

und

$$\hat{p}_{gix} = \frac{n_{gix}}{n_g} \qquad [2.10.2.2]$$

auszählt, worin n_g die Anzahl der Vpn bezeichnet, welche der Klasse g angehören, und n_{gix} für die Anzahl der Vpn steht, die Klasse g angehören *und* Item i in Kategorie x beantworten.

Fehlerstreuung der Parameterschätzungen

Für die Fehlerstreuung der Parameter ergibt sich entsprechend:

$$\sigma(\hat{p}_g) = \sqrt{\frac{p_g(1-p_g)}{n}} \qquad [2.10.2.3]$$

und

$$\sigma(\hat{p}_{gix}) = \sqrt{\frac{p_{gix}(1-p_{gix})}{n_g}}\,. \qquad [2.10.2.4]$$

Latente Klassen

Da wir es jedoch mit latenten Klassen zu tun haben, wissen wir nicht schon a priori, welche Vp zu welcher Klasse gehört, und die Häufigkeiten n_g und n_{gix} können daher nicht einfach ausgezählt werden. Sie sind unbekannt.

Eine zufriedenstellende Lösung des Problems wurde erst von Goodman (1974, 1979) gefunden. Später hat Andersen (1982) gezeigt, dass es sich dabei um einen Spezialfall des EM-Algorithmus handelt, der in seiner allgemeinen Form von Dempster et al. (1977) untersucht wurde.

Der EM-Algorithmus

Der EM-Algorithmus ist ein iteratives Verfahren zur ML-Schätzung, das z. B. auch zur Lösung der Schätzgleichungen für die Item- und Personenparameter des Rasch-Modells (vgl. Kap. 2.7) eingesetzt werden kann. Im Falle der Latent-Class-Analyse sieht die Vorgehensweise wie folgt aus:

1. Zunächst werden – z. B. mit Hilfe eines Zufallsgenerators – vorläufige Schätzwerte (Startwerte) der Parameter p_g und p_{gix} festgesetzt.
2. *E-Schritt* (Erwartungswerte): Mittels dieser Parameterwerte und der daraus ermittelten Membership-Wahrscheinlichkeiten werden in einem nächsten Schritt die erwarteten Häufigkeiten

$$e_g = \sum_{v=1}^{n} p_{g.x_v} \qquad [2.10.2.5]$$

und

$$e_{gix} = \sum_{v:x_{vi}=x} p_{g.x_v} \qquad [2.10.2.6]$$

berechnet, wobei sich das Summenzeichen in [2.10.2.6] über alle Vpn erstreckt, die Item i in Kategorie x beantwortet haben.

3. *M-Schritt* (Maximum-Likelihood): Aus den erwarteten Häufigkeiten werden dann mittels der Gleichungen

$$\hat{p}_g = \frac{e_g}{n} \qquad [2.10.2.7]$$

und

$$\hat{p}_{gix} = \frac{e_{gix}}{e_g} \qquad [2.10.2.8]$$

verbesserte Schätzwerte der Parameter p_g und p_{gix} berechnet.

4. Die verbesserten Schätzwerte werden schließlich als neue vorläufige Schätzwerte in den E-Schritt eingesetzt und das ganze Verfahren wird so lange wiederholt, bis der Logarithmus der marginalen Likelihood

$$\ln\{mL(X)\} = \sum_{v=1}^{n} \ln\left\{\sum_{g=1}^{h} p_g \prod_{i=1}^{k} \prod_{x=0}^{m_i-1} p_{gix}^{z_{vix}}\right\} \qquad [2.10.2.9]$$

nicht mehr weiter wächst und die Likelihood somit ihr Maximum erreicht hat.

Anzahl der unabhängigen Modellparameter

Insgesamt werden auf diese Art und Weise h Klassengrößen p_g und pro Klasse und Item jeweils m_i Kategorienwahrscheinlichkeiten p_{gix} berechnet. Aufgrund der Gleichungen

$$\sum_{g=1}^{h} p_g = 1 \qquad [2.10.2.10]$$

und

$$\sum_{x=0}^{m_i-1} p_{gix} = 1 \qquad [2.10.2.11]$$

sind aber nur h-1 Klassengrößen und jeweils m_i-1 Kategorienwahrscheinlichkeiten unabhängig voneinander, so dass sich die *Anzahl der unabhängigen Modellparameter* zu

$$n(P_{LCh}) = h - 1 + h\sum_{i=1}^{k}(m_i - 1) \qquad [2.10.2.12]$$

ergibt.

Freiheitsgrade der Parameterschätzung

Während die Anzahl der unabhängigen Modellparameter mit wachsender Klassenzahl linear zunimmt, ist die Anzahl der *zulässigen* Parameter jedoch beschränkt und darf jene des saturierten Modells

$$n(P_{sat}) = \prod_{i=1}^{k} m_i - 1 \qquad [2.10.2.13]$$

nicht übersteigen. In Frage kommen daher nur solche LCA-Modelle, deren Freiheitsgrade

$$df = n(P_{sat}) - n(P_{LCh}) \qquad [2.10.2.14]$$

nicht negativ sind.

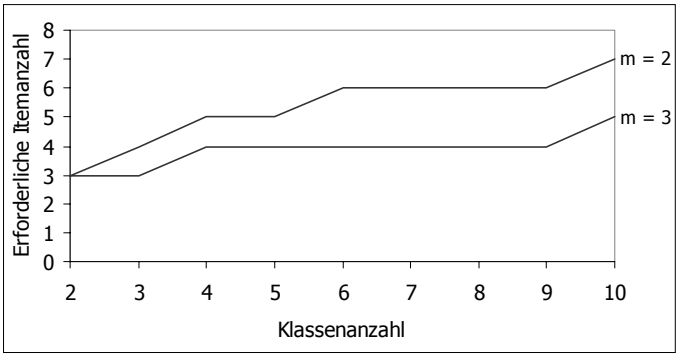


Abb. 2.10.2.1: Zur Identifikation von h Klassen erforderliche Itemanzahl

Wie Abbildung 2.10.2.1 zeigt, wächst die zur Identifikation von h Klassen erforderliche Itemanzahl mit wachsender Klassenzahl umso rascher an, je kleiner die Kategorienzahl der Items ist.

2.10.3 Identifikation der Klassenanzahl

Auch nach Lösung des Schätzproblems bleibt noch immer die Frage offen, wie viele latente Klassen tatsächlich zu unterscheiden sind. Um dies zu entscheiden, geht man so vor, dass man die LCA zunächst probeweise für die Klassenanzahlen h = 1,2,3,...

berechnet und daraus dann jenes Modell (= jene Klassenanzahl) auswählt, welches die Daten am besten beschreibt.

Der AIC-Index

Man kann aber nicht einfach die Klassenzahl wählen, unter welcher die Datenmatrix die größte (marginale) Likelihood besitzt. Da die Likelihood eine monoton wachsende Funktion der Anzahl der latenten Klassen ist, würde dies stets zu einer Entscheidung zugunsten des saturierten Modells führen.

Je mehr latente Klassen man annimmt, desto mehr Modellparameter sind aus den Daten zu schätzen; je mehr Modellparameter zu schätzen sind, desto weniger wird die Verteilung der Datenmatrix restringiert; und je weniger die Verteilung der Datenmatrix restringiert wird, desto größer wird ihre Likelihood.

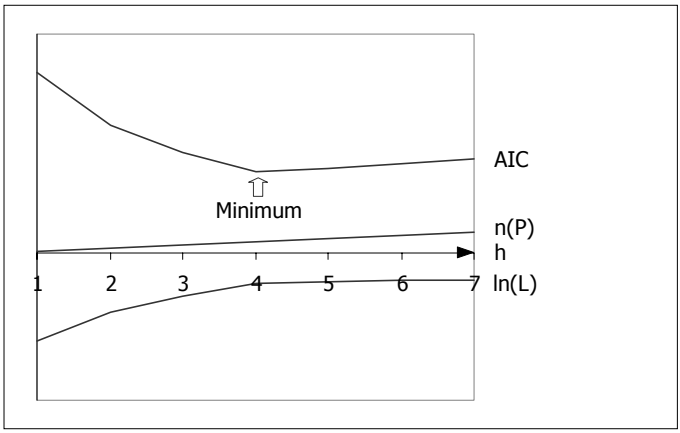
Als Kriterium für die Güte der Beschreibung muss man daher ein Maß wählen, das die Likelihood der Datenmatrix und die Anzahl der zu schätzenden Modellparameter zueinander in Beziehung setzt.

Ein solches Maß ist der auf Akaike (1987) zurückgehende AIC-Index

$$AIC = 2\left|\ln\{mL(X)\}\right| + 2\,n(P_{LCh}), \qquad [2.10.3.1]$$

worin $n(P_{LCh})$ die nach Gleichung [2.10.2.12] berechnete Anzahl der unabhängigen Modellparameter bezeichnet.

Abb. 2.10.3.1: Logarithmierte Likelihood, Parameteranzahl und AIC-Index



Da die Likelihood eine Zahl zwischen Null und Eins ist, ist ihr natürlicher Logarithmus $\ln\{mL(X)\} \leq 0$ und dem Betrage nach umso

kleiner je größer die Likelihood ist. Der AIC-Index ist daher umso kleiner,

- je genauer die Daten beschrieben werden (große Likelihood)
- und je sparsamer die Beschreibung ist (geringe Parameterzahl).

Als beste Beschreibung der Daten wird man daher jene Klassenanzahl auswählen, welche den AIC-Index *minimiert*. (vgl. Abb. 2.10.3.1).

Alternative Informationsmaße

Alternativ zu Akaikes Informationskriterium, das den Logarithmus der Likelihood und die Anzahl der Modellparameter gleich stark gewichtet, kann man sich auch anderer Informationsmaße bedienen, die je nach Gewichtung der Parameteranzahl eher größere oder kleinere Klassenanzahlen favorisieren (vgl. Abb. 2.10.3.2).

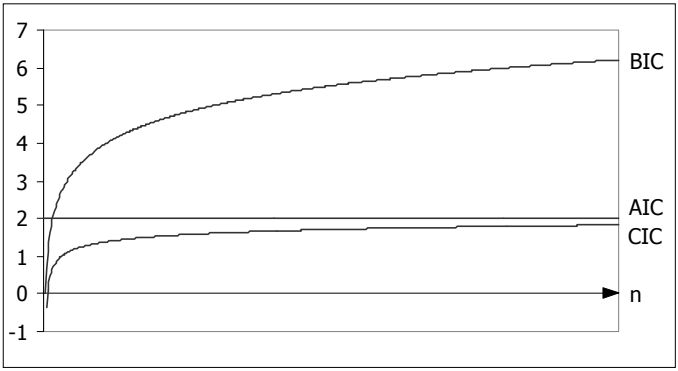


Abb. 2.10.3.2: Gewichtung der Parameteranzahl durch AIC-, BIC- und CIC-Index

So kann es sich in sehr großen Stichproben als nützlich erweisen, die bevorzugte Anzahl der Klassen niedrig zu halten, indem man den von Bozdogan (1987) vorgeschlagenen BIC-Index

$$\mathrm{BIC} = 2\left|\ln\{\mathrm{mL}(X)\}\right| + \ln(n)\, n(P_{LCh}) \qquad [2.10.3.2]$$

verwendet, der die Parameteranzahl mit dem Logarithmus des Stichprobenumfanges gewichtet. Der CIC-Index

$$\mathrm{CIC} = 2\left|\ln\{\mathrm{mL}(X)\}\right| + \ln\{\ln(n)\}\, n(P_{LCh}) \qquad [2.10.3.3]$$

dagegen favorisiert größere Klassenanzahlen als der AIC.

2.10.4 Modellvoraussetzungen

Ein Datenbeispiel

Tabelle 2.10.4.1 stellt die Goodness-of-Fit-Statistiken der Latent-Class-Analyse der in Kapitel 2.2.3 analysierten Daten von Immich et al. (1971) dar.

Tab. 2.10.4.1: Goodness-of-Fit-Statistiken der Latent-Class-Analyse

h	ln{mL(X)}	n(P)	df	AIC	BIC	CIC
1	-713.875	5	26	1437.750	1455.111	1436.249
2	-653.319	11	20	1328.637	1366.833	1325.334
3	-594.919	17	14	1223.838	1282.867	1218.733
4	-559.784	23	8	1165.567	1245.430	1158.661
5	-557.190	29	2	1172.379	1273.076	1163.671
6	–	35	-4	–	–	–
Sat.	-554.183	31				

Wie wir daraus ersehen können, erlauben die fünf dichotomen Variablen die Identifikation von maximal fünf latenten Klassen. Ab h = 6 wird die Anzahl der Freiheitsgrade negativ und die Modellparameter sind nicht mehr eindeutig identifizierbar.

Der Betrag der logarithmierten Likelihood nimmt mit wachsender Klassenanzahl monoton ab und die Informationsmaße verlaufen mit zunehmender Klassenanzahl zunächst ebenfalls monoton fallend, bis sie bei h = 4 ihr Minimum erreichen und danach wieder monoton steigen.

Lokale Maxima

Die Berechnung der LCA führt aber nicht in jedem Fall zu so idealen Ergebnissen, denn die (marginale) Likelihood der Datenmatrix hat mitunter nicht nur ein Maximum, sondern mehrere. Die (logarithmierte) Likelihoodfunktion sieht dann aus wie eine (multidimensionale) Gebirgslandschaft mit einem Hauptgipfel (= globales Maximum) und einem oder mehreren Nebengipfeln (= lokale Maxima). Welchen Gipfel man erklimmt, hängt davon ab, von welchem Ausgangspunkt (= Startwerte des EM-Algorithmus) man den Aufstieg beginnt (vgl. Abbildung 2.10.4.1).

Beseitigung lokaler Maxima

Dass man anstelle des globalen Maximums ein lokales Maximum gefunden hat, lässt sich oft schon daran erkennen, dass die Likelihoodfunktion mit wachsender Klassenanzahl nicht monoton zunimmt (vgl. Abbildung 2.10.4.2). In diesem Falle muss man die Berechnungen für die entsprechende Klassenanzahl so lange mit anderen Startwerten wiederholen, bis das lokale Maximum beseitigt ist.

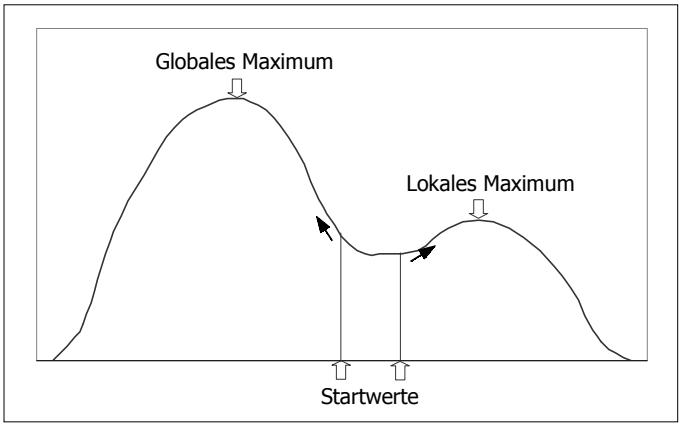


Abb. 2.10.4.1: Likelihoodfunktion mit lokalem Maximum

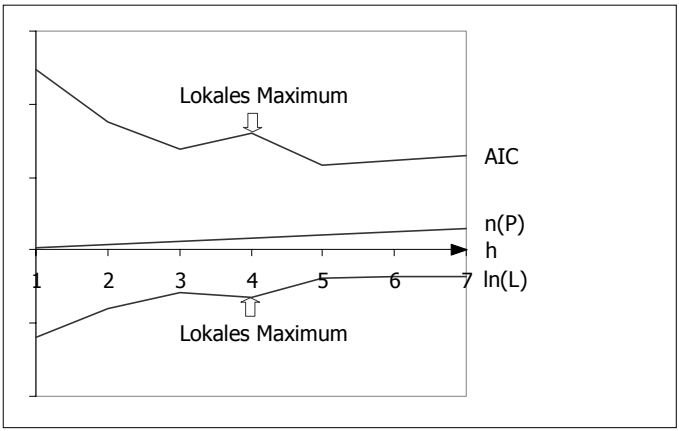


Abb. 2.10.4.2: Identifikation lokaler Maxima anhand der Likelihoodfunktion

Lokale Maxima sind aber nicht in jedem Falle an Irregularitäten der Likelihoodfunktion erkennbar, weshalb man die Berechnungen immer mit mehreren Startwerten durchführen sollte.

Eindeutigkeit der Latent-Class-Analyse

Im Extremfall ist es sogar möglich, dass die Likelihoodfunktion nicht nur ein, sondern *mehrere globale Maxima* besitzt. In diesem Falle liefert jedes der entsprechenden Berechnungsergebnisse eine gleichermaßen gute Beschreibung der Daten und es gibt kein (statistisches) Kriterium, anhand dessen man das eine gegenüber den anderen bevorzugen könnte.

Identifizierbarkeit der Modellparameter

In der Literatur wird das Eindeutigkeitsproblem unter dem etwas irreführenden Stichwort der Identifizierbarkeit der Modellparameter abgehandelt. Hinreichende Bedingungen dafür wurden von McHugh (1956, 1958) formuliert. Demnach sind die Modellparameter p_g und p_{gix} eindeutig identifizierbar, wenn

(i) $h_{max} - 1 \geq n(P_{LCh})$,

(ii) $\sum_{x_v} prob(x_v) = 1$,

(iii) die Funktionen $prob(x_v)$ stetig sind und stetige erste und zweite Ableitungen nach den Parametern p_g und p_{gix} besitzen, und

(iv) mindestens $n(P_{LCh})$ der $prob(x_v)$ funktional unabhängig sind.

Während die ersten beiden Bedingungen durch bloßes Zählen überprüft werden können, ist die Überprüfung der beiden anderen etwas komplizierter (vgl. Rindskopf, 1987, 84).

Bedingung (i) ist eine notwendige Bedingung, die daraus folgt, dass die Anzahl der unabgängigen Schätzgleichungen nicht kleiner sein darf als die Anzahl der zu schätzenden Parameter.[1] Bedingung (ii) ist redundant und folgt unmittelbar aus den Modellannahmen.

Bedingung (iii), deren Verletzung nach McHugh (1956, 227) jedoch eher unwahrscheinlich ist, kann erst nach erfolgter Parameterschätzung überprüft werden. Bedingung (iv) ergibt sich daraus, dass die

$$prob(x_v) = prob\{(x_{v1}, \ldots, x_{vk})\} \qquad [2.10.4.1]$$
$$= \sum_{g=1}^{h} p_g \prod_{i=1}^{k} \prod_{x=0}^{m_i - 1} p_{gix}^{z_{vix}}$$

Polynome in p_g und p_{gix} sind, so dass die Jacoby-Matrix

$$J = \frac{\partial\, prob(x_v)}{\partial\, p_u} \qquad [2.10.4.2]$$

(worin p_u den Vektor aller $n(P_h)$ unabhängigen Modellparameter bezeichnet) maximalen Rang haben muss.

1. Dies ist äquivalent mit der bereits in Kapitel 2.10.2 erwähnten Bedingung, wonach die Freiheitsgrade nicht negativ sein dürfen.

Normalerweise ist die Identifizierbarkeit der Modellparameter umso weniger problematisch, je größer die Anzahl der Probanden ist. Als Faustregel schlägt Formann (1984, 30) daher vor, dass der Stichprobenumfang n stets größer als die Anzahl der unabhängigen Modellparameter sein sollte.

Eine Faustregel

2.10.5 Weiterverrechnung der LCA-Ergebnisse

Klassenscores

Durch die Latent-Class-Analyse werden die Probanden in h nominal skalierte Klassen eingeteilt, so dass jede Vp v jener Klasse g zugeordnet wird, für welche sie die größte Membership-Wahrscheinlichkeit besitzt (vgl. Kap. 2.10.1). Unter Verwendung der Schreibweise

$$y_{vg} = \begin{cases} 1 & \text{wenn } p_{g.x_v} \text{ maximal ist} \\ 0 & \text{sonst} \end{cases} \qquad [2.10.5.1]$$

kann der Klassenscore q_V, den die Vp derart erhält, auch in der Form

$$q_v = \prod_{g=1}^{h} g^{y_{vg}} \qquad [2.10.5.2]$$

dargestellt werden.

Statistische Verfahren für qualitative Variablen

Für die statistische Weiterverrechnung der Klassenscores kann man dieselben Verfahren anwenden, wie sie auch sonst für qualitative Variablen gebräuchlich sind, z. B. Häufigkeitsverteilung, Modus, Kontingenztafeln, χ^2-Test und/oder log-lineares Modell.

Vergröberung der latenten Verteilung

Durch die Klassenscores werden die Probanden allerdings in manifeste Klassen eingeteilt, die mit den latenten Klassen nicht völlig identisch sind. Da in der Regel zumindest einige Klassenscores aufgrund von Membership-Wahrscheinlichkeiten vergeben werden, die kleiner als Eins sind, stellen sie eine Vergröberung der latenten Verteilung dar.

Aus diesem Grunde stimmt z. B. auch die Häufigkeitsverteilung der Klassenscores in der Regel nicht exakt mit den Klassengrößen überein.

Kontingenztafeln

Ähnliche Auswirkungen hat dies auch, wenn man die Kontingenzen der Klassenscore-Variable Q mit einer Kriteriumsvariable Y berechnet. Will man tatsächlich den Zusammenhang einer Krite-

riumsvariable mit den *latenten* Klassen untersuchen, dann kann dies nicht aufgrund der bivariaten Häufigkeiten n_{ab} erfolgen, mit welchen der Kriteriumswert $y_v = a$ und der Klassenscore $q_v = b$ zusammentreffen, sondern man muss die Häufigkeiten n_{ab} stattdessen mittels der Formel

$$n_{ab} = \sum_{v:y_v=a} p_{b.x_v} \qquad [2.10.5.3]$$

aus den Membership-Wahrscheinlichkeiten berechnen. Darin erstreckt sich das Summenzeichen über alle Vpn, die in der Kriteriumsvariable den Wert $y_v = a$ besitzen.

Kontrollfragen

- Wie erfolgt die Beurteilung der Testleistung eines Probanden in der Latent-Class-Analyse?
- Erläutern Sie die Anwendung des EM-Algorithmus zur Schätzung der Modellparameter der Latent-Class-Analyse!
- Wie groß darf die Anzahl der unabhängigen Modellparameter der Latent-Class-Analyse maximal sein?
- Wie kann die zur Beschreibung der Daten erforderliche Klassenanzahl bestimmt werden?
- Welche Informationsmaße zum Vergleich verschiedener Latent-Trait- und Latent-Class-Modelle kennen Sie und wie sind diese definiert?
- Wie wirkt sich die Wahl der verschiedenen Informationsmaße auf die Anzahl der latenten Klassen aus, die in der Latent-Class-Analyse als optimale Beschreibung der Daten erachtet wird?
- Was versteht man unter einem lokalen Maximum der Likelihoodfunktion und wie kann man es erkennen?
- Warum kommt es bei der Zuordnung der Vpn zu den latenten Klassen zu einer Vergröberung der latenten Verteilung?

Übungsaufgaben

1. Die Latent-Class-Analyse eines Tests aus drei dichotomen Items ergab in einer Stichprobe von n = 397 Vpn zwei Klassen. Die Klassengrößen und die klassenspezifischen Lösungswahrscheinlichkeiten der Items sind in der nachstehenden Tabelle wiedergegeben:

	p_g	p_{g1}	p_{g2}	p_{g3}
g=1	0.7	0.1	0.1	0.8
g=2	0.3	0.5	0.5	0.5

a) Mit welcher (Membership-)Wahrscheinlichkeit kann ein Proband mit dem Antwortmuster (0,0,1) den beiden Klassen zugeordnet werden

und
b) welche diagnostische Entscheidung würden Sie treffen?

2. Berechnen Sie die Fehlerstreuung der Parameterschätzungen in Aufgabe 1!
3. Berechnen Sie die Anzahl der unabhängigen Modellparameter des LCA-Modells in Aufgabe 1!
4. Berechnen Sie die Anzahl der Freiheitsgrade des LCA-Modells in Aufgabe 1!
5. Die Latent-Class-Analyse eines Tests aus 4 dreistufigen Items ergab die folgenden Goodness-of-Fit-Statistiken:

h	ln{mL(X)}	$n(P_{LCh})$	AIC
1	-5234.641	8	10485,282
2	-5104.846	17	10243.692
3	-5106.956	26	10265.912
4	-5085.653	35	10241.306
5	-5083.132	44	10254.264

Untersuchen Sie, ob es in dieser Tabelle ein lokales Maximum gibt und begründen Sie ihre Antwort!

6. Die Latent-Class-Analyse eines Tests aus 5 binären Items ergab die folgenden Likelihood-Werte:

h	1	2	3	4	5	Sat. Modell
ln{mL(X)}	-4234.471	-4094.864	-4088.456	-4083.132	-4082.653	-4081.467

a) Wie viele Klassen würden Sie aufgrund des AIC-Index annehmen und warum?
b) Berechnen Sie einen Likelihood-Quotienten-Test zur Überprüfung der Annahme, dass die von Ihnen angenommene Klassenzahl die Daten ebenso gut zu beschreiben vermag wie das saturierte Modell!
c) Berechnen Sie einen Likelihood-Quotienten-Test zur Überprüfung der Annahme, dass die von Ihnen angenommene Klassenzahl die Daten besser zu beschreiben vermag als die Ein-Klassen-Lösung!

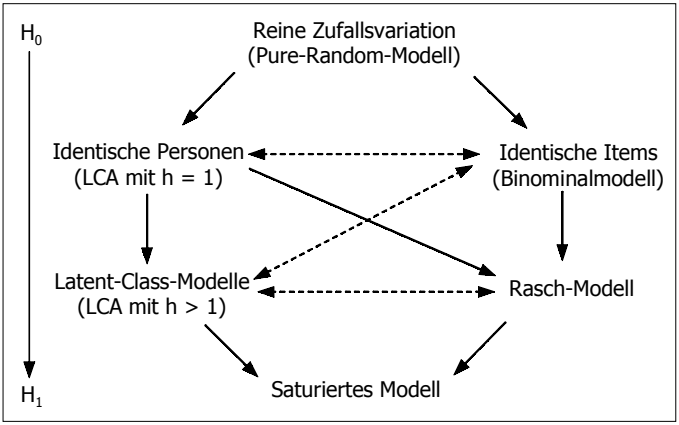
2.11
Modellvergleiche

Fragestellungen

Ganz gleich welches der in den vorangegangenen Kapiteln diskutierten stochastischen Testmodelle wir der Testkonstruktion zugrunde legen, stellt sich stets die Frage nach der Angemessenheit des Modells. Und zwar in dreierlei Hinsicht:

1. Ist das Modell überhaupt geeignet, die Daten zu beschreiben?
2. Wäre ein restriktiveres Modell (mit weniger Parametern) nicht gleichermaßen gut geeignet?
3. Wäre ein anderer Modelltypus (Latent-Trait statt Latent-Class oder umgekehrt) nicht besser geeignet?

Abb. 2.11.1: Systematik der Modellvergleiche



Nullhypothese und Alternativhypothese

Bei der Untersuchung der beiden ersten Fragen geht es um den Vergleich zwischen in der Hierarchie der Modellspezifikationen höher stehenden Modellen mit solchen, die in der Hierarchie der Modellspezifikationen tiefer stehen (und weniger strenge Modellannahmen treffen), wobei das höher stehende Modell jeweils ein Spezialfall des tiefer stehenden Modells ist. In Abbildung 2.11.1 ist diese hierarchische Beziehung mittels durchgezogener Pfeile veranschaulicht.

Beim Vergleich von je zwei solchen Modellen entspricht dem höher stehenden Modell die Null-Hypothese (H_0), dem tiefer stehenden die Alternativhypothese (H_1).

Likelihood-Quotienten-Test

Welches der beiden Modelle zu favorisieren ist, kann mittels Likelihood-Quotienten-Test, anhand der asymptotisch nach χ^2 mit

$$df = n(P_1) - n(P_0) \qquad [2.11.1]$$

verteilten Prüfgröße

$$-2\ln(\lambda) = 2\left|\ln(L_0) - \ln(L_1)\right| \qquad [2.11.2]$$

entschieden werden. Darin bezeichnen $n(P_0)$ und $n(P_1)$ die Anzahl der unter H_0 bzw. unter H_1 zu schätzenden unabhängigen Modellparameter. L_0 und L_1 stehen für die *marginale Likelihood* der Daten unter Annahme von Null- und Alternativhypothese.

Vergleich mit der Ein-Klassen-Lösung der Latent-Class-Analyse

Von unmittelbarer psychodiagnostischer Relevanz ist der Vergleich sowohl des Rasch-Modells als auch der Latent-Class-Analyse (als H_1) mit der Ein-Klassen-Lösung der LCA (als H_0), der darüber entscheidet, ob der Test überhaupt eine latente Personenvariable misst.

Vergleich mit dem saturierten Modell

Da der Likelihood-Quotienten-Test voraussetzt, dass das unter der Alternativhypothese angenommene Modell empirisch gültig ist, ist zuvor in jedem Fall das entsprechende Modell (als H_0) gegen das saturierte Modell (als H_1) zu testen und sicherzustellen, dass das Modell die Daten nicht signifikant schlechter beschreibt als die bestmögliche Beschreibung der Daten, wie sie durch das saturierte Modell gegeben ist.

Informationsmaße

Alternativ zur Berechnung eines Likelihood-Quotienten-Tests kann der Vergleich der Modelle auch mittels der Informationsmaße AIC, BIC und CIC erfolgen, wobei wieder jenem Modell der Vorzug zu geben ist, für welches das entsprechende Informationsmaß den kleinsten Wert annimmt.

Vergleich unterschiedlicher Modelltypen

Anders als der Likelihood-Quotienten-Test ist der Modellvergleich mit Hilfe von Informationsmaßen auch dann zulässig, wenn keines der zu vergleichenden Modelle einen Spezialfall des anderen darstellt. In Abbildung 2.11.1 sind diese Modellvergleiche durch gestrichelte Pfeile repräsentiert.

Von unmittelbarer diagnostischer Relevanz ist der Vergleich zwischen Rasch-Modell und Latent-Class-Analyse, der darüber entscheidet, ob die latente Variable quantitativ (Rasch-Modell) oder qualitativ (LCA) ist.

Reproduzierbarkeit der Patternhäufigkeiten

Auf alle Modellvarianten anwendbar ist auch die auf Aitkin et al. (1981) zurückgehende Modellkontrolle mittels Resimulation der Datenmatrix. Zu diesem Zweck berechnet man die bei Geltung des jeweiligen Modells erwarteten (theoretischen) Häufigkeiten $e(x_v)$ der möglichen Antwortmuster und vergleicht sie mittels der asymptotisch nach χ^2 mit

$$df = \prod_{i=1}^{k} m_i - n(P) - 1 \qquad [2.11.3]$$

Freiheitsgraden verteilten Testgröße

$$\chi^2 = \sum_{x_v} \frac{\left(n(x_v) - e(x_v)\right)^2}{e(x_v)} \qquad [2.11.2.4]$$

mit den tatsächlich beobachteten (empirischen) Häufigkeiten $n(x_v)$. Darin bezeichnet n(P) die Anzahl der unabhängigen Modellparameter und das Summenzeichen in Gleichung [2.11.4] erstreckt sich über alle möglichen Antwortvektoren.

Da bei großem k die Anzahl der möglichen Antwortmuster enorm groß ist, sind die erwarteten Häufigkeiten jedoch in der Regel zu klein, um die Prüfgröße gegen die χ^2-Verteilung zu testen. In der Praxis bedient man sich daher auch hier wieder eines Bootstrap-Verfahrens.

Kontrollfragen

- Welche Fragen stellen sich bezüglich der Angemessenheit stochastischer Testmodelle?
- Stellen Sie die Hierarchie der verschiedenen Latent-Trait- und Latent-Class-Modelle zwischen den beiden Polen Pure-Random- und saturiertes Modell dar und erläutern Sie
  a) welche Modellvergleiche mittels Likelihood-Quotienten-Test durchgeführt werden können,
  b) welches dabei jeweils die Nullhypothese und welches die Alternativhypothese ist sowie
  c) welche Modellvergleiche nicht mittels Likelihood-Quotienten-Test sondern nur mittels Informationsmaßen durchgeführt werden können

und
d) warum dies der Fall ist!

- Welche Art von Likelihood wird den Modellvergleichen zugrunde gelegt?
- Auf welche Modellvarianten ist die Modellkontrolle mittels Resimulation der Datenmatrix anwendbar?

Übungsaufgaben

1. Ein Tests aus 4 binären Items wurde von 146 Vpn bearbeitet. Die psychometrische Analyse des Tests ergab die folgenden marginalen Likelihoodfunktionen:

Modell	ln{mL(X)}
Pure Random	-401.9131
Identische Personen	-396.7988
Identische Items	-384.1581
LCA (h=3 Klassen)	-372.9344
Rasch-Modell	-377.5100
Saturiertes Modell	-371.4418

a) Berechnen Sie die Anzahl der zur Darstellung der Likelihoodfunktionen erforderlichen unabhängigen Modellparameter sowie die Informationsmaße AIC, BIC und CIC!
b) Untersuchen Sie, welche Modelle durch die verschiedenen Informationsmaße favorisiert werden und
c) entscheiden Sie ggf. mittels Likelihood-Quotienten Test, welchem Modell der Vorzug zu geben ist!

2.12 Validität

2.12.1 Augenscheinvalidität, theoretische Validität und Kriteriumsvalidität

Begriffsklärung

Die Untersuchung der Validität von Tests ist ein Grenzgebiet der Psychometrie. Manche Testtheoretiker wie z. B. Fischer (1974, 71) zweifeln an, ob es denn überhaupt Gegenstand der Theorie des Messens psychischer Eigenschaften ist oder ob es nicht eher der angewandten Psychologie zugeordnet werden sollte. Immerhin kann die Psychometrie dafür jedoch eine Reihe von methodischen Hilfsmitteln bereitstellen, deren wichtigste im Folgenden dargestellt werden.

Allgemeiner Konsens ist es, dass es so etwas wie *die* Validität eines Tests gar nicht gibt. Validität ist ein mehrstelliger Prädikator. Valide ist ein Test nie an sich, sondern immer nur in Bezug auf ein bestimmtes Verhaltensmerkmal oder eine latente Eigenschaft der Probanden, die man damit erfassen und/oder prognostizieren will. Z. B. sind Intelligenztests ziemlich valide für die Prognose des Schulerfolgs. Für die Prognose sexueller Vorlieben jedoch weisen sie erwartungsgemäß so gut wie keinerlei Validität auf.

In der testtheoretischen Literatur unterscheidet man zwischen Augenscheinvalidität, theoretischer Validität, Kriteriumsvalidität und Konstruktvalidität.

Augenscheinvalidität

Das Konzept der Augenscheinvalidität (engl. *face validity*) meint nicht mehr als die Plausibilität der Auswahl der Testitems: Wenn man die Intelligenz messen will, stellt man den Probanden Denkaufgaben; will man antisemitische Einstellungen messen, so fragt man die Probanden z. B., ob sie damit einverstanden wären, wenn ihre Tochter einen Juden heiratet, und wenn man Depressivität messen will, fragt man sie nach depressiven Symptomen wie Antriebslosigkeit etc.

Theoretische Validität

Theoretische Validität weist ein Test dagegen dann auf, wenn die Aufgabenkonstruktion bestimmten Regeln folgt, die theoretisch

begründet werden können bzw. die aus einer Theorie des jeweiligen Gegenstandsbereiches abgeleitet sind.

Theoretische Validität in diesem Sinne haben z. B. die von Siegler (1976) konstruierten Testaufgaben zur Untersuchung der kognitiven Entwicklung im Kindesalter (vgl. Bd. 1, Kap. 2.4.7)

Kriteriumsvalidität

So stringent konstruierte Testaufgaben sind in der Testpsychologie aber eher die Ausnahme. Meist begnügt man sich bei der Konstruktion der Testitems mit deren Augenscheinvalidität und untersucht ihre Validität dann empirisch, indem man die Testergebnisse X_{ot} mit einer Kriteriumsvariable $Y_{ot'}$ korreliert.

Veränderung der Validität mit wechselnder Testlänge

Für die so definierte Kriteriumsvalidität ρ_{XY} haben wir bereits in Kapitel 2.2.2 gesehen, dass sie niemals größer werden kann als die Quadratwurzel aus der Reliabilität des Tests. Die Reliabilität ist also eine Voraussetzung für die (Kriteriums-)Validität eines Tests, und unter denselben Voraussetzungen, unter denen man die Veränderung der Reliabilität bei wechselnder Testlänge prognostizieren kann, ist dies auch für die Validität möglich:

$$\rho(X_{ot}, Y_{ot'}) = \rho(X_{oo}, Y_{ot'})\sqrt{\frac{n}{1+(n-1)\rho^2_{XT_o}}}\ , \qquad [2.12.1.1]$$

worin $\rho(X_{oo}, Y_{ot'}) = \rho(X_{o1}, Y_{ot'}) = \ldots = \rho(X_{on}, Y_{ot'})$ die Validität der (zumindest essentiell) parallelen Teiltests bezeichnet, aus welchen der Test zusammengesetzt ist.

Beweis

Aus Gründen der mathematischen Einfachheit leiten wir die Gleichung [2.12.1.1] hier nur für den Spezialfall n = 2 (doppelte Testlänge) her, für den allgemeinen Fall vgl. Lord & Novick (1968).

Zunächst zeigen wir, dass die Kovarianz des zusammengesetzten Tests mit der Kriteriumsvariable gleich

$$\begin{aligned}\sigma(X_{ot}, Y_{ot'}) &= \sigma((X_{o1}+X_{o2}), Y_{ot'}) \\ &= E((X_{o1}+X_{o2})Y_{ot'}) - E(X_{o1}+X_{o2})E(Y_{ot'}) \\ &= E(X_{o1}Y_{ot'}) + E(X_{o2}Y_{ot'}) - E(X_{o1})E(Y_{ot'}) - E(X_{o2})E(Y_{ot'}) \\ &= \sigma(X_{o1}, Y_{ot'}) + \sigma(X_{o2}, Y_{ot'})\end{aligned} \qquad [2.12.1.2]$$

ist. Für die Validität des zusammengesetzten Tests folgt daraus

$$\rho(X_{ot}, Y_{ot'}) = \frac{\sigma(X_{ot}, Y_{ot'})}{\sigma(X_{ot})\,\sigma(Y_{ot'})} \qquad [2.12.1.3]$$

$$= \frac{\sigma(X_{o1}, Y_{ot'}) + \sigma(X_{o2}, Y_{ot'})}{\sigma(X_{ot})\,\sigma(Y_{ot'})}\,.$$

Seien j = 1 und j = 2 essentiell parallele Teiltests, dann ist darin wegen Gleichung [2.2.2.22]

$$\sigma(X_{ot}) = \sqrt{2\sigma^2(X_{oo})(1+\rho^2_{XT_o})} \qquad [2.12.1.4]$$

Ferner korrelieren die essentiell parallelen Teiltests wegen Gleichung [2.2.2.3] gleich hoch mit der Kriteriumsvariable, so dass $\sigma(X_{o1},Y_{ot'}) = \sigma(X_{o2},Y_{ot'}) = \sigma(X_{oo},Y_{ot'})$. Einsetzen in Gleichung [2.12.1.3] ergibt schließlich:

$$\rho(X_{ot}, Y_{ot'}) = \frac{2\,\sigma(X_{oo}, Y_{ot'})}{\sqrt{2\sigma^2(X_{oo})(1+\rho^2_{XT_o})}\;\sigma(Y_{ot'})}$$

$$= \frac{\sqrt{2}\,\sigma(X_{oo}, Y_{ot'})}{\sigma(X_{oo})\sqrt{1+\rho^2_{XT_o}}\;\sigma(Y_{ot'})}$$

$$= \frac{\sigma(X_{oo}, Y_{ot'})}{\sigma(X_{oo})\,\sigma(Y_{ot'})}\;\frac{\sqrt{2}}{\sqrt{1+\rho^2_{XT_o}}}$$

$$= \rho(X_{oo}, Y_{ot'})\sqrt{\frac{2}{1+\rho^2_{XT_o}}}\,.$$

Maximal erzielbare Validität

Anders als die Reliabilität kann die Validität eines Tests durch Testverlängerung jedoch nicht beliebig wachsen, sondern ist durch die Reliabilität der Kriteriumsmessung und die Korrelation zwischen den True-Scores in Test und Kriterium nach oben hin beschränkt. Die Obergrenze der erzielbaren Validität

$$|\rho(X_{ot}, Y_{ot'})| \le |\rho(T_{ot}, T_{ot'})|\,\rho(Y_{ot'}, T_{ot'}) \qquad [2.12.1.5]$$

folgt unmittelbar aus der Gleichung [2.2.2.3]. Der Beweis sei dem Leser als Übungsaufgabe überlassen.

Correction for attenuation

Unmittelbar aus dieser Gleichung ergibt sich auch die sogenannte Verdünnungsformel (engl. *correction for attenuation*), mit der nicht nur in der Testpsychologie eine Menge Missbrauch getrieben wurde und heute noch getrieben wird. Sie beschreibt die *fiktive Validität* des Tests unter der Annahme idealer Reliabilität des Tests und der Kriteriumsmessung.

Der deutsche Name Verdünnungsformel kommt daher, dass die Korrelation zweier Variablen durch die Messfehler verwässert oder eben "verdünnt" wird. Diese Verdünnung wird durch die correction for attenuation fiktiv rückgängig gemacht. Letztlich ist die Verdünnungsformel

$$\rho(T_{ot}, T_{ot'}) = \frac{\rho(X_{ot}, Y_{ot'})}{\sqrt{\rho^2_{XT}\rho^2_{YT'}}} \qquad [2.12.1.6]$$

aber nichts anderes als ein Trick, mit dem man schlechte Korrelationen schönrechnen kann.

Missbrauch der Verdünnungsformel

So hat bereits Spearman (1904) die Annahme einer generellen Intelligenz (vgl. Bd. 1, Kap. 2.4.7) gegen die empirischen Befunde zu "beweisen" versucht, indem er die niedrigen bis mäßigen Korrelationen zwischen verschiedenen Intelligenzleistungen mittels der Verdünnungsformel hochgerechnet hat.

So wird z. B. aus einer Korrelation von 0.38 zwischen Schulintelligenz und Diskriminationsleistungen nach Division durch $\sqrt{0.55 \times 0.25}$ die ideale Korrelation von Eins, was Spearman (1904, 272) zu der euphorischen Feststellung veranlasst hat:

"On the whole, then, we reach the profoundly important conclusion that there really exists a something that we may provisionally term 'General Sensory Discrimination' and similarly a 'General Intelligence', and further that the functional correspondence between these two is not appreciably less than absolute".

Geradezu eine Verletzung der wissenschaftlichen Integrität ist die Anwendung der Verdünnungsformeln, wenn die tatsächliche Korrelation der beiden Variablen und ihre Reliabilitäten nicht dokumentiert werden, sondern in der Publikation empirischer Befunde lediglich angegeben wird, wie hoch zwei Variablen "nach correction for attenuation" miteinander korrelieren.

Andernfalls kann man mittels der Gleichung

$$\rho(X_{ot}, Y_{ot'}) = \rho(T_{ot}, T_{ot'})\sqrt{\rho^2_{XT}\rho^2_{YT'}} \qquad [2.12.1.7]$$

zumindest noch zurückrechnen, wie niedrig die beiden Variablen tatsächlich korrelieren.

Überschätzung der Validität

Da die Methoden der Reliabilitätsbestimmung in der Regel nur eine untere Schranke für die Reliabilität angeben (vgl. Kap.

2.2.2), führt die Verdünnungsformel im allgemeinen zu einer Überschätzung der fiktiven "wahren" Korrelation zwischen den Variablen, was so weit gehen kann, dass man am Ende Korrelationen erhält, die dem Betrage nach größer als Eins sind.

Auch wenn man Spearmans oben zitierte Korrelations- und Reliabilitätskoeffizienten zugrunde legt, erhält man bei genauem Nachrechnen einen Wert größer Eins, was Spearman zum Nachdenken hätte anregen müssen.

2.12.2 Konstruktvalidität

Die Korrelationsmatrix

Die Konstruktvalidität ist gewissermaßen eine Verallgemeinerung der Kriteriumsvalidität, wobei aber nicht nur die Korrelation zwischen einem Test und einer Kriteriumsmessung, sondern die Beziehungen *mehrerer* Tests und/oder Kriteriumsmessungen zueinander untersucht werden.

Allgemein gesprochen haben wir es also mit n Maßzahlvariablen $X_{o1},\ldots,X_{on}$ zu tun, deren paarweise Korrelationen die sog. Korrelationsmatrix bilden, wie sie in Tabelle 2.12.2.1 exemplarisch dargestellt ist.

Tab. 2.12.2.1: Korrelationsmatrix der Subtests Wortauswahl, Zahlenreihen und Würfelaufgaben des IST-70 (nach Kempf & Meder, 1993)

	WA	ZR	WÜ
WA	1	0.304	0.300
ZR	0.304	1	0.278
WÜ	0.300	0.278	1

Wegen der Invarianz der Korrelation gegenüber linearen Transformationen können wir o.B.d.A. voraussetzen, dass die Variablen zu $E(X_{ot}) = 0$ und $\sigma^2(X_{ot}) = 1$ standardisiert sind. Dann ist die Korrelation je zweier Variablen t = 1 und t = 2 gleich ihrem Produktmoment

$$\rho_{12} = \rho(X_{o1}, X_{o2}) = \sigma(X_{o1}, X_{o2}) = E(X_{o1}X_{o2}) \qquad [2.12.2.1]$$

und in der Hauptdiagonale der Korrelationsmatrix stehen die zu 1 normierten zweiten Momente

$$\rho_{tt} = \rho(X_{ot}, X_{ot}) = \sigma(X_{ot}, X_{ot}) = \sigma^2(X_{ot}) = E(X_{ot}^2) = 1 \qquad [2.12.2.2]$$

Das Interpretationsproblem

Nicht nur, weil psychologische Tests in der Regel nur mäßig hoch korrelieren und weil die Korrelationsmatrix bei größerem n sehr

unübersichtlich wird, gibt ihre bloße Betrachtung allerdings kaum Aufschluss über die Struktur der Beziehungen zwischen den Variablen. Da so gut wie jeder Test mit jedem anderen korreliert, können die Korrelationen auch nicht unabhängig voneinander interpretiert werden.

Die geometrische Interpretation des Korrelationskoeffizienten

Einen Ausweg aus diesem Dilemma bietet der Umstand, dass sich die Korrelation je zweier Variablen X_{o1} und X_{o2} als Winkel zwischen zwei Vektoren $a_1 = (a_{11}, a_{12})$ und $a_2 = (a_{21}, a_{22})$ in einem rechtwinkligen Koordinatensystem darstellen lässt. Allgemein gesprochen ist die Korrelation gleich

$$\rho_{12} = h_1 h_2 \cos\varphi_{12} \ , \qquad [2.12.2.3]$$

worin

$$h_t = \sqrt{a_{t1}^2 + a_{t2}^2}$$

die Länge der Vektoren und φ_{12} den Winkel zwischen ihnen bezeichnet (vgl. Abbildung 2.12.2.1). Wegen $\rho_{tt} = 1$, $\varphi_{tt} = 0$ und weil der Kosinus von Null gleich Eins ist, ist aber

$$\rho_{tt} = h_t^2 = 1$$

und die Länge der Vektoren ist zu Eins normiert, so dass

$$\rho_{12} = \cos\varphi_{12} \ . \qquad [2.12.2.4]$$

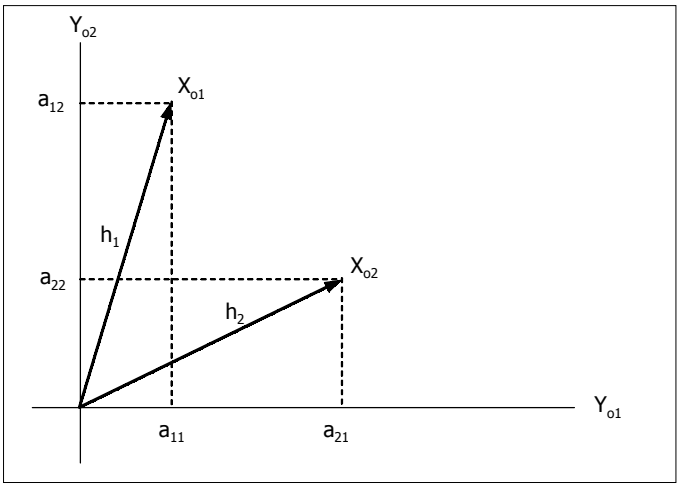


Abb. 2.12.2.1: Geometrische Darstellung der Korrelation zweier Variablen X_{o1} und X_{o2}

Beweis

Zum Beweis stellen wir die beiden Variablen o.B.d.A. in der Form

$$X_{ot} = a_{t1}Y_{o1} + a_{t2}Y_{o2}$$

dar, wobei wir die Gewichtszahlen a_{t1} und a_{t2} so wählen, dass die Hilfsvariablen Y_{o1} und Y_{o2} unkorreliert und zu $E(Y_{oj}) = 0$ und $\sigma^2(Y_{oj}) = 1$ standardisiert sind.

Ist dies der Fall, so nimmt die Korrelation zwischen den beiden Tests die Form

$$\rho_{12} = E(X_{o1}X_{o2})$$
$$= E\{(a_{11}Y_{o1} + a_{12}Y_{o2})(a_{21}Y_{o1} + a_{22}Y_{o2})\}$$
$$= E(a_{11}a_{21}Y_{o1}^2 + a_{11}a_{22}Y_{o1}Y_{o2} + a_{12}a_{21}Y_{o2}Y_{o1} + a_{12}a_{22}Y_{o2}^2)$$
$$= a_{11}a_{21} + a_{12}a_{22}$$

an. Wie wir zeigen, kann auch der Ausdruck auf der rechten Seite der Gleichung [2.12.2.3] in der selben Form

$$h_1h_2\cos\varphi_{12} = a_{11}a_{21} + a_{12}a_{22}$$

dargestellt werden. Der Winkel φ_{12} zwischen den beiden Vektoren ist nämlich gleich der Differenz der Winkel α (zwischen der Abszisse und X_{o1}) und β (zwischen der Abszisse und X_{o2}), so dass sein Kosinus wegen des Additionstheorems der Winkelfunktionen gleich

$$\cos\varphi_{12} = \cos(\alpha - \beta)$$

$$= \cos\alpha\cos\beta + \sin\alpha\sin\beta$$

ist. Einsetzen von

$$\cos\alpha = \frac{a_{11}}{h_1}, \qquad \sin\alpha = \frac{a_{12}}{h_1}$$

$$\cos\beta = \frac{a_{21}}{h_2} \quad \text{und} \quad \sin\alpha = \frac{a_{22}}{h_2}$$

ergibt schließlich

$$\cos\varphi = \frac{a_{11}}{h_1}\frac{a_{21}}{h_2} + \frac{a_{12}}{h_1}\frac{a_{22}}{h_2}$$

und nach Multiplikation der Gleichung mit h_1h_2 erhalten wir

$$h_1h_2\cos\varphi_{12} = a_{11}a_{21} + a_{12}a_{22}.$$

Schon um die paarweisen Korrelationen von nur drei Variablen darzustellen, reicht ein zweidimensionales Koordinatensystem nur dann aus, wenn sich die Winkel zu $\varphi_{12} + \varphi_{23} = \varphi_{13}$ summieren. In der Regel ist die Summe aus $\varphi_{12} + \varphi_{23}$ jedoch größer als φ_{13}, so dass wir in die dritte Dimension ausweichen müssen (vgl. Abb. 2.12.2.2).

Die geometrische Darstellung der paarweisen Korrelationen von mehr als zwei Variablen

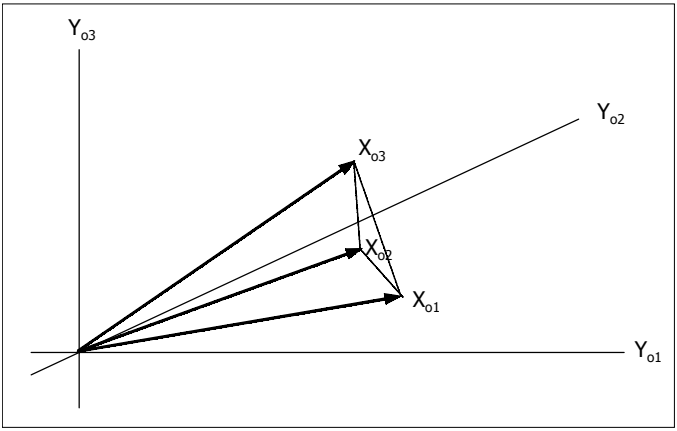


Abb. 2.12.2.2: Darstellung der Korrelation dreier Variablen X_{o1}, X_{o2} und X_{o3} im dreidimesionalen Raum.

In der Regel wird man somit zur Darstellung der Interkorrelationen von n Variablen ebenso viele Dimensionen benötigen.

Dies macht sich die auf Hotelling (1933) zurückgehende Hauptkomponentenanalyse zu Nutze, welche die n Variablen $X_{o1},\ldots,X_{on}$ mittels der Grundgleichungen

Die Hauptkomponentenanalyse

$$X_{ot} = a_{t1}Y_{o1} + \ldots + a_{tj}Y_{oj} + \ldots + a_{tq}Y_{tq} \qquad [2.12.2.5]$$

als Linearkombination von $q \leq n$ standardisierten Hilfsvariablen $Y_{o1},\ldots,Y_{oq}$ darstellt. Die Hilfsvariablen sind unkorreliert, so dass sie die Achsen eines orthogonalen (rechtwinkligen) Koordinatensystems bilden, und werden Hauptkomponenten (engl. *principal components*) genannt. Die Gewichtszahlen a_{tj} beschreiben die Korrelationen der Variablen $X_{o1},\ldots,X_{on}$ mit den Hauptkomponenten $Y_{o1},\ldots,Y_{oq}$ und werden als die Ladungen der Variablen auf den Hauptkomponenten bezeichnet:

$$a_{tj} = \rho\left(X_{ot}, Y_{oj}\right). \qquad [2.12.2.6]$$

Beweis

Wegen der Standardisierung der Variablen und Hauptkomponenten ist

$$\rho(X_{ot}, Y_{oj}) = E(X_{ot} Y_{oj})$$

und durch Einsetzen der Grundgleichungen [2.12.2.5] für X_{ot} erhalten wir

$$\rho(X_{ot}, Y_{oj}) = E\left\{\left(\sum_{i=1}^{q} a_{ti} Y_{oi}\right) Y_{oj}\right\}$$

$$= \sum_{i=1}^{q} E(a_{ti} Y_{oi} Y_{oj})$$

$$= \sum_{i=1}^{q} a_{ti} E(Y_{oi} Y_{oj}).$$

Wegen der Orthogonalität der Hauptkomponenten sind die Produktmomente $E(Y_{oi}Y_{oj})$ aber für $i \neq j$ alle gleich Null, so dass

$$\rho(X_{ot}, Y_{oj}) = a_{tj} E(Y_{oj}^2)$$

Wegen der Standardisierung der Hauptkomponenten sind ihre zweiten Momente $E(Y_{oj}^2)$ alle gleich Eins und es folgt

$$\rho(X_{ot}, Y_{oj}) = a_{tj}.$$

Die Matrix $A = ((a_{tj}))$ heißt die (vollständige) Ladungsmatrix.

Die Berechnung der Ladungsmatrix

Seien $X = (X_{o1},\ldots,X_{on})$ und $Y = (Y_{o1},\ldots,Y_{oq})$, dann können die Grundgleichungen [2.12.2.5] unter Verwendung der Matrix- und Vektorschreibweise zu

$$X' = AY' \qquad [2.12.2.7]$$

zusammengefasst und die Ladungsmatrix kann durch Lösung des Gleichungssystems

$$Q = AA' \qquad [2.12.2.8]$$

gefunden werden, worin Q die Korrelations- bzw. Kovarianzmatrix der Variablen $X_{o1},\ldots,X_{on}$ bezeichnet[1] und sich die Anzahl der Dimensionen, welche zur Darstellung der Korrelationsmatrix erforderlich sind, gemäß

1. Auf die technischen Details der Lösung des Gleichungssystems braucht hier nicht näher eingegangen zu werden. Vgl. hierzu aber z. B. Kähler (2002).

$$q = \text{Rang}(Q) \leq n \ , \qquad [2.12.2.9]$$

aus dem Rang der Korrelationsmatrix ergibt.

Der tautologische Charakter der Hauptkomponentenanalyse

Anders als die Faktorenanalyse, die die Interkorrelationsmatrix durch q latente Faktoren *erklären* will, ist die Hauptkomponentenanalyse ein rein tautologisches Verfahren. Das Gleichungssystem [2.12.2.8] besitzt stets eine Lösung und die Hauptkomponenten $Y_{o1},\ldots,Y_{oq}$ sind keine latenten Variablen, sondern durch

$$F' = A^{-1}X' \qquad [2.12.2.10]$$

definierte Linearkombinationen der manifesten Variablen $X_{o1},\ldots,X_{on}$.

Das Eindeutigkeitsproblem

Das Gleichungssystem [2.12.2.8] hat aber nicht nur eine Lösung, sondern es gibt unendlich viele Möglichkeiten, wie der q-dimensionale Vektorraum durch rechtwinklige Koordinaten aufgespannt werden kann. Die Ladungsmatrix ist daher nur bis auf orthogonale lineare Transformationen der Form

$$B = AT \qquad [2.12.2.11]$$

bestimmt, wobei es genau eine Transformationsmatrix T gibt, welche je zwei gleichwertige Ladungsmatrizen A und B ineinander überführt. Dieses ist die Transformationsmatrix

$$T = \left(A'A\right)^{-1}A'B \ \cdot \qquad [2.12.2.12]$$

Das Rotationsproblem

Für welche der unendlich vielen Rotationen des rechtwinkligen Koordinatensystems man sich entscheidet, ist eine pragmatische Frage, die darauf hinausläuft, bei welcher Rotation die Struktur der Korrelationsmatrix am besten erkennbar wird.

Bereits die zur Lösung des Gleichungssystems [2.12.2.8] gebräuchliche Hauptachsenmethode berechnet nicht irgendeine Ladungsmatrix, sondern legt das Koordinatensystem so fest, dass die erste Hauptkomponente einen möglichst großen Anteil der Gesamtvarianz der Variablen beschreibt. Die zweite Hauptkomponente wird dann so festgelegt, dass sie einen möglichst großen Anteil der verbleibenden Restvarianz beschreibt, usw. Die Varianzanteile der Hauptkomponenten berechnen sich nach

$$\text{Varianzanteil} = \frac{1}{n}\lambda_j, \qquad [2.12.2.13]$$

worin

$$\lambda_j = \sum_{t=1}^{n} a_{tj}^2 \qquad [2.12.2.14]$$

den zum j-ten Eigenvektor der Korrelationsmatrix gehörenden Eigenwert darstellt.

Interpretation der Analyseergebnisse

In unserem Zahlenbeispiel oben erhalten wir derart die Ladungsmatrix in Tabelle 2.12.2.2, wonach die erste Hauptkomponente etwas mehr als die Hälfte und die anderen beiden jeweils etwas weniger als ein Viertel der Gesamtvarianz beschreiben.

Tab. 2.12.2.2: Hauptkomponentenanalyse der Subtests Wortauswahl, Zahlenreihen und Würfelaufgaben des IST-70

	Y_{o1}	Y_{o2}	Y_{o3}
WA	0,73	0,33	-0,59
ZR	0,72	0,36	+0,59
WÜ	0,72	-0,69	0,00
Varianzanteil	0.52	0.24	0.23

Im Unterschied zu den voneinander abhängigen Interkorrelationen der Tests sind die Ladungen (= die Korrelationen der Tests mit den Achsen des Koordinatensystems) nun unabhängig voneinander interpretierbar.

- Wie Abbildung 2.12.2.3 veranschaulicht, handelt es sich bei der ersten Hauptkomponente um eine unidirektionale Dimension, die etwas beschreibt, das allen drei Tests gemeinsam ist und das vielleicht als "allgemeine Intelligenz" umschrieben werden könnte.
- Die zweite Hauptkomponente ist bidirektional. Sie differenziert zwischen dem Wortauswahl- und dem Zahlenreihentest einerseits und den Würfelaufgaben andererseits. Entsprechend könnte sie etwa auf die Polarität von analytischem Denken und (räumlicher) Vorstellung verweisen.
- Die dritte Hauptkomponente ist ebenfalls bidirektional. Während die Würfelaufgaben auf dieser Dimension überhaupt nicht laden, differenziert sie jedoch sehr deutlich zwischen dem Wortauswahl- und dem Zahlenreihentest. Entsprechend könn-

te sie etwa die Polarität von formal-analytischem (ZR) und material-analytischem (WA) Denken zum Ausdruck bringen.

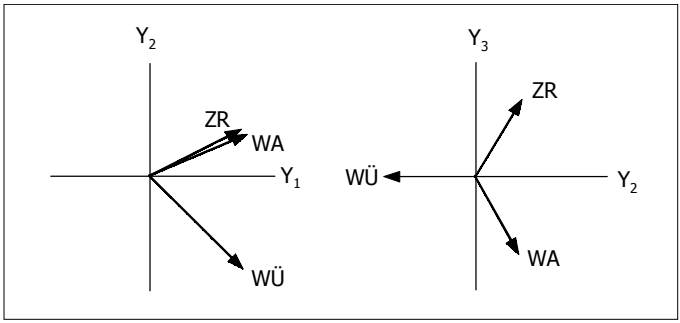


Abb. 2.12.2.3: Geometrische Darstellung der Analyseergebnisse

Schiefwinklige Rotation

Grundsätzlich kann ein q-dimensionaler Raum nicht nur durch ein rechtwinkliges, sondern auch durch ein schiefwinkliges Koordinatensystem aufgespannt werden. Folglich kann man die Ladungsmatrix auch schiefwinklig rotieren und in der Tat liefert jede Multiplikation der Ladungsmatrix mit einer beliebigen regulären Transformationsmatrix T wieder eine Lösung des Gleichungssystems [2.12.2.8]. Durch eine schiefwinklige Rotation erhält man aber miteinander korrelierende Hauptkomponenten, so dass sich das oben angesprochene Interpretationsproblem erneut einstellt und der praktische Nutzen der Hauptkomponentenanalyse verloren geht.

Unvollständige Ladungsmatrizen

Wenn die Anzahl der analysierten Variablen groß ist, kann es vorkommen, dass bereits durch relativ wenige Hauptkomponenten ein relativ großer Varianzanteil beschrieben wird und die weiteren Komponenten nur schwer zu interpretieren sind. Man beschränkt sich in der Praxis daher oft auf die ersten $m < q$ Hauptkomponenten und gibt sich z. B. damit zufrieden, dass die berücksichtigten Komponenten nur ca. 80% der Gesamtvarianz der Testergebnisse beschreiben.

Unter pragmatischem Aspekt mag dies durchaus sinnvoll sein. Man muss sich aber darüber im Klaren sein, dass eine unvollständige Ladungsmatrix, die nur die ersten m Komponenten enthält, die Korrelationsmatrix nur deshalb *sparsamer* beschreibt als die vollständige Ladungsmatrix, *weil* sie die Korrelationsmatrix *unvollständig* beschreibt.

Hauptkomponentenanalyse und Faktorenanalyse

Die formale Ähnlichkeit zwischen Hauptkomponentenanalyse und Faktorenanalyse sowie der Umstand, dass man sich in der Hauptkomponentenanalyse oft mit unvollständigen Ladungsmatrizen zufrieden gibt, haben immer wieder zu einer Verwechslung der beiden Verfahren geführt. Namentlich in der neueren Literatur wird die Hauptkomponentenanalyse gerne als explorative Faktorenanalyse bezeichnet, was jedoch auf einem grundlegenden (wenngleich kaum auszurottenden) Missverständnis beruht.

- Während die von Thurstone (1931, 1947) begründete Faktorenanalyse eine Erklärung von n manifesten Variablen durch eine geringere Anzahl von q < n latenten Variablen (= Faktoren) anstrebt,
- handelt es sich bei den Hauptkomponenten ganz einfach um (in der Regel) q = n Linearkombinationen der manifesten Variablen, die weder latent sind noch irgendeinen Erklärungswert besitzen, sondern lediglich die Interpretation der Korrelationsmatrix erleichtern.

Schon dies macht die Hauptkomponentenanalyse zu einer Methode von großem heuristischem Wert.

Das Modell der Faktorenanalyse

Ähnlich wie der Hauptkomponentenanalyse liegt auch der Faktorenanalyse ein lineares Gleichungssystem der Form

$$X_{ot} = a_{t1}F_{o1} + \ldots + a_{tj}F_{oj} + \ldots + a_{tm}F_{tm} \qquad [2.12.2.15]$$

zugrunde. Anders als bei den Hauptkomponenten Y_{oj} handelt es sich bei den Faktoren F_{oj} jedoch um *latente* Variable, welche die manifesten Variablen *erklären* sollen. Außerdem unterscheiden sich Hauptkomponenten- und Faktorenanalyse auch dahingehend, dass die Anzahl der latenten Variablen m die Anzahl der manifesten Variablen übersteigt.

Gemeinsame und spezifische Faktoren

Innerhalb der latenten Variablen wird zwischen gemeinsamen und spezifischen Faktoren unterschieden. Geht man – ähnlich wie in der klassischen Testtheorie – davon aus, dass die Messung der manifesten Variablen fehlerbehaftet ist, so haben wir es in der Regel mit n spezifischen Faktoren zu tun, die neben Messfehlern auch systematische Einflussgrößen beinhalten können und daher als Einzigartigkeitsfaktoren zu interpretieren sind. Insgesamt hat man es in der Faktorenanalyse daher mit

$$m = q + n \qquad [2.12.2.16]$$

Faktoren zu tun, von denen man sich aber nur für die $q < n$ gemeinsamen Faktoren weiter interessiert.

Die reduzierte Kovarianzmatrix

Setzt man voraus, dass die Faktoren nicht miteinander korrelieren, dann kann die reduzierte Ladungsmatrix $\overline{A}$, welche die Ladungen auf den q gemeinsamen Faktoren enthält, ähnlich wie in der Hauptkomponentenanalyse wieder durch die Lösung eines Gleichungssystems der Form

$$\overline{Q} = \overline{A}\,\overline{A}' \qquad [2.12.2.17]$$

gefunden werden. Darin ist aber $\overline{Q}$ nicht die oben dargestellte Kovarianzmatrix der manifesten Variablen, sondern die reduzierte Kovarianzmatrix, die man erhält, wenn man die Hauptdiagonalelemente von Q durch die Kommunalitäten

$$h_t^2 = \sum_{j=1}^{q} a_{tj}^2 \qquad [2.12.2.18]$$

ersetzt, die den Varianzanteil der Variablen beschreiben, welcher durch die gemeinsamen Faktoren erklärt wird.

Das Kommunalitätenproblem

Die Kommunalitäten sind zu Beginn der Analyse noch nicht bekannt. Deshalb behilft man sich mit einer Schätzung der Kommunalitäten und verbessert diese Schätzungen iterativ. Ein solches Vorgehen ist nur dann statthaft, wenn das Verfahren gegen die richtigen Werte konvergiert. Dies ist jedoch nur möglich, wenn die Kommunalitäten – wenigstens theoretisch – durch die Kovarianzen eindeutig bestimmt sind. Wie aus Arbeiten von Kalveram (1970) und Kempf (1972b) hervorgeht, ist dies aber nicht der Fall. Zu einer Kovarianzmatrix mit ausgesparter Hauptdiagonale kann man Ladungsmatrizen finden, die zwar die Kovarianzen richtig reproduzieren, aber zu unterschiedlichen Kommunalitäten führen. Es gibt keine reguläre lineare Transformation, welche solche Ladungsmatrizen ineinander überführt, und im allgemeinen Faktorenmodell mit $q \leq n$ gemeinsamen und n spezifischen Faktoren ist daher weder die Anzahl der gemeinsamen Faktoren eindeutig bestimmt noch die Auffindbarkeit einer vorgegebenen Ladungsmatrix sichergestellt.

Konfirmatorische Faktorenanalyse

In der Praxis der Faktorenanalyse macht man aus der Not eine Tugend: Wenn die Anzahl der gemeinsamen Faktoren schon

nicht durch die Daten bestimmt ist, kann man sie so festsetzen, dass sie den jeweiligen theoretischen Vorstellungen entspricht.

Voraussetzung dafür ist, dass man bereits *a priori* über Hypothesen bezüglich (a) der Anzahl der zugrunde liegenden Faktoren, (b) bezüglich der Beziehungen der Variablen zu den Faktoren sowie (c) bezüglich der Beziehungen der Faktoren untereinander verfügt.

Mittels der auf Jöreskog (1969) zurückgehenden konfirmatorischen Faktorenanalyse kann man dann überprüfen, ob eine hinreichend gute Übereinstimmung der empirischen Daten mit dem theoretischen Modell besteht oder ob das Modell verworfen werden muss. Unter der Zusatzannahme, dass die Variablen multivariat normalverteilt sind, können die Modellparameter mittels Maximum-Likelihood-Methode geschätzt und die Modellgeltung mittels Likelihood-Quotienten-Test überprüft werden.

Kontrollfragen

- Worin besteht der Unterschied zwischen Augenscheinvalidität, theoretischer Validität und Kriteriumsvalidität?
- Warum führt die Verdünnungsformel in der Regel zu einer Überschätzung der fiktiven "wahren" Korrelation der Variablen?
- Worin bestehen Gemeinsamkeiten und Unterschiede zwischen Kriteriumsvalidität und Konstruktvalidität?
- Warum gibt die bloße Betrachtung der Korrelationsmatrix kaum Aufschluss über die Struktur der Beziehungen zwischen den Variablen?
- Wie kann die Korrelation zweier Variablen geometrisch interpretiert werden?
- Wie viele Hauptkomponenten sind zur Darstellung der Korrelationen zwischen n Variablen normalerweise erforderlich?
- Was beschreiben die Ladungen der Variablen auf den Hauptkomponenten?
- Wodurch wird die Anzahl der Dimensionen bestimmt, die zur Darstellung der Korrelationsmatrix erforderlich sind?
- Warum geht der praktische Nutzen der Hauptkomponentenanalyse verloren, wenn man das Koordinatensystem schiefwinkelig rotiert?
- Worin bestehen Gemeinsamkeiten und Unterschiede zwischen Hauptkomponentenanalyse und Faktorenanalyse?
- Was beschreiben in der Faktorenanalyse die sog. Kommunalitäten?
- Woran scheitert die iterative Verbesserung der Kommunalitätenschätzung?

Übungsaufgaben

1. Ein Eignungstest korreliert mit dem Ausbildungserfolg in der Programmiererausbildung zu 0.68. Die Reliabilität des Eignungstests wird mit 0.95 angegeben. Mit welcher verbleibenden Validität ist zu rechnen, wenn man an Stelle des Tests eine Kurzform verwendet, die nur 30% der ursprünglichen Testitems umfasst?
2. Zeigen Sie, dass die Ungleichung [2.12.1.5] stets erfüllt ist!
3. Leiten Sie die Verdünnungsformel [2.12.1.6] her!
4. Spearman berichtet eine Korrelation von 0.38 zwischen Schulintelligenz und Diskriminationsleistungen. Die Reliabilität der zugrundeliegenden Tests wird mit 0.55 für die Schulintelligenz und 0.25 für die Diskriminationsleistungen angegeben. Berechnen Sie mittels der Verdünnungsformel die fiktive Korrelation zwischen den True-Score-Variablen der beiden Tests und diskutieren Sie das Ergebnis!
5. Leiten Sie die Gleichung [2.12.1.7] her!
6. Eine Studie berichtet über eine fiktive "wahre" Korrelation zwischen Studienmotivation und dem Studienerfolg von 0.7. Die Reliabilität der zugrunde liegenden Tests wird mit 0.6 für die Studienmotivation und 0.5 für den Studienerfolg angegeben. Wie hoch korrelieren die beiden Tests tatsächlich (empirisch) miteinander?

Literatur

Aitkin, M., Andersen, D. & Hinde, J. (1981). Statistical modelling of data on teaching styles. Journal of the Royal Statistical Society A, 144, 419-461.

Akaike, H. (1987). Factor Analysis and AIC. Psychometrika, 52, 317-332.

Amthauer, R. (1973). Intelligenz-Struktur-Test, IST-70. 4. Auflage. Göttingen: Hogrefe

Andersen, E.B. (1973). Conditional inference and models for measuring. Copenhagen: Mentalhygiejnisk Forlag.

Andersen, E.B. (1982). Latent structure analysis – A survey. Scandinavian Journal of Statistics, 9, 1-12.

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 2, 581-594.

Andrich, D. (1978b). A rating formulation for ordered response categories. Psychometrika, 43, 561-573.

Appelsmeyer, H., Kochinka, A. & Straub, J. (1997). Qualitative Methoden. In: Straub, J., Kempf, W. & Werbik, H. (eds.), Psychologie. Eine Einführung. Grundlagen, Methoden, Perspektiven. München: Deutscher Taschenbuchverlag, 709-734.

Austrian Study Center for Peace and Conflict Resolution – ASPR (ed.) (2003). Constructive conflict coverage. A social psychological approach. Berlin: verlag irena regener.

Baros, W. (2001). Familien in der Migration. Frankfurt/Main: Peter Lang.

Baros, W. & Reetz, K.-D. (2002). Sozialpsychologische Rekonstruktion und empirische Migrationsforschung. Conflict & communication online, 1/2.

Berelson, B. (1952). Content analysis in communications research. Glencoe, Ill: The Free Press.

Bergmann, J.R. (2000). Konversationsanalyse. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 524-537.

Bergmann, J.R. (2000b). Ethnomethodologie. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 118-135.

Binet, A. & Simon, Th. (1905). Methodes nouvelles pour le diagnostique du niveau intellectuel des anormaux. Année Psychologique, 11, 191-244.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M. & Novick, M.R. (1968), Statistical theories of mental test scores. Reading (MA): Addison-Wesley. Part 5, 397-479.

Bläsi, B., Jaeger, S., Kempf, W., Kondopoulou, M. & Paskoski, D. (2005). Konstruktive Aspekte des serbischen, deutschen und griechischen Nachkriegsdiskurses – qualitative Vergleichsstudien. In: Projektgruppe Friedensforschung Konstanz (ed.), Nachrichtenmedien als Mediatoren von Peace-Building, Demokratisierung und Versöhnung in Nachkriegsgesellschaften. Berlin: verlag irena regener, 149-200.

Bläsi, B., Jaeger, S., Kempf, W. & Möckel, J. (2004). A catalog of escalation- and de-escalation-oriented aspects of conflict coverage. Diskussionsbeiträge der Projektgruppe Friedensforschung Konstanz, Nr. 53. Berlin: verlag irena regener.

Blaikie, N.W. (1991). A critique of the use of triangulation in social research. Quality and Quantity, 25. 115-136.

Bloor, M. (1997). Techniques of validation in qualitative research: A critical commentary. In: Miller, G., & Dingwall, R. (eds.), Context and method in qualitative research. London: Thousand Oaks, 37-50.

Bogdan, R. & Taylor, St. J. (1975). Introduction to qualitative research methods. A phenomenological approach to the social sciences. New York: Wiley.

Bohnsack, R. (2000). Gruppendiskussion. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 360-368.

Bortz, J. (1993). Statistik für Sozialwissenschaftler. 4. Auflage. Berlin: Springer.

Bos, W. & Tarnai, C. (eds.) (1989). Angewandte Inhaltsanalyse in empirischer Pädagogik und Psychologie. Münster: Waxmann.

Bozdogan, H. (1987). Model selection for Akaike's information criterion (AIC). Psychometrika, 53, 345-370.

Bühner, M. (2006). Einführung in die Test- und Fragebogenkonstruktion. 2. Auflage. München: Pearson.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational psychological measurement, 20, 37-46.

Colonius, H. (1977). On Keat's generalization of the Rasch-Model. Psychometrika, 42, 443-445.

Cook, G. (1990). Transcribing infinity – Problems of context presentation. Journal of Pragmatics, 14, 1-24.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Dechmann, M. (1978). Teilnahme und Beobachtung als soziologisches Basisverhalten. Bern: Haupt.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM-algorithm. Journal of the Royal Statistical Society, 39, 1-22.

Denzin, N.K. (1978). The research act. A theoretical introduction to sociological methods. New York: Mc Graw Hill.

Denzin, N.K. (1989). The research act. A theoretical introduction to sociological methods. 3. Auflage. Englewood Cliffs, N.J.: Prentice Hall.

Denzin, N.K. (2000). Symbolischer Interaktionismus. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 136-150.

Denzin, N.K. & Lincoln, Y.S. (eds.) (1994). Handbook of qualitative research. London: Sage.

Deutsch, M. (1973). The resolution of conflict. New Haven: Yale University Press.

Dijk, T. A. van (ed.) (1985). Discourse and Communication: New Approches to the Analysis of Mass Media Discourse and Communication. Berlin; New York: de Gruyter.

Eckes, Th. (2006). Multifacetten-Rasch-Analyse von Personenbeurteilungen. Zeitschrift für Sozialpsychologie, 37, 185-195.

Efron, B. & Tishirani, R.J. (1993). An introduction to the bootstrap. New York: Chapman & Hall.

Ehlich, K. & Switalla, B. (1976). Transkriptionssysteme. Eine exemplarische Übersicht. Studium Linguistik, 2, 78-105.

Fischer, G.H.. (1972). A step towards dynamic test-theory. Psychologisches Institut der Universität Wien. Abteilung für Methodik: Research Bulletin No. 10, February 1972.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. Acta Psychologica, 37, 359-374.

Fischer, G. H. (1974). Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen. Bern: Huber.

Fischer, G.H., (1983). Logistic latent trait models with linear constraints. Psychometrika, 48, 3-26.

Fischer, G.H. & Parzer, P. (1991a). An extension of the rating scale model with an application to the measurement of treatment effects. Psychometrika, 56, 637-651.

Fischer, G.H. & Parzer, P. (1991b). LRSM: Parameter estimation for the linear rating scale model. Applied Psychological Measurement, 15, 138.

Fischer, G.H. & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. Psychometrika, 59, 177-192.

Fisher, R. & Ury, W. (1984). Das Harvard-Konzept. Sachgerecht verhandeln – erfolgreich verhandeln. Frankfurt/Main: Campus.

Fisz, M. (1970). Wahrscheinlichkeitsrechnung und mathematische Statistik. Berlin: VEB Deutscher Verlag der Wissenschaften.

Flick, U. (1995). Qualitative Forschung – Theorien, Methoden, Anwendung in Psychologie und Sozialwissenschaften. 4. Auflage. Reinbek: Rowohlt.

Flick, U., Kardoff, E. v. & Steinke, I. (2000). Was ist qualitative Forschung? Einleitung und Überblick. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 13-29.

Flick, U. (2000a). Konstruktivismus. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 150-164.

Flick, U. (2000b). Triangulation in der qualitativen Forschung. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 309- 318.

Formann, A.K. (1984). Die Latent-Class-Analyse. Weinheim: Beltz.

Freire, P. (1981). Der Lehrer ist Politiker und Künstler. Neue Texte zur befreienden Bildungsarbeit. Reinbek: Rowohlt.

Friebertshäuser, B. (1997). Interviewtechniken – ein Überblick. In: Friebertshäuser, B. & Prengel, A. (eds.), Handbuch qualitative Forschungsmethoden in der Erziehungswissenschaft. Weinheim: Juventa, 371-395.

Friedrich, J. (1973). Methoden empirischer Sozialforschung. Reinbek: Rowohlt.

Früh, W. (2001). Inhaltsanalyse. Theorie und Praxis. 5., überarbeitete Auflage. Konstanz: UVK.

Girtler, R. (1984). Methoden der qualitativen Sozialforschung. Wien: Böhlau.

Glaser, B.G. (1978). Theoretical sensitivity: Advances in the methodology of grounded theory. Mill Valley, Ca: Sociology Press.

Glaser, B.G. & Strauss, A.L. (1967). The discovery of grounded theory. Strategies for qualitative research. Chicago: Aldine.

Glaser, B.G. & Strauss, A.L. (1979). Die Entdeckung gegenstandsbezogener Theorie: Eine Grundstrategie qualitativer Sozialforschung. In: Hopf, C. & Weingarten, E. (eds.), Qualitative Sozialforschung. Stuttgart: Klett-Cotta, 91-111.

Glasl, F. (1992). Konfliktmanagement. Bern: Haupt.

Goodman, L.A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable, Part I. A modified latent structure approach. American Journal of Sociology, 79, 1179-1259.

Goodman, L.A. (1979). On the estimation of parameters in latent sructure analysis. Psychometrika, 44, 123-128.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika, 10, 255-282.

Guttman, L. (1950). The basis of scalogram analysis. In: Stouffer, S.A., Guttman, L., Suchman, E.A., Lazarsfeld, P.F., Star, S.A. & Clausen, J.A. (eds.), Studies in social psychology in world war II, Vol. IV. Measurement and Prediction. Princeton, N.J.: Princeton Universtity Press, 60-90.

Hall, P. & Heyde, C.C. (1980). Martingale limit theory and its application. New York: Academic Press.

Hamdorf, D. (2001). Zweidimensionales Bild der Wirklichkeit. Das Nordirland Friedensabkommen in der Berliner Zeitung als ein Beispiel für Friedensjournalismus? In: Richter, J. (ed.), Deutschland: (un-)bewältigte Vergangenheiten. Tübingen: dgvt-Verlag, 141-150.

Heinze, T. & Klusemann, H.W. (1979). Ein biographisches Interview als Zugang zu einer Bildungsgeschichte. In: Baacke, D. & Schulz, T. (eds.), Aus Geschichten lernen. München: Juventa, 195-201.

Heinze, T. & Klusemann, H.W. (1980). Versuch einer sozialwissenschaftlichen Paraphrasierung am Beispiel des Ausschnittes aus einer Bildungsgeschichte. In: Heinze, T., Klusemann, H.W. & Soeffner, H.G. (eds.), Interpretationen einer Bildungsgeschichte. Überlegungen zur sozialwissenschaftlichen Hermeneutik. Bensheim: päd.-extra Buchverlag, 97-152.

Hitzler, R. & Eberle, T.S. (2000). Phänomenologische Lebensweltanalyse. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 109-118.

Hoffmann-Riem, Ch. (1980). Die Sozialforschung einer interpretativen Soziologie. Der Datengewinn. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 32, 339-372.

Holsti, O.R. (1969). Content analysis for the social sciences and humanities. Reading, Mass: Addison-Wesley.

Holzkamp, K. (1972). Wissenschaftstheoretische Voraussetzungen kritisch-emanzipatorischer Psychologie. In: Holzkamp, K. (ed.), Kritische Psychologie. Vorbereitende Arbeiten. Frankfurt/Main: Fischer, 75-146.

Honer, A. (1994). Einige Probleme lebensweltlicher Ethnographie. Zur Methodologie und Methodik einer interpretativen Sozialforschung. In: Schröer, N. (ed.), Interpretative Sozialforschung. Auf dem Wege zu einer hermeneutischen Wissenssoziologie. Opladen: Westdeutscher Verlag, 90-98.

Hopf, C. (2000). Qualitative Interviews – ein Überblick. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 349-360.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24, 417-441, 498-520.

Immich, H., Eckmann, F., Neumann, H., Schärperle, O., Schwarz, H. & Tempel, H. (1971). Grundlegende Probleme bei der Antidepressiva-Prüfung. Arzneimittelforschung (Drug Res), 21, 525-528, 644-649.

Jaeger, S. (2005). Von Nachhall und Überwindung deutsch-französischer Gegnerschaft – qualitative Fallstudien. In: Projektgruppe Friedensforschung Konstanz (ed.), Nachrichtenmedien als Mediatoren von Peace-Building, Demokratisierung und Versöhnung in Nachkriegsgesellschaften. Berlin: verlag irena regener, 81-126.

Jaeger, S., Mattenschlager, A. & Meder, G. (1999). Dokumentation der Datenbasis der Bosnien-Studie im "Journalism in the New World Order"-Projekt. Diskussionsbeiträge der Projektgruppe Friedensforschung Konstanz, Nr. 47. Konstanz: KOPS.

Jäger, S. (1994). Kritische Diskursanayse. Eine Einführung. Duisburg: Duisburger Institut f. Sprach- u. Sozialforschung.

Jahoda, M., Lazarsfeld, P.F. & Zeisel, H. (1933). Die Arbeitslosen von Marienthal. Leipzig: Hirzel.

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34, 183-202.

Jonkisz, E. & Moosbrugger, H. (2007). Planung und Entwicklung von psychologischen Tests und Fregebogen. In: Moosbrugger, H. & Kelava, A. (eds.), Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer, 27-72.

Kalveram, K. Th. (1970). Über Faktorenanalyse. Kritik eines theoretischen Konzepts und seine mathematische Neuformulierung. Arch. Psychol., 122, 92-118.

Kähler, W.-M. (2002). Statistische Datenanalyse. Verfahren verstehen und mit SPSS gekonnt einsetzen. 2. Auflage. Braunschweig: Vieweg.

Keats, J.A. (1974). Applications of projective transformations to test theory. Psychometrika, 39, 359-360.

Kelderman, H., & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. Journal of Educational Measurement, 27, 307-327.

Kelle, U. & Erzberger, Ch. (2000). Qualitative und quantitative Methoden: kein Gegensatz. In: Flick, U.; Kardoff, E.v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 299-309.

Kempf, W. (1972a). Probabilistische Modelle experimentalpsychologischer Versuchssituationen. Psychologische Beiträge, 14, 16-37.

Kempf, W. (1972b). Zur Bewertung der Faktorenanalyse als psychologische Methode. Psychologische Beiträge, 14, 610-625.

Kempf, W. (1974). Dynamische Modelle zur Messung sozialer Verhaltensdispositionen. In: Kempf, W. (ed.), Probabilistische Modelle in der Sozialpsychologie. Bern: Huber.

Kempf, W. (1977). Dynamic Models for the Measurement of "Traits" in Social Behavior. In: Kempf, W. & Repp, B. (eds.), Mathematical Models for Social Psychology. New York: Wiley.

Kempf, W. (1987). Psychologische Forschung als Begegnung. In: Brandtstädter, J. (ed.), Struktur und Erfahrung in der psychologischen Forschung. Berlin: de Gruyter, 169-207.

Kempf, W. (1992). Zum Verhältnis von qualitativen und quantitativen Methoden in der psychologischen Forschung. Forum Kritische Psychologie, 29, 89-108.

Kempf, W. (1994). Manipulierte Wirklichkeiten. In: Kempf, W. (ed.), Manipulierte Wirklichkeiten. Medienpsychologische Untersuchungen

der bundesdeutschen Presseberichterstattung im Golfkrieg. Münster: LIT, 3-27.

Kempf, W. (2000). Gewaltursachen und Gewaltdynamiken. In: Österreichisches Studienzentrum für Frieden und Konfliktlösung – ÖSFK (ed.), Konflikt und Gewalt. Ursachen – Entwicklungstendenzen – Perspektiven. Münster: Lit.

Kempf, W. (2001). News media and conflict escalation: A comparative study of the Gulf War coverage in American and European media. In: Nohrstedt, S.A. & Ottosen, R. (eds.), Journalism and the New World Order. Vol. 1. Gulf War, national news discourses and globalization. Göteborg: Nordicom, 149-174.

Kempf, W. (2002) Integration of quantitative and qualitative content analysis in media research. In: Kempf, W. & Luostarinen, H. (eds.), Journalism in the New World Order. Vol. II. Studying War and the Media. Göteborg: Nordicom, 145-172.

Kempf, W. (2003). Constructive conflict coverage. A social psychological research and development program. Conflict & communication online, 2/2.

Kempf, W., Baros, W. & Regener, I. (2000). Sozialpsychologische Rekonstruktion – Integration quantitativer und qualitativer Forschungsmethoden in der psychologischen Konflikt- und Friedensforschung. Forum Qualitative Sozialforschung.

Kempf, W. & Baumgärtner, U. (1996). Sozial-psychologische Rekonstruktion als textinterpretatives Verfahren. Diskussionsbeiträge der Projektgruppe Friedensforschung Konstanz, Nr. 31. Konstanz: KOPS.

Kempf, W., & Hampapa, P. (1977). Conditional inference for the dynamic test model. In: Kempf, W., Repp, B. (eds.), Mathematical models for social psychology. New York: Wiley, 81-100.

Kempf, W. & Meder, G. (1993). Theoretische und empirische Untersuchungen zur Restriktivität der Klassischen Testtheorie. Diagnostica, 39, 1, 1-21.

Kempf, W. & Palmbach, U. (1994). Sozial-psychologische Rekonstruktion massenmedialer Einflussnahme, In: Kempf, W. (ed.), Manipulierte Wirklichkeiten. Medienpsychologische Untersuchungen der bundesdeutschen Presseberichterstattung im Golfkrieg. Münster: LIT, 28-46.

Kempf, W. & Reimann, M. (1994). Die Berichterstattung über alliierte Kriegsgefangene. In: Kempf, W. (ed.), Manipulierte Wirklichkeiten. Medienpsychologische Untersuchungen der bundesdeutschen Presseberichterstattung im Golfkrieg. Münster: LIT, 82-101.

Kempf, W. & Reimann, M. (2002). The presentation of alternative ways to settle the Gulf Conflict in German, Norwegian and Finnish media. In: Kempf, W. & Luostarinen, H. (eds.), Journalism in the New World Order. Vol. II. Studying War and the Media. Göteborg: Nordicom, 203-226.

Kempf, W., Reimann, M. & Luostarinen, H. (1996). Qualitative Inhaltsanalyse von Kriegspropaganda und Kritischem Friedensjournalismus.

Diskussionsbeiträge der Projektgruppe Friedensforschung Konstanz, Nr. 32. Konstanz: KOPS.

Keupp, H. (1991). Für eine diskursive Psychologie jenseits des "possesiven Individualismus". In: Jüttemann, G. (eds.), Individuelle und soziale Regeln des Handelns. Beiträge zur Weiterentwicklung geisteswissenschaftlicher Ansätze in der Psychologie. Heidelberg, Asanger, 185-207.

Keupp, H. (1992). Das Subjekt und die Psychologie in der Krise der Moderne: Die Chancen postmoderner Provokationen. Psychologie und Gesellschaftskritik, 3-4, 16, 17-41.

Klauer, K.J. (1983). Kriteriumsorientierte Tests. In: Feger, H. & Bredenkamp, J. (eds.), Enzyklopädie der Psychologie. Band 3: Messen und Testen. Göttingen: Hogrefe.

Klauer, K.J. (1987). Kriteriumsorientierte Tests. Göttingen: Hogrefe.

König, H.D. (2000). Tiefenhermeneutik. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 556-569.

Koutsopoulos, C.J., (1962). The mathematical foundations of classical test theory. An axiomatic approach. I. Research Bulletin 62-17. Princeton, N.J.: Educational Testing Service.

Kowal, S. & O'Connell, D.C. (2000). Zur Transkription von Gesprächen. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 437-447.

Kracauer, S. (1952). The challenge of qualitative content analysis. Public Opinion Quaterly, 16, 631-642.

Kraus, W. (2001). Text. In: Keupp, H. & Weber, K. (eds.), Psychologie. Ein Grundkurs. Reinbek: Rowohlt.

Kreyszig, E. (1968). Statistische Methoden und ihre Anwendungen. Göttingen: Vandenhoeck & Ruprecht.

Krippendorff, K. (1980). Content analysis. An introduction to its methodology. Beverly Hills: Sage.

Kuder, G.F. & Richardson, M.W. (1937). The theory of the estimation of test reliability. Psychometrika, 2, 151-160.

Lamnek, S. (1988). Qualitative Sozialforschung. Band 1. Methodologie. München: Psychologie Verlags Union.

Lamnek, S. (1989). Qualitative Sozialforschung. Band 2. Methoden und Techniken. München: Psychologie Verlags Union.

Lamnek, S. (2005a). Qualitative Sozialforschung. 4. vollst. überarbeitete Auflage. Weinheim: Beltz.

Lamnek, S. (2005b). Gruppendiskussion. Theorie und Praxis. 2. überarbeitete und erweiterte Auflage. Weinheim: Beltz.

Lang, K. (1998). Entwicklung eines Kodiersystems für die Inhaltsanalyse von Konfliktkommunikation. Universität Konstanz: Psychol. Diplomarbeit.

Lasswell, H. (1927). Propaganda technique in the World War. London: Kegan Paul.

Lazarsfeld, P.F. (1950). Logical and mathematical foundations of latent structure analysis. In: Stouffer, S.A., Guttman, L., Suchman, E.A., Lazarsfeld, P.F., Star, S.A. & Clausen, J.A. (eds.), Studies in social psychology in world war II, Vol. IV. Measurement and Prediction. Princeton/N.J.: Princeton University Press, 362-412.

Lazarsfeld, P.F. (1980). Vorspruch zur neuen Auflage. In: Jahoda, M., Lazarsfeld, P.F. & Zeisel, H., Die Arbeitslosen von Marienthal. Frankfurt/Main: Suhrkamp.

Leithäuser, T. & Volmerg, B. (1977). Die Entwicklung einer empirischen Forschungsperspektive aus der Theorie des Alltagsbewußtseins. In: Leithäuser, T., Volmerg, B., Salje, G., Volmerg, U. & Wutka, B. Entwurf zu einer Empirie des Alltagsbewußtseins. Frankfurt/Main: Suhrkamp.

Leithäuser, T. & Volmerg, B. (1979). Anleitung zur empirischen Hermeneutik. Psychoanalytische Textinterpretation als sozialwissenschaftliches Verfahren. Frankfurt/Main: Suhrkamp.

Leithäuser, T. & Volmerg, B. (1988). Psychoanalyse in der Sozialforschung. Eine Einführung. Opladen: Westdeutscher Verlag.

Linacre, J.M. (1989). Many-facet Rasch measurement. Chicago: MESA Press.

Linacre, J.M. & Wright, B.D. (2002). Construction of measures from many-facet data. Journal of Applied Measurement, 3, 484-509.

Lindeberg, J.W. (1922). Eine neue Herleitung des Exponentialgesetzes der Wahrscheinlichkeitsrechnung. Math. Z., 15, 221.

Lindsay, J. & O'Connell (1995). How do transcribers deal with audio recordings of spoken discourse? Journal of psycholinguistic research, 24, 101-115.

Lippmann, W. (1922). Public Opinion. New York: McMillan.

Lissmann, U. (1997). Inhaltsanalyse von Texten. Landau: Verlag empirische Pädagogik.

Ljapunoff, A.M. (1901). Nouvelle forme du théorème sur la limite de probabilité. Mem. Acad. Sci. St. Pétersburg, 12, 5.

Lord, F.M. (1953). The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 13, 517-548.

Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Reading (Mass.): Addison-Wesley.

Lorenzer, A. (1981). Zum Beispiel "Der Malteser Falke". Analyse der psychoanalytischen Untersuchung literarischer Texte. In: Urban, B. & Kudszus, W. (eds.), Psychoanalytische und psychopathologische Literaturinterpretation. Darmstadt: Wissenschaftliche Buchgesellschaft, 23-46.

Lorenzer, A. (1986). Tiefenhermeneutische Kulturanalyse. In: König, H.D., Lorenzer, A., Lüdde, H., Naghol, S., Prokop, U., Schmid Noerr,

G. & Eggert, A. (eds.), Kultur-Analysen. Psychoanalytische Studien zur Kultur. Frankfurt/Main: Fischer, 11-98.

Lumsdaine, A.A. & Janis, I.L. (1953). Resistance to "counter propaganda" produced by one-sided and two-sided "propaganda" presentations. Public Opinion Quaterly, 17, 311-318.

Mattenschlager, A. & Riedle, H. (2003). Media construction of national identities in Germany and Switzerland, 1946-1995. Conflict & communication online, 2/1.

Mayring, P. (1983). Qualitative Inhaltsanalyse. Weinheim: Beltz.

Mayring, P. (2000). Qualitative Inhaltsanalyse. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 468-475.

Mayring, P. & Gläser-Zikuda, M. (eds.) (2005). Die Praxis der qualitativen Inhaltsanalyse. Weinheim: Beltz.

McHugh, R.B. (1956). Efficient estimation and local identification in Latent Class Analysis. Psychometrika, 21, 331-347.

McHugh, R.B. (1958). Note on "Efficient estimation and local identification in Latent Class Analysis". Psychometrika, 23, 273-274.

McNemar, Q. (1949). Psychological Statistics. New York: Wiley.

Meder, G. (1998). Zur Neubestimmung der Rolle der Bundeswehr in den deutschen Printmedien. In: Kempf, W. & Schmidt-Regener, I (eds.), Krieg, Nationalismus, Rassismus und die Medien. Münster: LIT, 201-210.

Mergenthaler, E. (1989). Die Transkription von Gesprächen. Ulm: Ulmer Textbank.

Merten, K. (1983). Inhaltsanalyse. Einführung in Theorie, Methode und Praxis. Opladen: Westdeutscher Verlag.

Merton, R.K. & Kendall, P.L. (1979). Das fokussierte Interview. In: Hopf, C. & Weingarten, E. (eds.), Qualitative Sozialforschung. Stuttgart: Klett-Cotta, 171-204.

Meschkowski, H. (1986). Mathematik verständlich dargestellt. 3. Auflage. München: Piper.

Micko, H.C. (1970). Eine Verallgemeinerung des Messmodells von Rasch mit einer Anwendung auf die Psychophysik der Reaktionen. Psychologische Beiträge 12, 4-22.

Mislevy, R.J. & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. Psychometrica, 55, 195-215.

Mitchell, C.R. (1981). Peacemaking and the consultant's role. New York: Nichols Publishing Company.

Mollenhauer, K. & Rittelmyer, C. (1977). Methoden der Erziehungswissenschaft. München: Juventa.

Moosbrugger, H. & Kelava, A. (eds.) (2007). Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer.

Moser, H. (1995). Grundlagen der Praxisforschung. Freiburg: Lambertus.

Neyman, J. & Scott, E.L. (1948). Consistent estimates based on partially consistent observations. Econometrica, 16, 1-32.

Nießen, M. (1977). Gruppendiskussion. Interpretative Methodologie – Methodenbegründung – Anwendung. München: Fink.

Nohrstedt, S.A. (2001). US dominance in Gulf War news. Propaganda relations between news discourses in US and European media. In: Nohrstedt, S.A. & Ottosen, R. (eds.), Journalism and the New World Order. Vol. 1. Gulf War, national news discourses and globalization. Göteborg: Nordicom, 175-215.

Novick, M. (1966). The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3, 1-18.

O'Connell, D.C. & Kowal, S. (1994). The transcriber as language user. In: Bartelt, G. (ed.), The dynamic of language processes: Essays in honor of Hand W. Dechert. Tübingen: Gunter Narr, 119-142.

Oevermann, U. (1983). Zur Sache: Die Bedeutung von Adornos methodologischem Selbstverständnis für die Begründung einer materialen soziologischen Strukturanalyse. In: Friedeburg, L.v. & Habermas, J. (eds.), Adorno-Konferenz 1983. Frankfurt/Main: Suhrkamp, 234-289.

Oevermann, U. (1993). Die objektive Hermeneutik als unverzichtbare methodologische Grundlage für die Analyse von Subjektivität. Zugleich eine Kritik der Tiefenhermeneutik. In: Jung, T. & Müller-Doohm, S. (eds.), „Wirklichkeit" im Deutungsprozess. Verstehen in den Kultur- und Sozialwissenschaften. Frankfurt/Main: Suhrkamp, 106-189.

Parker, I. (2000). Die diskursanalytische Methode. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 546-556.

Pfanzagl, J. (1966). Allgemeine Methodenlehre der Statistik, Band 2. 2. Auflage. Berlin: de Gruyter.

Pfanzagl, J. (1993). A case of asymptotic equivalence between conditional and marginal maximum likelihood estimators. Journal of Statistical Planning and Inference, 35, 301-307.

Plontz, A. (2005). Mediale Verarbeitung eines politischen Konfliktes: Berichterstattung in der deutschen und in der französischen Presse über den deutsch-französischen Konflikt über den Vorsitz der Europäischen Zentralbank. Universität Konstanz: Psychol. Diplomarbeit.

Plontz,A. (2006). Der deutsch-französische Konflikt um die Präsidentschaft der Europäischen Zentralbank in der deutschen und französischen Presse. Conflict & communication online 5/1.

Potter, J. & Wetherell, M. (1987). Discourse and Social Psychology: Beyond attitudes and behaviour. London: Sage.

Potter, J. & Wetherell, M. (1995). Soziale Repräsentation, Diskursanalyse und Rassismus. In: Flick, U. (ed.), Psychologie des Sozialen. Repräsentationen in Wissen und Sprache. Reinbek: Rowohlt, 177-200.

Projektgruppe Friedensforschung Konstanz – PFK (ed.) (2005). Nachrichtenmedien als Mediatoren von Peace-Building, Demokratisierung und Versöhnung in Nachkriegsgesellschaften. Berlin: verlag irena regener.

Psathas, G. (ed.) (1979). Everyday language: Studies in ethnomethodology. New York: Irvington.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. Berkeley: University of California Press.

Rasch, G. (1965). Unveröffentlichtes Vorlesungsmanuskript. Verfasst von J. Stene, übersetzt von B. Repp.

Rasch, G. (1968). A mathematical theory of objectivity and its consequences for model construction. Proceedings of the European Meeting on Statistic, Econometrics and Management Science, Amsterdam, September 2-7, 1968.

Reetz, K.-D. (2007). Migration und schulischer Misserfolg italienischer Kinder. Berlin: verlag irena regener.

Reichertz, J. (2000). Objektive Hermeneutik und hermeneutische Wissenssoziologie. In: Flick, U., Kardoff, E. v. & Steinke, I. (eds.), Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, 514-524.

Reimann, M. (1997). Alternative Konfliktlösungsoptionen im Golfkrieg. Eine qualitative Inhaltsanalyse deutscher Medien (1990-1993). Diskussionsbeiträge der Projektgruppe Friedensforschung Konstanz, Nr. 36. Konstanz: KOPS.

Reimann, M. (2002). Communication disorders in conflict coverage. In: Kempf, W. & Luostarinen, H. (eds.), Journalism in the New World Order. Vol. II. Studying War and the Media. Göteborg: Nordicom, 173-181.

Richardson, S.A., Snell Dohrendwend, B. & Klein, D. (1979). Die "Suggestivfrage". Erwartungen und Unterstellungen im Interview. In: Hopf, C. & Weingarten, E. (eds.), Qualitative Sozialforschung. Stuttgart: Klett-Cotta, 205-231.

Rindskopf, D. (1987). Using Latent Class Analysis to test developmental models. Developmental Review, 7, 66-85.

Rogers, C. (1972). Die klientenzentrierte Gesprächspsychotherapie. München: Kindler.

Rohracher, H. (1971). Einführung in die Psychologie. 10. Auflage. Wien: Urban & Schwarzenberg.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. Applied Psychological Measurement, 14, 271-282.

Rost, J. (1996). Lehrbuch Testtheorie Testkonstruktion. Bern: Huber.

Scheiblechner, H. (1970). Eine parameterfreie Modellkontrolle des eindimensionalen probabilistischen Meßmodells von Rasch. Univ. Wien: Hektographiertes Manuskript.

Scheele, B. & Groeben, N. (1988). Dialog-Konsens-Methoden zur Rekonstruktion subjektiver Theorien. Tübingen: Francke.

Schenkein, J. (ed.) (1978). Studies in the organization of conversational interaction. New York: Academic Press.

Schluroff, M. (1998). Grundlagen der gegenwärtigen deutschen Militärpolitik. Die "Verteidigungspolitischen Richtlinien" von 1992. In: Kempf, W. & Schmidt-Regener, I (eds.), Krieg, Nationalismus, Rassismus und die Medien. Münster: LIT, 195-200.

Schmetterer, L. (1966). Einführung in die mathematische Statistik. 2. Auflage. Wien: Springer.

Schulz, W. (1976). Die Konstruktion von Realität in den Nachrichtenmedien. Freiburg: Alber.

Schulz von Thun, F. (1981). Miteinander reden 1. Störungen und Klärungen. Allgemeine Psychologie der Kommunikation. Reinbek: Rowohlt.

Schwemmer, O. (1987). Das Allgemeine unseres Handelns: Zum Sinn der Rede von Gesetzen und Regeln für die Darstellung menschlichen Handelns. In: Brandtstädter, J. (ed.), Struktur und Erfahrung in der psychologischen Forschung. Berlin: de Gruyter, 104-124.

Schütze, F. (1977). Die Technik des narrativen Interviews in Interaktionsfeldstudien – dargestellt an einem Projekt zur Erforschung von kommunalen Machtstrukturen. Universität Bielefeld, Fakultät für Soziologie: Arbeitsberichte und Forschungsmaterialien, Nr. 1.

Schütze, F. (1982). Narrative Repräsentation kollektiver Schicksalsbetroffenheit. In: Lämmert, C. (ed.), Erzählforschung: ein Symposium. Stuttgart: Metzler.

Schütze, F. (1983). Biographieforschung und narratives Interview. Neue Praxis, 13, 283-293.

Sève, L. (1978). Marxistische Analyse der Entfremdung. Frankfurt/Main: Verlag Marxistische Blätter.

Siegler, R.S. (1976). Three aspects of cognitive development. Cognitive Psychology, 8, 481-520.

Sommer, J. (1987). Dialogische Forschungsmethoden. München: Psychologie Verlags Union.

Spearman, C. (1904). "General intelligence", objectively determined and measured. American Journal of Psychology, 15, 201-293.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. American Journal of Psychology, 18, 161-169.

Spöhring, W. (1989). Qualitative Sozialforschung. Stuttgart: Teubner.

Staab, J.F. (1990). Nachrichtenwert-Theorie: formale Struktur und empirischer Gehalt. Freiburg: Alber.

Strauss, A.L. (1991). Grundlagen qualitativer Sozialforschung – Datenanalyse und Theoriebildung in der empirischen soziologischen Forschung. München: Fink.

Strauss, A.L. & Corbin, J. (1996). Grundlagen qualitativer Sozialforschung. Weinheim: Psychologie Verlags Union.

Sudnow, D. (ed.) (1972). Studies in social interaction. New York: Free Press.

Tarnai, C. & Bos, W. (1989). Emanzipation und Emanzipatoren. Definition des Begriffs Emanzipation durch Studierende der Pädagogik. In: Bos, W. & Tarnai, C. (eds.), Angewandte Inhaltsanalyse in empirischer Pädagogik und Psychologie. Münster: Waxmann, 155-172.

Tausch, R. & Tausch, A.M. (1981). Gesprächs-Psychotherapie. Einfühlsame hilfreiche Gruppen- und Einzelgespräche in Psychotherapie und im alltäglichen Leben. 8. Auflage. Göttingen: Hogrefe.

Terhart, E. (1981). Intuition – Interpretation – Argumentation. Zum Problem der Geltungsbegründung von Interpretationen. Zeitschrift für Pädagogik, 27/5, 769-793.

Thurstone, L.L. (1931). Multiple factor analysis. Psychological Review, 38, 406-427.

Thurstone, L.L. (1947). Multiple factor analysis. Chicago: University of Chicago Press.

Tijms, H. (2004). Understanding Probability: Change Rules in Everyday Life. Cambridge: Cambridge University Press.

Turner, R. (ed.) (1974). Ethnomethodology: Selected readings. Harmondsworth: Penguin.

Verhelst, N.D. & Glas, C.A.W. (1993). A dynamic generalization of the Rasch model. Psychometrika, 58, 395-415.

Volmerg, B. (1980). Das Verfahren der psychoanalytischen Textinterpretation am Beispiel eines Gruppendiskussionsprotokolls. In: Heinze, T., Klusemann, H.-W. & Soeffner, H.-G. (eds.), Interpretationen einer Bildungsgeschichte. Überlegungen zur sozialwissenschaftlichen Hermeneutik. Bensheim: päd.-extra Buchverlag, 202-234.

Volmerg, B. (1988). Wie interpretiert man Texte tiefenhermeneutisch? In: Leithäuser, Volmerg (eds.), Psychoanalyse in der Sozialforschung. Opladen: Westdeutscher Verlag, 251-261.

Volmerg, B., Volmerg, U. & Leithäuser, Th. (1983). Kriegsängste und Sicherheitsbedürfnis. Zur Sozialpsychologie des Ost-West-Konflikts im Alltag. Frankfurt/Main: Fischer.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response models. Psychometrika, 54, 427-450.

Watzlawick, P., Beavin, J.K. & Jackson, D.D. (1967). Pragmatics of human communication. A study of interactional patterns, pathologies, and paradoxes. New York: Norton & Company.

Wilke, S. (1992). Die erste Begegnung. Eine konversations- und inhaltsanalytische Untersuchung der Interaktion im psychoanalytischen Erstgespräch. Heidelberg: Asanger.

Wirtz, M. & Caspar, F. (2002). Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen. Göttingen: Hogrefe.

Witzel, A. (1982). Verfahren der qualitativen Sozialforschung. Überblick und Alternativen. Frankfurt/Main: Campus.

Anhang

Über den Autor

Wilhelm Kempf, Prof. Dr. phil. habil., wurde 1947 in Klagenfurt/ Österreich geboren. 1965-1970 Studium der Psychologie, Philosophie und Statistik an der Universität Wien. 1969/70 Postgraduiertenstudium der Soziologie am Institut für Höhere Studien. 1970 Promotion an der Universität Wien. 1977 Habilitation an der Universität Erlangen-Nürnberg, *venia legendi* für Psychologie.

Seit 1977 Professor für psychologische Methodenlehre und Leiter der Projektgruppe Friedensforschung an der Universität Konstanz. Lehr- und Forschungsaufenthalte an der Universidad de Costa Rica, an der Universität Tampere (Finnland) und an der York University (Toronto, Kanada).

Arbeitsschwerpunkte: Psychologische Methodologie, Friedens- und Konfliktforschung.

Aktuelle Buchpublikationen: *Constructive Conflict Coverage – A Social Psychological Approach* (hrsg. vom Austrian Study Center for Peace and Conflict Resolution, 2003); *Nachrichtenmedien als Mediatoren von Peace-Building, Demokratisierung und Versöhnung in Nachkriegsgesellschaften* (hrsg. für die Projektgruppe Friedensforschung Konstanz, 2005); *Forschungsmethoden der Psychologie, Band I: Theorie und Empirie* (2. überarbeitete und erweiterte Auflage, 2006)

Adresse: Universität Konstanz, FB Psychologie
D – 78457 Konstanz
eMail: Wilhelm.Kempf@uni-konstanz.de
Website: www.uni-konstanz.de/FuF/SozWiss/fg-psy/ag-meth/

Abkürzungen und Notationen

AV	Abhängige Variable	Abkürzungen
BM	Binomialmodell	
CML	Conditional Maximum-Likelihood	
DTM	Dynamisches Testmodell	
EEG	Elektroenzephalogramm	
KT	Klassische Testtheorie	
LCA	Latent-Class-Analyse	
LLTM	Lineares logistisches Testmodell	
MEG	Magnetenzephalogramm	
ML	Maximum-Likelihood	
MML	Marginal Maximum Likelihood	
N(0,1)	Standardnormalverteilung	
RM	Rasch-Modell	
SV	Störvariable	
UML	Unconditional Maximum-Likelihood	
UV	Unabhängige Variable	
VB	Versuchsbedingung	
VL	Versuchsleiter	
Vp	Versuchsperson	
WML	Weighted Maximum-Likelihood	
$a = 1,\ldots,z$	Experimentelle Faktoren	Indizes
$b = 0,\ldots,c-1$	Faktorstufen	
$g = 1,\ldots,h$	Klassen von Vpn	
$i = 1,\ldots,k$	Items	
$j = 1,\ldots,m$	Basiskomponenten (LLTM)	
$t = 1,\ldots,n$	Tests (KT)	
$v = 1,\ldots,n$	Vpn	
$x = 0,\ldots,m_i-1$	Antwortkategorien	
$Q = ((q_{ij}))$	Designmatrix (LLTM)	Matrizen und Vektoren
Q_*	Erweiterte Designmatrix	
Q_0	Reduzierte Designmatrix	
$X = ((x_{vi}))$	Antwortmatrix	
$x_v = (x_{v1},\ldots,x_{vk})$	Antwortvektor	

Variablen	F_{vt}, $F_{vt'}$	Individuelle Fehlervariablen (KT)
	X_{vt}, $X_{vt'}$	Individuelle Maßzahlvariablen (KT)
	F_{ot}, $F_{ot'}$	Fehlervariablen in der Referenzpopulation (KT)
	X_{ot}, $X_{ot'}$	Maßzahlvariablen in der Referenzpopulation (KT)
	Y_{ot}, $Y_{ot'}$	Kriteriumsvariablen in der Referenzpopulation (KT)
	T_{ot}, $T_{ot'}$	True-Score-Variablen in der Referenzpopulation (KT)
	X_{vi}	Antwortvariablen
		$X_{v\bullet}$ Itemunabhängig
		$X_{\bullet i}$ Personenunabhängig
		$X_{\bullet\bullet}$ Item- und personenunabhängig
Beobachtungen	x_{vi}	Itemantwort
	x_{vt}	Testscore (KT)
	x_{vo}	Summenscore
	$x_{vo}^{(i)}$	Partieller Summenscore
	$x_{v\bullet}$	Gewichteter Summenscore
	x_{oi}	Gesamtzahl der bei Item i erzielten Punkte
Häufigkeiten	e	Erwartete Häufigkeiten
	n	Empirische Häufigkeiten
		n_g Anzahl der Vpn in Klasse g
		n_{oox} Häufigkeit, mit welcher in Kategorie x geantwortet wird
		n_{oix} Häufigkeit, mit welcher Item i in Kategorie x beantwortet wird
		n_{vox} Häufigkeit, mit welcher Vp v in Kategorie x antwortet
Wahrscheinlichkeitsparameter	p_{vi}	Lösungswahrscheinlichkeiten
	$p_{vi.r}$	Bedingte Lösungswahrscheinlichkeiten (DTM)
	p_{vix}	Kategorienwahrscheinlichkeiten
		$p_{v\bullet x}$ Itemunabhängig
		$p_{\bullet ix}$ Personenunabhängig
		$p_{\bullet\bullet x}$ Item- und personenunabhäng
	p_g	Klassengrößen
	p_{gix}	Klassenspezifische Kategorienwahrscheinlichkeiten (LCA)
	$p_{g.x_v}$	Membership-Wahrscheinlichkeiten (LCA)
Itemparameter	α_i	Trennschärfe (Birnbaum-Modell)
	α_{ix}	Schwellenparameter (Ordinales RM)

δ_i	Itemschwierigkeit (RM)	
δ_{ix}	Kategorienschwierigkeit (Polytomes RM)	
δ_{ig}	Klassenspezifische Itemschwierigkeit (Mischverteilungs RM)	
ε_i	Itemleichtigkeit (delogarithmiert) (RM, DTM)	
σ_i	Itemschwierigkeit (delogarithmiert) (RM, DTM)	
θ_v	Personen-Fähigkeit (RM)	Personenparameter
ξ_v	Personen-Fähigkeit (delogarithmiert) (RM, DTM)	
τ_{vt}	True-Score (KT)	
β_j	Effekt der VB (delogarithmiert) (Mehrfaktorielles RM)	Effektparameter
η_j	Leichtigkeit der Basiskomponente (LLTM)	
$\kappa_{b,b'}^{(a,a')}$	Interaktionseffekt der Stufen b und b' der Faktoren a und a' (Mehrfaktorielles RM)	
ψ_r	Transferparameter (DTM)	
$\omega_b^{(a)}$	(Haupt-) Effekt der Faktorstufe b des Faktors a (Mehrfaktorielles RM)	
$\omega_{b,b'}^{(a,a')}$	Gemeinsamer Effekt der Stufen b und b' der Faktoren a und a' (Mehrfaktorielles RM)	
$f_{vi}(x)$	Wahrscheinlichkeitsdichte der Antwortvariable X_{vi} $f_{v\bullet}(x)$ Itemunabhängig $f_{\bullet i}(x)$ Personenunabhängig $f_{\bullet\bullet}(x)$ Item- und personenunabhängig	Wahrscheinlichkeits- und Dichtefunktionen
$f_{ix}(\theta_v)$, $f_i(\theta_v)$	Itemcharakteristikfunktionen	
$f_{ix.r}(\theta_v)$, $f_{i.r}(\xi_v)$	Bedingte Itemcharakteristikfunktionen	
$g(\theta)$	Wahrscheinlichkeitsdichte der latenten Personenvariable Θ	
$mL(X) = L(X)$	Marginale Likelihood der Antwortmatrix	Likelihoodfunktionen
$cL(X) = L(X \mid x_{1o},\ldots,x_{no})$	Bedingte Likelihood der Antwortmatrix	
$uL(X) = L(X \mid \theta_1,\ldots,\theta_n)$	Unbedingte Likelihood der Antwortmatrix	
$L(x_v \mid \theta_v)$	(Unbedingte) Likelihood des Antwortvektors	
$L(\theta_v \mid x_v)$	Likelihood des Personenparameters bei gegebenem Antwortmuster	
$n(P)$	Anzahl der unabhängigen Modellparameter, die zur Darstellung der Likelihoodfunktion erforderlich sind $n(P_{BM})$ Binomialmodell	Anzahl unabhängiger Modellparameter

		$n(P_{CR})$ Rasch-Modell, bedingte Likelihood $n(P_{LC1})$ Ein-Klassen-Lösung der LCA $n(P_{LCh})$ Latent-Class-Analyse $n(P_{MR})$ Rasch-Modell, marginale Likelihood $n(P_{PR})$ Pure-Random-Modell $n(P_{sat})$ Saturiertes Modell
Informations-funktionen und Fehlerstreuung der Parameterschätzung	$I(\vartheta)$	Informationsfunktion des Parameters ϑ $I(\delta_i)$ Itemparameter (RM) $I(\theta_v)$ Personenparameter (RM)
	$\sigma(\hat{\vartheta})$	Fehlerstreuung der Schätzfunktion für den Parameter ϑ $\sigma(\hat{\tau}_{vt})$ True-Score (KT) $\sigma(\hat{p}_g)$ Klassengrößeparameter $\sigma(\hat{p}_{gix})$ Klassenspezifische Kategorienwahrscheinlichkeiten (LCA) $\sigma(\hat{p}_{v\bullet x})$ Itemunabhängige Kategorienwahrscheinlichkeiten (z. B. Binomialmodell) $\sigma(\hat{p}_{\bullet ix})$ Personenunabhängige Kategorienwahrscheinlichkeiten (Ein-Klassen-LCA) $\sigma(\hat{p}_{\bullet\bullet x})$ Item- und personenunabhängige Kategorienwahrscheinlichkeiten (Pure-Random-Modell) $\sigma(\hat{\delta}_i)$ Itemparameter (RM) $\sigma(\hat{\theta}_v)$ Personenparameter (RM)
Kennzahlen der klassischen Testtheorie	α_2, α_n, α_k	Koeffizient Alpha nach Cronbach
	$\rho_{XY} = \rho(X_{ot}, Y_{ot'})$	Validität des Tests t bezüglich der Kriteriumsvariable Y
	$\rho^2_{XT} = \rho^2(X_{ot}, T_{ot})$	Reliabilität des Tests t
	$\rho^2_{XT_o} = \rho(X_{ot}, X_{ot'})$	Reliabilität der (essentiell) parallelen Subtests t und t'
	$\sigma(F_{ot})$	Standardmessfehler des Tests t

Das griechische Alphabet

Alpha	α	Α
Beta	β	Β
Gamma	γ	Γ
Delta	δ	Δ
Epsilon	ε	Ε
Zeta	ζ	Ζ
Eta	η	Η
Theta	θ, ϑ	Θ
Iota	ι	Ι
Kappa	κ	Κ
Lambda	λ	Λ
My	μ	Μ
Ny	ν	Ν
Xi	ξ	Ξ
Omikron	ο	Ο
Pi	π	Π
Rho	ρ	Ρ
Sigma	σ	Σ
Tau	τ	Τ
Ypsilon	υ	Υ
Phi	φ	Φ
Chi	χ	Χ
Psi	ψ	Ψ
Omega	ω	Ω

Terminologische Tabellen

1 Bewertung des Koeffizienten κ

Über 0.75	Sehr gute Übereinstimmung
0.60 – 0.75	Gute Übereinstimmung
0.40 – 0.60	Akzeptable Übereinstimmung

2 Bescheibung von LCA-Ergebnissen

0% – 1%	Nie (so gut wie nie)
Unter 10%	Kaum/selten
10% – 25%	Gelegentlich/mitunter/nicht selten
25% – 75%	(Relativ bis sehr) deutlich/häufig/oft
75% – 90%	Meistens
Über 90%	(Ohne Qualifizierung)/in der Regel
99% – 100%	(So gut wie ausnahmslos)/ausnahmslos

Statistische Tabellen

1 Binomialkoeffizienten

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

x	n=1	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	n=11
0	1	1	1	1	1	1	1	1	1	1	1
1	1	2	3	4	5	6	7	8	9	10	11
2		1	3	6	10	15	21	28	36	45	55
3			1	4	10	20	35	56	84	120	165
4				1	5	15	35	70	126	210	330
5					1	6	21	56	126	252	462
6						1	7	28	210	462	462
7							1	8	36	120	330
8								1	9	45	165
9									1	10	55
10										1	11
11											1

x	n=12	n=13	n=14	n=15	n=16	n=17	n=18	n=19	n=20
0	1	1	1	1	1	1	1	1	1
1	12	13	14	15	16	17	18	19	20
2	66	78	91	105	120	136	153	171	190
3	220	286	364	455	560	680	816	969	1140
4	495	715	1001	1365	1820	2380	3060	3876	4845
5	792	1287	2002	3003	4368	6188	8568	11628	15504
6	924	1716	3003	5005	8008	12376	18564	27132	38760
7	792	1716	3432	6435	11440	19448	31824	50388	77520
8	495	1287	3003	6435	12870	24310	43758	75582	125970
9	220	715	2002	5005	11440	24310	48620	92378	167960
10	66	286	1001	3003	8008	19448	43758	92378	184756
11	12	78	364	1365	4368	12376	31824	75582	167960
12	1	13	91	455	1820	6188	18564	50388	125970
13		1	14	105	560	2380	8568	27132	77520
14			1	15	120	680	3060	11628	38760
15				1	16	136	816	3876	15504
16					1	17	153	969	4845
17						1	18	171	1140
18							1	19	190
19								1	20
20									1

2 Binomialverteilung für p = 0.5

$$F(x) = \sum_{i=0}^{x} \binom{n}{x} p^x (1-p)^{n-x}$$

n	Werte von x für									
	prob(X ≤ x) = F(x) ≤					prob(X ≥ x) = 1 - F(x-1) ≤				
	0.001	0.005	0.01	0.025	0.05	0.05	0.025	0.01	0.005	0.001
5	–	–	–	–	0	5	–	–	–	–
6	–	–	–	0	0	6	6	–	–	–
7	–	–	0	0	0	7	7	7	–	–
8	–	0	0	0	1	7	8	8	8	–
9	–	0	0	1	1	8	8	9	9	–
10	0	0	0	1	1	9	9	10	10	10
11	0	0	1	1	2	9	10	10	11	11
12	0	1	1	2	2	10	10	11	11	12
13	0	1	1	2	3	10	11	12	12	13
14	1	1	2	2	3	11	12	12	13	13
15	1	2	2	3	3	12	12	13	13	14
16	1	2	2	3	4	12	13	14	14	15
17	1	2	3	4	4	13	13	14	15	16
18	2	3	3	4	5	13	14	15	15	16
19	2	3	4	4	5	14	15	15	16	17
20	2	3	4	5	5	15	15	16	17	18
21	3	4	4	5	6	15	16	17	17	18
22	3	4	5	5	6	16	17	17	18	19
23	3	4	5	6	7	16	17	18	19	20
24	4	5	5	6	7	17	18	19	19	20
25	4	5	6	7	7	18	18	19	20	21
	Näherung für n > 25									
	prob(X ≤ x) = Φ(z)					prob(X ≥ x) = 1 - Φ(z)				
	$z = \frac{2x+1-n}{\sqrt{n}}$					$z = \frac{2x-1-n}{\sqrt{n}}$				

3 Standardnormalverteilung

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}} du$$

Werte von z für									
prob(Z ≤ z) = Φ(z) =					prob(Z ≥ z) = 1 - Φ(z) =				
0.001	0.005	0.01	0.025	0.05	0.05	0.025	0.01	0.005	0.001
-3.090	-2.576	-2.326	-1.960	-1.645	1.645	1.960	2.326	2.576	3.090

Werte von z für		
prob(-z ≤ Z ≤ z) = Φ(z) - Φ(-z) =		
0.95	0.99	0.999
1.960	2.576	3.291

4 t-Verteilung

$$F(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \int_{-\infty}^{x} \frac{du}{\left(1+\frac{u^2}{n}\right)^{(n+1)/2}}$$

	Werte von x für									
	prob(X ≤ x) = F(x) =					prob(X ≥ x) = 1 - F(x) =				
	0.001	0.005	0.01	0.025	0.05	0.05	0.025	0.01	0.005	0.001
1	-318.3	-63.7	-31.8	-12.7	-6.31	-6.31	12.7	31.8	63.7	318.3
2	-22.3	-9.93	-6.97	-4.30	2.92-	2.92	4.30	6.97	9.93	22.3
3	-10.2	-5.84	-4.54	-3.18	-2.35	2.35	3.18	4.54	5.84	10.2
4	-7.17	-4.60	-3.75	-2.78	-2.13	2.13	2.78	3.75	4.60	7.17
5	-5.89	-4.03	-3.37	-2.57	-2.02	2.02	2.57	3.37	4.03	5.89

	Werte von x für									
	prob(X ≤ x) = F(x) =					prob(X ≥ x) = 1 - F(x) =				
	0.001	0.005	0.01	0.025	0.05	0.05	0.025	0.01	0.005	0.001
6	-5.21	-3.71	-3.14	-2.45	-1.94	1.94	2.45	3.14	3.71	5.21
7	-4.79	-3.50	-3.00	-2.37	-1.90	1.90	2.37	3.00	3.50	4.79
8	-4.50	-3.36	-2.90	-2.31	-1.86	1.86	2.31	2.90	3.36	4.50
9	-4.30	-3.25	-2.82	-2.26	-1.83	1.83	2.26	2.82	3.25	4.30
10	-4.14	-3.17	-2.76	-2.23	-1.81	1.81	2.23	2.76	3.17	4.14
11	-4.03	-3.11	-2.72	-2.20	-1.80	1.80	2.20	2.72	3.11	4.03
12	-3.93	-3.06	-2.68	-2.18	-1.78	1.78	2.18	2.68	3.06	3.93
13	-3.85	-3.01	-2.65	-2.16	-1.77	1.77	2.16	2.65	3.01	3.85
14	-3.79	-2.98	-2.62	-2.15	-1.76	1.76	2.15	2.62	2.98	3.79
15	-3.73	-2.95	-2.60	-2.13	-1.75	1.75	2.13	2.60	2.95	3.73
16	-3.69	-2.92	-2.58	-2.12	-1.75	1.75	2.12	2.58	2.92	3.69
17	-3.65	-2.90	-2.57	-2.11	-1.74	1.74	2.11	2.57	2.90	3.65
18	-3.61	-2.88	-2.55	-2.10	-1.73	1.73	2.10	2.55	2.88	3.61
19	-3.58	-2.86	-2.54	-2.09	-1.73	1.73	2.09	2.54	2.86	3.58
20	-3.55	-2.85	-2.53	-2.09	-1.73	1.73	2.09	2.53	2.85	3.55
22	-3.51	-2.82	-2.51	-2.07	-1.72	1.72	2.07	2.51	2.82	3.51
24	-3.47	-2.80	-2.49	-2.06	-1.71	1.71	2.06	2.49	2.80	3.47
26	-3.44	-2.78	-2.48	-2.06	-1.71	1.71	2.06	2.48	2.78	3.44
28	-3.41	-2.76	-2.47	-2.05	-1.70	1.70	2.05	2.47	2.76	3.41
30	-3.39	-2.75	-2.46	-2.04	-1.70	1.70	2.04	2.46	2.75	3.39
40	-3.31	-2.70	-2.42	-2.02	-1.68	1.68	2.02	2.42	2.70	3.31
50	-3.26	-2.68	-2.40	-2.01	-1.68	1.68	2.01	2.40	2.68	3.26
100	-3.17	-2.63	-2.37	-1.98	-1.66	1.66	1.98	2.37	2.63	3.17
200	-3.13	-2.60	-2.35	-1.97	-1.65	1.65	1.97	2.35	2.60	3.13
∞	-3.09	-2.58	-2.33	-1.96	-1.65	1.65	1.96	2.33	2.58	3.09

5 χ^2-Verteilung

$$F(x) = K_n \int_0^x u^{(n-2)/2} e^{-u/2} du \qquad \text{für } x \geq 0$$

	Werte von x für						Werte von x für				
df	prob(X ≥ x) = 1 – F(x) =					df	prob(X ≥ x) = 1 – F(x) =				
	0.05	0.025	0.01	0.005	0.001		0.05	0.025	0.01	0.005	0.001
1	3.84	5.02	6.63	7.88	10.83	23	35.17	38.08	41.64	44.18	49.73
2	5.99	7.38	9.21	10.60	13.82	24	36.42	39.36	42.98	45.56	51.18
3	7.81	9.35	11.34	12.84	16.27	25	37.65	40.65	44.31	46.93	52.62
4	9.49	11.14	13.28	14.86	18.47	26	38.89	41.92	45.64	48.29	54.05
5	11.07	12.83	15.09	16.75	20.52	27	40.11	43.19	46.96	49.64	55.48
6	12.59	14.45	16.81	18.55	22.46	28	41.34	44.46	48.28	50.99	56.89
7	14.07	16.01	18.48	20.28	24.32	29	42.56	45.72	49.59	52.34	58.30
8	15.51	17.53	20.09	21.96	26.13	30	43.77	46.98	50.89	53.67	59.70
9	16.92	19.02	21.67	23.59	27.88	40	55.76	59.34	63.69	66.77	73.40
10	18.31	20.48	23.21	25.19	29.59	50	67.50	71.42	76.15	79.49	86.66
11	19.68	21.92	24.73	26.76	31.26	60	79.08	83.30	88.38	91.95	99.61
12	21.03	23.34	26.22	28.30	32.91	70	90.53	95.02	100.43	104.22	112.32
13	22.36	24.74	27.69	29.82	34.53	80	101.88	106.63	112.33	116.32	124.84
14	23.68	26.12	29.14	31.32	36.12	90	113.15	118.14	124.12	128.30	137.21
15	25.00	27.49	30.58	32.80	37.70	100	124.34	129.56	135.81	140.17	149.45
16	26.30	28.85	32.00	34.27	39.25		Näherung für df > 100				
17	27.59	30.19	33.41	35.72	40.79		$x = \frac{1}{2}\left(\sqrt{2df - 1} + z\right)^2$				
18	28.87	31.53	34.81	37.16	42.31						
19	30.14	32.85	36.19	38.58	43.82		Werte von z für				
20	31.41	34.17	37.57	40.00	45.32		prob(X ≥ x) = 1 – F(x) =				
21	32.67	35.48	38.93	41.40	46.80		0.05	0.025	0.01	0.005	0.001
22	33.92	36.78	40.29	42.80	48.27		1.645	1.960	2.326	2.576	3.090

Aufgabenlösungen

Zu Kapitel 2.1

1. a) LA = 9.33
   b) IA = 8.2
   c) IQ = 87.89
   d) Rückstand
2. x_v = (1,1,1,1,1,1,0,1,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0)
3. x_{vt} = 11
4. a) (x_{oi}) = (2,3,6,7,7)
   b) (x_{vo}) = (1,2,3,4,1,2,4,3,2,3)
5. IQ = 131.43

Zu Kapitel 2.2

1. ρ_{XY} = 0.318
2. a) Retest-Reliabilität = 0.808
   b) (Zumindest essentielle) Parallelität von Test und Retest
   c) Wenn Test und Retest (zumindest essentiell) τ-äquivalent sind, ist die Retest-Reliabilität gleich dem geometrischen Mittel der Reliabilitäten der beiden Testungen. Ist auch diese Voraussetzung nicht erfüllt, so ist die Retest-Reliabilität kleiner als das geometrische Mittel der beiden Reliabilitäten.
3. a) Paralleltest-Reliabilität = 0.974
   b) im Falle i) ist die Reliabilität = 0.974, im Falle ii) ist das geometrische Mittel der Reliabilitäten der beiden Tests = 0.974 und im Falle iii) ist das geometrische Mittel der beiden Reliabilitäten > 0.974.
4. Maximal 0.949
5. 90% der Varianz der Testscores ist systematische Varianz, 10% sind Fehlervarianz.
6. $\sigma(F_{ot})$ = 3.674
7. a) Es ist eine Reliabilität von ρ^2_{XT} = 0.907 zu erwarten.
   b) Die beiden Parallelformen müssen tatsächlich (zumindest essentiell) parallel sein.
8. a) Für die Kurzform ist nur noch mit einer Reliabilität von 0.73 zu rechnen.
   b) Um eine Reliabilität von 0.8 zu gewährleisten, muss die Kurzform mindestens 15 Items umfassen.
   c) Bei Erweiterung auf 50 Items ist mit einer Reliabilität von 0.931 zu rechnen.
   d) Um eine Reliabilität von 0.95 zu erzielen, müsste der Test auf 71 Items (2.35-fache Testlänge) erweitert werden.
9. ρ^2_{XT} = 0.788
10. a) $\rho^2_{XT_o}$ = 0.506
    b) ρ^2_{XT} = 0.754
11. Die Reliabilität des Tests beträgt mindestens α_2 = 0.425.

12. Die Reliabilität des Tests beträgt mindestens $\alpha_k = 0.83$.

13. a) Paralleltest-Reliabilität: zumindest essentielle Parallelität des Tests und seiner Parallelform (Form A des IST-70);
Split-Half: Mindestens essentielle Parallelität der gebildetenTesthälften;
Koeffizient Alpha (α_2): Mindestens essentielle τ-Äquivalenz der Testhälften;
Kuder-Richardson 20 (α_k): gleiche Lösungswahrscheinlichkeiten (da essentielle τ-Äquivalenz mit binären Items nicht kompatibel und gewöhnliche τ-Äquivalenz bei binären Items wegen $\sigma^2(X_{oi}) = p_{oi}(1-p_{oi})$ inhaltsgleich mit Paralllelität und/oder Gleichheit der Lösungswahrscheinlichkeiten ist).
b) Da die Reliabilität des Tests mindestens $\max(\alpha_2, \alpha_k) = 0.89$ beträgt, sind die Voraussetzungen der Paralleltest-Reliabilität und der Kuder-Richardson 20 offensichtlich nicht erfüllt. Wegen Split-Half = Koeffizient Alpha (α_2) spricht nichts dagegen, dass die Voraussetzungen der Split-Half-Reliabilität erfüllt sind.
c) Die Reliabilität des Tests beträgt mindestes 0.89 (α_2). Da die Split-Half-Reliabilität den selben Wert hat, spricht nichts dagegen, dass dieser Wert gleich der Reliabilität des Tests ist.

14. Einsetzen von $e^{\theta_v}/(1+e^{\theta_v})$ für $p_{v\bullet}$ in $\ln(p_{v\bullet}/(1-p_{v\bullet}))$ ergibt:

$$\ln\left(\frac{p_{v\bullet}}{1-p_{v\bullet}}\right) = \ln(p_{v\bullet}) - \ln(1-p_{v\bullet})$$

$$= \ln\left(\frac{e^{\theta_v}}{1+e^{\theta_v}}\right) - \ln\left(\frac{1}{1+e^{\theta_v}}\right)$$

$$= \theta_v - \ln(1+e^{\theta_v}) - 0 + \ln(1+e^{\theta_v}) = \theta_v .$$

Umgekehrt ergibt Einsetzen von $\ln(p_{v\bullet}/(1-p_{v\bullet}))$ für θ_v in $e^{\theta_v}/(1+e^{\theta_v})$:

$$\frac{e^{\theta_v}}{1+e^{\theta_v}} = \frac{\frac{p_{v\bullet}}{1-p_{v\bullet}}}{1+\frac{p_{v\bullet}}{1-p_{v\bullet}}}$$

$$= \frac{\frac{p_{v\bullet}}{1-p_{v\bullet}}}{\frac{1-p_{v\bullet}}{1-p_{v\bullet}} + \frac{p_{v\bullet}}{1-p_{v\bullet}}} = p_{v\bullet} .$$

Also gilt $p_{v\bullet} = e^{\theta_v}/(1+e^{\theta_v})$ genau dann, wenn $\theta_v = \ln(p_{v\bullet}/(1-p_{v\bullet}))$.

15. Die 4-Felder-Tafel lautet:

	+	−	
+	42	63	105
−	18	27	45
	60	90	150

16. r = 0.76

17. prob{x_v} = 0.072

1. Die Klassen unterscheiden sich qualitativ. Zu Kapitel 2.3
   (Itemrangreihe = 2 > 1 > 4 > 3; Profillinien sind nicht überschneidungsfrei).
2. Einsetzen von [2.3.2.32] in

$$\text{prob}\{X_{vi} = x \mid x-1 \le X_{vi} \le x\} = \frac{\text{prob}\{X_{vi} = x\}}{\text{prob}\{X_{vi} = x-1\} + \text{prob}\{X_{vi} = x\}}$$

   ergibt

$$\frac{\dfrac{(\xi_v \varepsilon_i)^x}{x!} e^{-\xi_v \varepsilon_i}}{\dfrac{(\xi_v \varepsilon_i)^{x-1}}{(x-1)!} e^{-\xi_v \varepsilon_i} + \dfrac{(\xi_v \varepsilon_i)^x}{x!} e^{-\xi_v \varepsilon_i}}$$

$$= \frac{\dfrac{\xi_v \varepsilon_i}{x}}{1 + \dfrac{\xi_v \varepsilon_i}{x}}$$

$$= \frac{\xi_v \alpha_{ix}}{1 + \xi_v \alpha_{ix}},$$

   worin $\alpha_{ix} = \varepsilon_i x^{-1}$.
3. a) $\gamma_0(\varepsilon) = 1;\ \gamma_1(\varepsilon) = 3.5\ ;\ \gamma_2(\varepsilon) = 3.5\ ;\ \gamma_3(\varepsilon) = 1$
   b) $\text{prob}\{(0,1,1) \mid x_{vo} = 2\} = 0{,}571$
4. a) Aus der Definition [2.3.3.24] folgt

$$\xi_v (1 - f_{1.0}) = f_{1.0}$$
$$\xi_v - \xi_v f_{1.0} = f_{1.0}$$
$$\xi_v = \xi_v f_{1.0} + f_{i.0}$$
$$\xi_v = f_{1.0}(\xi_v + 1)$$

   und daher

$$f_{1.0} = \frac{\xi_v}{\xi_v + 1}.$$

   b) Einsetzen von [2.3.3.24] in die Definition [2.3.3.25] ergibt

$$\sigma_2 = \xi_v \frac{1 - f_{2.0}}{f_{2.0}}$$

   und daher

$$f_{2.0}\, \sigma_2 = \xi_v (1 - f_{2.0})$$
$$f_{2.0}\, \sigma_2 = \xi_v - \xi_v f_{2.0}$$
$$\xi_v f_{2.0} + f_{2.0}\, \sigma_2 = \xi_v$$
$$f_{2.0}(\xi_v + \sigma_2) = \xi_v$$
$$f_{2.0} = \frac{\xi_v}{\xi_v + \sigma_2}.$$

1. $\sigma(\hat{\tau}_{vt}) \cong \sigma(F_{ot}) = 1.35$ Zu Kapitel 2.4
2. a) t = 22.94, df = 29; prob(t ≥ 22.94) < 0.001 < $\alpha/2$ = 0.025. Die

Annahme der t-Äquivalenz muss verworfen werden.
b) Bei Teilung der Stichprobe in Vpn mit $y \leq 1.87$ vs. $y \geq 1.93$ (Medianspilt) ist $t = 1.98$ bei $df = 28$; $prob(t \geq 1.98) > 0.025 = \alpha/2$. Die Annahme der essentiellen t-Äquivalenz kann beibehalten werden.
c) Da die beiden Tests nur essentiell t-äquivalent sind, können sie auch nicht parallel, sondern allenfalls essentiell parallel sein.
d) Vergleich der Maßzahlvarianzen ergibt $t = 0.152$ bei $df = 29$; $prob(t \geq 0.152) > 0.025 = \alpha/2$. Die Annahme der essentiellen Parallelität kann beibehalten werden.
Vergleich der Korrelationen mit der Abiturnote ergibt $z = -1.04$.
$prob(z \leq -1.04) > 0.025 = \alpha/2$. Die Annahme der essentiellen Parallelität kann beibehalten werden.

3. $KONF_{99\%}\{99.54 \leq \tau_{vt} \leq 118.46\}$
4. $z = -0.54$; $prob(z \leq -0.54) > \alpha/2 = 0.025$. Die Annahme, dass die Leistungsfähigkeit der Vp dem Durchschnitt entspricht, kann beibehalten werden.
5. $z = 2.12$; $prob(z \geq 2.12) < \alpha/2 = 0.025$. Die Vpn unterscheiden sich in ihrer Leistungsfähigkeit signifikant voneinander.
6. $n_{10} = 2$; $n = 14$. $prob(X \leq 2) \leq 0.025 = \alpha/2$. Die Testleistung der Probanden ist signifikant verschieden.

Zu Kapitel 2.5

1. a) und c)

g	1	2	3	4	5	6	7	8
$\hat{p}_g$	0.0230	0.0805	0.0690	0.1034	0.1954	0.1724	0.1379	0.2184
$\sigma(\hat{p}_g)$	0.0161	0.0292	0.0272	0.0327	0.0425	0.0405	0.0370	0.0443

b) $n(P_{sat}) = 7$
d) $\ln(L(X)) = -168.4522$

2. a) $h_{max} = 32$
b) $h_{max} = 125$
c) $h_{max} = 4000$

Zu Kapitel 2.6

1. a) und b)

	x = 0	x = 1	x = 2
$\hat{p}_{\bullet\bullet x}$	0.2222	0.4815	0.2963
$\sigma(\hat{p}_{\bullet\bullet x})$	0.0800	0.0962	0.0879

c) $\ln\{L(X)\} = -28.2572$
d) $n(P_{PR}) = 2$

2. a)

	x = 0	x = 1	x = 2
i = 1	0.4444	0.4444	0.1111
i = 2	0.1111	0.6667	0.2222
i = 3	0.1111	0.3333	0.5556

b)

	x = 0	x = 1	x = 2
i = 1	0.1656	0.1656	0.1048
i = 2	0.1048	0.1571	0.1386
i = 3	0.1048	0.1571	0.1656

c) ln{L(X)} = –24.7548
d) $n(P_{LC1}) = 6$
e) Da die Kategorienwahrscheinlichkeiten für alle Items gleich groß angenommen werden, beschreibt das Pure-Random-Modell die Likelihood der Datenmatrix durch eine geringere Anzahl von unabhängigen Modellparametern, deren Schätzung eine geringere Fehlerstreuung aufweist. Andererseits ist das Pure-Random-Modell dadurch jedoch restriktiver als die Ein-Klassen-Lösung der LCA, weshalb es die Daten weniger genau zu beschreiben vermag und die Likelihood der Datenmatrix kleiner ausfällt.

3. a) und b)

	x = 0	x = 1	x = 2
$\hat{p}_{v\bullet x}$	0.2000	0.3500	0.4500
$\sigma(\hat{p}_{v\bullet x})$	0.0894	0.1067	0.1112

c) $\ln\{L(x_v|\theta_v)\} = -20.9731$
d) n(P) = 2

4. a)

x_{vo}	0	1	2	3	4	5
$\hat{\theta}_v$	0	0.2	0.4	0.6	0.8	1
$\sigma(\hat{\theta}_v)$	0	0.1789	0.2191	0.2191	0.1789	0

b)

x_{vo}	0	1	2	3	4	5
$\sigma(\hat{p}_g)$	0.10	0.17	0.28	0.24	0.15	0.06
$\sigma(\hat{p}_g)$	0.0300	0.0376	0.0449	0.0427	0.0357	0.0237

c) $\ln\{L(X)\} = -339.6167$
d) $n(P_{BM}) = 5$

5. a) $n(P_{sat}) = 7$; $n(P_{PR}) = 1$
b) Saturiertes Modell:

g	1	2	3	4	5	6	7	8
x_g	000	001	010	100	011	101	110	111
$\hat{p}_g$	0.025	0.125	0.050	0.050	0.200	0.250	0.200	0.100

Pure Random Modell: $\hat{p}_{\bullet\bullet 0} = 0.3917$; $\hat{p}_{\bullet\bullet 1} = 0.6083$

c) Itemunabhängigkeit: $\chi^2 = 1.329$, df = 2.
prob($\chi^2 \geq 1.329$) > 0.05 = α. Die Annahme der Itemunabhängigkeit der Antwortvariablen kann beibehalten werden.
Personenunabhängigkeit: $\chi^2 = 4.516$, df = 3.
prob($\chi^2 \geq 4.516$) > 0.05 = α. Die Annahme der Personenunabhängigkeit der Antwortvariablen kann beibehalten werden.

Zu Kapitel 2.7

1. $\ln\{uL(X)\} = -7.0228$
2. a) $\ln\{cL(X)\} = -8.6621$
   b) $n(P_{CR}) = 2$
   c) $I(\delta_1) = 2.122,\ I(\delta_2) = 2.216,\ I(\delta_3) = 2.216$
   d) $\sigma(\hat{\delta}_1) = 0.686,\ \sigma(\hat{\delta}_2) = 0.672,\ \sigma(\hat{\delta}_3) = 0.672$
3. a) $\ln\{mL(X)\} = -22.3096$
   b) $n(P_{MR}) = 5$
4. Die Herleitung der Gleichungen ergibt sich wegen der Symmetrie zwischen ξ und ε in direkter Analogie zu den Gleichungen [2.7.2.2], [2.7.2.6] und [2.7.2.10], indem man darin v und i, x_{vo} und x_{oi} sowie ξ und ε miteinander vertauscht.
5. a) $I(\theta_0) = 0.327,\ I(\theta_1) = 0.697,\ I(\theta_2) = 0.697,\ I(\theta_3) = 0.327$
   b) $\sigma(\hat{\theta}_0) = 1.750,\ \sigma(\hat{\theta}_1) = 1.198,\ \sigma(\hat{\theta}_2) = 1.198,\ \sigma(\hat{\theta}_3) = 1.750$
6. Für kleine θ ist $p_{vi} < 0.5$ und daher $(1 - 2p_{vi}) > 0$, so dass die Summanden im Zähler des Korrekturterms positive Werte annehmen. Für große θ ist dagegen $p_{vi} > 0.5$ und daher $(1 - 2p_{vi}) < 0$, so dass die Summanden im Zähler des Korrekturterms negative Werte annehmen.

Zu Kapitel 2.8

1. Einsetzen von [2.3.3.5] in

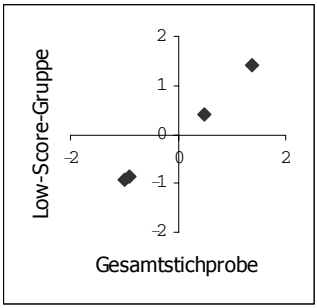
$$\text{prob}(x_{vi} = 1 | x_{vo} = g) = \frac{\text{prob}(x_{vi} = 1)\text{prob}\left(\sum_{j \neq i} x_{vj} = g - 1\right)}{\text{prob}(x_{vo} = g)}$$

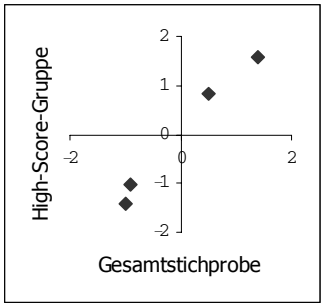
ergibt wegen [2.3.3.8]

$$\text{prob}(x_{vi} = 1 | x_{vo} = g) = \frac{\dfrac{\xi_v \varepsilon_i}{1 + \xi_v \varepsilon_i} \dfrac{\xi_v^{g-1}}{\prod_{j \neq i}(1 + \xi_v \varepsilon_j)} \gamma_{g-1}^{(i)}(\varepsilon)}{\dfrac{\xi_v^{g}}{\prod_{i}(1 + \xi_v \varepsilon_i)} \gamma_g(\varepsilon)}$$

$$= \frac{\varepsilon_i \gamma_{g-1}^{(i)}(\varepsilon)}{\gamma_g(\varepsilon)}.$$

2. a) In der High-Score-Gruppe zeigt sich eine deutliche Abweichung von der 45°-Geraden: Leichte Items sind in der High-Score-Gruppe noch leichter und schwierige Items noch schwieriger als in der Gesamtstichprobe. Die spezifische Objektivität scheint verletzt zu sein.





b) $-2\ln(\lambda) = 4.097$; df = 3; $\text{prob}(\chi^2 \geq 4.097) > 0.05$. Das Rasch-Modell kann trotz der Abweichungen von der 45°-Geraden beibehalten werden.

3. $\chi_o^2 = 5.028$; df = 6; $\text{prob}(\chi^2 \geq 5.028) > 0.05$. Das Rasch-Modell kann beibehalten werden.
4. $\chi_o^2 = 24.661$; df = 9; $\text{prob}(\chi^2 \geq 24.661) < 0.005$. Das Rasch-Modell muss verworfen werden. Insbesondere in der Gruppe jener Vpn, die nur ein Item lösen (und somit noch relativ stark in traditionellen Geschlechtsrollenvorstellungen verhaftet sind), unterscheiden sich männliche und weibliche Vpn sehr deutlich in ihrem Antwortverhalten ($\chi_1^2 = 20.100$; df = 3, $\text{prob}(\chi^2 \geq 20.100) < 0.001$).
5. $\omega_1^2 = 14.37$; $\omega_2^2 = 11.05$. (Wie man mittels Bootstrapping zeigen kann, sind beide Werte signifikant: $\text{prob}(\omega_1^2 \geq 14.37) = 0.018$; $\text{prob}(\omega_2^2 \geq 11.05) = 0.003$. Die Annahme der lokalen Unabhängigkeit muss verworfen werden).
6. a) $\text{KONF}_{95\%}\{2.515 \leq \theta_v \leq 6.125\}$
   b) z = 3.724; $\text{prob}(Z \geq 3.724) < 0.001 < \alpha/2 = 0.005$. Die Leistungsfähigkeit der Vp ist signifikant überdurchschnittlich.
7. z = 2.697; $\text{prob}(Z \geq 2.697) < 0.005 = \alpha/2$. Die Leistungsfähigkeit der Vpn ist signifikant verschieden.
8. Einsetzen von [2.3.3.6] in

$$\text{prob}\{(1,0)|x_{oi} = 1\} = \frac{\text{prob}\{(1,0)\}}{\text{prob}\{(1,0) + \text{prob}(0,1)\}}$$

ergibt

$$\text{prob}\{(1,0)|x_{oi} = 1\} = \frac{\dfrac{\xi_v\varepsilon_i}{1+\xi_v\varepsilon_i}\dfrac{1}{1+\xi_w\varepsilon_i}}{\dfrac{\xi_v\varepsilon_i}{1+\xi_v\varepsilon_i}\dfrac{1}{1+\xi_w\varepsilon_i} + \dfrac{1}{1+\xi_v\varepsilon_i}\dfrac{\xi_w\varepsilon_i}{1+\xi_w\varepsilon_i}}$$

$$= \frac{\xi_v}{\xi_v + \xi_w}.$$

Zu Kapitel 2.9

1. a) Gewichtszahlenmatrix und Effektparameter

Versuchsplan			Gewichtszahlen $g_{ab}^{(j)}$						$\ln(\beta_j)$
j	F1	F2	a = 1			a = 2			
			b_1 =			b_2 =			
			0	1	2	0	1	2	
1	0	0	1	0	0	1	0	0	$\omega_0^{(1)} + \omega_0^{(2)}$
2	0	1	1	0	0	0	1	0	$\omega_0^{(1)} + \omega_1^{(2)}$
3	0	2	1	0	0	0	0	1	$\omega_0^{(1)} + \omega_2^{(2)}$
4	1	0	0	1	0	1	0	0	$\omega_1^{(1)} + \omega_0^{(2)}$
5	1	1	0	1	0	0	1	0	$\omega_1^{(1)} + \omega_1^{(2)}$
6	1	2	0	1	0	0	0	1	$\omega_1^{(1)} + \omega_2^{(2)}$
7	2	0	0	0	1	1	0	0	$\omega_2^{(1)} + \omega_0^{(2)}$
8	2	1	0	0	1	0	1	0	$\omega_2^{(1)} + \omega_1^{(2)}$
9	2	2	0	0	1	0	0	1	$\omega_2^{(1)} + \omega_2^{(2)}$

b) 4 unabhängige Effektparameter

2. a) Gewichtszahlenmatrix und Effektparameter

Versuchsplan			Gewichtszahlen $g_{a*b*}^{(j)}$				$\ln(\beta_j)$
j	Faktoren		b* = (b,b')				
	F1=a	F2=a'	(0,0)	(0,1)	(1,0)	(1,1)	
1	0	0	1	0	0	0	$\omega_{0,0}^{(a,a')}$
2	0	1	0	1	0	0	$\omega_{0,1}^{(a,a')}$
3	1	0	0	0	1	0	$\omega_{1,0}^{(a,a')}$
4	1	1	0	0	0	1	$\omega_{1,1}^{(a,a')}$

b) 3 unabhängige Effektparameter

3. a) Haupteffekte

	0	1	2	3
a	-1.400	1.075	0.300	-0.075
a'	0.975	-1.275	-0.625	0.825

b) Interaktionseffekte

	b' = 0	b' = 1	b' = 2	b' = 3
b = 0	0.425	-0.125	0.025	-0.225
b = 1	0.250	0.600	0.250	-1.000
b = 2	-0.175	0.675	0.425	-0.825
b = 3	-0.400	-1.050	-0.600	2.150

4. a) Haupteffekte und gemeinsame Effekte: Gewichtszahlenmatrix

Versuchsplan				Gewichtszahlen $g_{ab}^{(j)}$								$\ln(\beta_j)$
j	F1	F2	F3	a = 1			a = 2			a = 3		
				b_1 =			b_2 =			b_3 =		
				0	1	2	0	1	2	1	2	
1	0	0	–	1	0	0	1	0	0	0	0	$\omega_0^{(1)} + \omega_0^{(2)}$
2	0	1	–	1	0	0	0	1	0	0	0	$\omega_0^{(1)} + \omega_1^{(2)}$
3	0	2	–	1	0	0	0	0	1	0	0	$\omega_0^{(1)} + \omega_2^{(2)}$
4	1	0	–	0	1	0	1	0	0	0	0	$\omega_1^{(1)} + \omega_0^{(2)}$
5	1	1	–	0	1	0	0	1	0	0	0	$\omega_1^{(1)} + \omega_1^{(2)}$
6	(1,2)		1	0	0	0	0	0	0	1	0	$\omega_1^{(3)} = \omega_{1,2}^{(1,2)}$
7	2	0	–	0	0	1	1	0	0	0	0	$\omega_2^{(1)} + \omega_0^{(2)}$
8	(2,1)		2	0	0	0	0	0	0	0	1	$\omega_2^{(3)} = \omega_{2,1}^{(1,2)}$
9	2	2	–	0	0	1	0	0	1	0	0	$\omega_2^{(1)} + \omega_2^{(2)}$

b) Haupteffekte und Wechselwirkungen: Gewichtszahlenmatrix

Versuchsplan				Gewichtszahlen $g_{ab}^{(j)}$								$\ln(\beta_j)$
j	F1	F2	F3	a = 1			a = 2			a = 3		
				b_1 =			b_2 =			b_3 =		
				0	1	2	0	1	2	1	2	
1	0	0	–	1	0	0	1	0	0	0	0	$\omega_0^{(1)} + \omega_0^{(2)}$
2	0	1	–	1	0	0	0	1	0	0	0	$\omega_0^{(1)} + \omega_1^{(2)}$
3	0	2	–	1	0	0	0	0	1	0	0	$\omega_0^{(1)} + \omega_2^{(2)}$
4	1	0	–	0	1	0	1	0	0	0	0	$\omega_1^{(1)} + \omega_0^{(2)}$
5	1	1	–	0	1	0	0	1	0	0	0	$\omega_1^{(1)} + \omega_1^{(2)}$
6	(1,2)		1	0	1	0	0	0	1	1	0	$\omega_1^{(1)} + \omega_2^{(2)} + \omega_1^{(3)} =$ $\omega_1^{(1)} + \omega_2^{(2)} + \kappa_{1,2}^{(1,2)}$
7	2	0	–	0	0	1	1	0	0	0	0	$\omega_2^{(1)} + \omega_0^{(2)}$
8	(2,1)		2	0	0	1	0	1	0	0	1	$\omega_2^{(1)} + \omega_1^{(2)} + \omega_2^{(3)} =$ $\omega_2^{(1)} + \omega_1^{(2)} + \kappa_{2,1}^{(1,2)}$
9	2	2	–	0	0	1	0	0	1	0	0	$\omega_2^{(1)} + \omega_2^{(2)}$

c) Jeweils 4 Haupteffekte + 2 gemeinsame bzw. Interaktionseffekte.

5. Einsetzen von [2.9.2.1] in [2.7.2.7] und Nullsetzen der ersten partiellen Ableitungen nach den Basisparametern ergibt für j = 1,…,m:

$$\frac{\partial \ln\{cL(X)\}}{\partial \eta_j} = \sum_{i=1}^{k} \frac{\partial \ln\{cL(X)\}}{\partial \varepsilon_i} \frac{\partial \varepsilon_i}{\partial \eta_j}$$

$$= \sum_{i=1}^{k} \left\{ \frac{x_{oi}}{\varepsilon_i} - \sum_{g=0}^{k} \frac{\gamma_{g-1}^{(i)}(\varepsilon)}{\gamma_g(\varepsilon)} \right\} q_{ij} = 0 .$$

Dieser Ausdruck ist der Gleichung [2.9.2.3] äquivalent.

6. a) Designmatrix:

	η_1	η_2
i = 1	1	0
i = 2	0	1
i = 3	0	1
i = 4	1	0

b) $-2\ln(\lambda) = 42.0764$; df = 2; $p < 0.001$. Die Hypothese muss verworfen werden.

7. Soll x_{vo} eine erschöpfende Statistik für θ_v sein, dann muss die bedingte Wahrscheinlichkeit

$$\text{prob}\{(1,0)|x_{vo}\} = \frac{\text{prob}\{(1,0)\}}{\text{prob}\{(1,0)\} + \text{prob}\{(0,1)\}}$$
$$= \frac{f_1(1-f_2)}{f_1(1-f_2) + (1-f_1)f_2}$$

von θ_v unabhängig sein. Setzt man für die Itemcharakteristiken f_i den Ausdruck [2.9.5.1] ein, dann ergibt sich aber

$$\text{prob}\{(1,0)|x_{vo}\} = \frac{1-p_2}{1 + \frac{1-p_1}{1-p_2} \frac{p_2 + \exp(\theta_v - \delta_2)}{p_1 + \exp(\theta_v - \delta_1)}} .$$

Daraus lässt sich θ_v nur dann wegkürzen, wenn die Ratewahrscheinlichkeiten $p_1 = p_2 = 0$ sind.

8. Soll x_{vo} eine erschöpfende Statistik für θ_v sein, dann muss die bedingte Wahrscheinlichkeit

$$\text{prob}\{(1,0)|x_{vo}\} = \frac{\text{prob}\{(1,0)\}}{\text{prob}\{(1,0)\} + \text{prob}\{(0,1)\}}$$
$$= \frac{f_{1.0}(1-f_{2.1})}{f_{1.0}(1-f_{2.1}) + (1-f_{1.0})f_{2.0}}$$

von θ_v unabhängig sein. Setzt man für die bedingten Itemcharakteristiken $f_{i.r}$ den Ausdruck [2.9.5.2] ein, dann ergibt sich aber

$$\text{prob}\{(1,0)|x_{vo}\} = \frac{1}{1 + \frac{\exp(-\delta_2 + \beta_{20})}{\exp(-\delta_1 + \beta_{10})} \frac{1 + \exp(\theta_v - \delta_2 + \beta_{21})}{1 + \exp(\theta_v - \delta_2 + \beta_{20})}} .$$

Daraus lässt sich θ_v nur dann wegkürzen, wenn $\beta_{21} = \beta_{20}$ ist und also kein erfolgskontingentes Lernen stattfindet.

Zu Kapitel 2.10

1. a) $p_{1.x_v} = 0.92364$; $p_{2.x_v} = 0.07636$

b) Die Vp kann mit einer Sicherheit von 92.364% der Klasse 1 zugeordnet werden.

2. Fehlerstreuung der Parameterschätzungen:

	$\sigma(\hat{p}_g)$	$\sigma(\hat{p}_{g1})$	$\sigma(\hat{p}_{g2})$	$\sigma(\hat{p}_{g3})$
g = 1	0.023	0.018	0.018	0.024
g = 2	0.023	0.046	0.046	0.046

3. $n(P_{LC2}) = 7$
4. df = 0
5. Lokales Maximum bei h = 3, weil ln{mL(X)} für h = 3 kleiner ist als für h = 2.
6. a) h = 3 Klassen.
   b) $-2\ln(\lambda) = 13.978$, df = 14, n.s. Die Daten werden durch 3 latente Klassen ebenso gut beschrieben wie durch das saturierte Modell.
   c) $-2\ln(\lambda) = 292.03$, df = 12, $p < 0.001$. Die Daten werden durch 3 latente Klassen signifikant besser beschrieben.

1. a) Anzahl der unabhängigen Modellparameter und Informationsmaße: Zu Kapitel 2.11

Modell	$\ln\{mL(X)\}$	n(P)	AIC	BIC	CIC
Pure Random	-401.9131	1	805.8262	808.8097	805.4323
Identische Personen	-396.7988	4	801.5976	813.5319	800.0221
Identische Items	-384.1581	4	776.3162	788.2507	774.7409
LCA (h=3 Klassen)	-372.9344	14	773.8688	815.6393	768.3549
Rasch-Modell	-377.5100	7	769.0200	789.9053	766.2631
Saturiertes Modell	-371.4418	15	772.8836	817.6378	766.9760

   b) Während AIC und CIC das Rasch-Modell favorisieren, spricht der BIC zugunsten der Annahme identischer Items (Binomialmodell).
   c) Vergleich von Binomialmodell (H_0) und Rasch-Modell (H_1) mittels Likelihood-Quotienten-Tests ergibt:
   $-2\ln(\lambda) = 13.2962$, df = 3, $p < 0.005$. Das Binomialmodell wird zugunsten des Rasch-Modells verworfen.

1. Es ist mit einer verbleibenden Validität von $\rho(X_{ot}, Y_{ot'}) = 0.643$ zu rechnen. Zu Kapitel 2.12.
2. Da sowohl $\rho(X_{ot}, T_{ot})$ als auch $\rho(Y_{ot'}, T_{ot'})$ nichtnegative Zahlen sind, folgt aus Gleichung [2.2.2.3] der Ausdruck
   $|\rho(X_{ot}, Y_{ot'})| = \rho(X_{ot}, T_{ot})\,|\rho(T_{ot}, T_{ot'})|\,\rho(Y_{ot'}, T_{ot'})$.
   Die Ungleichung [2.12.1.5] ergibt sich daraus, weil $\rho(X_{ot}, T_{ot}) \leq 1$ ist.
3. Die Verdünnungsformel [2.12.1.6] ergibt sich, indem man die linke und die rechte Seite der Gleichung [2.2.2.3] durch den Ausdruck $\sqrt{\rho^2_{XT}\,\rho^2_{YT'}}$ dividiert.
4. Einsetzen in die Verdünnungsformel [2.12.1.6] ergibt mit $\rho(T_{ot}, T_{ot'}) = 1.02$ einen Wert, der die fiktive Korrelation nicht nur überschätzt, sondern größer als Eins ist und daher als Korrelation gar nicht in Frage kommt.
5. Die Gleichung [2.12.1.7] ergibt sich, indem man die linke und die

rechte Seite der Gleichung [2.12.1.6] mit dem Ausdruck $\sqrt{\rho^2_{XT}\,\rho^2_{YT'}}$ multipliziert.

6. Die tatsächliche (empirische) Korrelation der beiden Tests beträgt lediglich $\rho(X_{ot}, Y_{ot'}) = 0.383$.

Sachregister

Ablaufschema
- der qualitativen Inhaltsanalyse 80f.
- der quantitativen Inhaltsanalyse 112ff.
- der Textinterpretation 46f.
Abhängigkeit, serielle 193, 212, 227, 282, 307, 310
AIC-Index 322ff., 331
Analyse
- explorative 81f., 86, 89
- inspektive 81, 86, 101
Ankerbeispiel 90-94, 103, 107, 116
Antwortkategorie 141, 248, 254, 291, 365
Antwortmatrix 141f., 243, 247, 249, 252, 259f., 262, 266ff., 274, 277, 279, 282, 308, 311, 365, 367
- bedingte 283f.
Antwortmuster siehe Antwortvektor
Antwortvariable 145, 181, 199-205, 212, 218, 224, 227, 290, 305, 307, 318, 366
- itemunabhängige 204, 247-255, 366f.
- ordinale 216, 310
- personenunabhängige 206, 247-255, 366f.
- polytome 215, 310
Antwortvektor 141f., 190, 203, 206, 208ff., 212, 218-230, 240, 243ff., 250f., 258f., 266, 269f., 278, 285, 307, 309, 311f., 317, 332, 365, 367
Appellfunktion (der Sprache) siehe parade aspect
Augenscheinvalidität 334
Aufgabenlösung
- trait-bedingte 213f.
- transfer-bedingte 213f.
Ausdrucksfunktion (der Sprache) siehe command aspect
Axiome von Gulliksen 149ff., 159, 199, 203

Basisparameter 297-301
Bayes-Theorem 270
Beobachtung, teilnehmende 20
BIC-Index 323f., 331
Binomialmodell siehe Identische Items
Binomialverteilung 209, 240f., 371, 373
Birnbaum-Modell 185f., 212, 262, 310ff., 366
Bootstrap-Verfahren 282, 284, 332
Bruch, thematischer 68

CIC-Index 323f., 331
Codematrix 97, 99
Codemuster 97
Command aspect 38ff., 42f., 48f., 52-60, 76, 82, 130-132
Correction for attenuation 336f.

Darstellungsfunktion (der Sprache) siehe report aspect
Datenaggregation 76, 79, 281
Datensimulation 282, 284
Designmatrix 297-305, 309, 365
- erweiterte 300, 365

- reduzierte 304f., 365
Dichtefunktion siehe Wahrscheinlichkeitsdichte
Differenz, kritische (der Testleistungen zweier Vpn) 239, 285
Diskursanalyse 31, 42
Distanz (zwischen Forscher und Forschungssubjekt) 23

Echtheit 25
Effektparameter 291-294, 301-305, 367
Eigenschaftszuschreibung 86-93
Ein-Klassen-Lösung der LCA siehe Identische Personen
EM-Algorithmus 319, 324
Entfremdung 40, 71
Erhebungsinstrument 44f.
- qualitatives 19f.
- standardisiertes 16, 32
Erwartungswert 143, 146, 150f., 199, 203f., 234f., 265, 281, 319
Experiment 9, 16, 19, 293

Faktorenanalyse 343, 346ff.
Fehlerstreuung 125, 159ff., 244, 248-252, 264, 268, 285, 319, 368
Fehlervariable 150, 160, 366
Fragen
- exmanente 22
- immanente 22
Form, thematische 62
Forschung, interpretative 18ff.
Freiheitsgrade 236f., 254, 278, 280f., 299f., 321, 324, 326, 332
F-Test 237

Gehalt
- exterritorialisierter 39, 41, 67, 71
- latenter 39, 41, 48, 66f., 71
Gespräch, partnerzentriertes 22f., 44f., 68, 70
Gesamtstichprobe 129, 191f., 276, 278
- reduzierte 278f.
Gesprächsführung 20, 26f., 31
- klientenzentrierte 23, 25f.
Gewinn, sekundärer 39, 41, 67f., 71
Grounded Theory 31
Grundhaltung, personenzentrierte 25
Grundfunktion, symmetrische 219, 226, 263, 269
Gruppendiskussion 20, 27, 35, 37, 45, 48, 68, 70
Guttman-Skala 177ff., 185

Handlungsanalyse 44ff., 70
Handlungsbeschreibung 21, 86-92, 115f.
Handlungsprämisse 37, 41, 44f., 70
Hauptkomponentenanalyse 341, 343, 345ff.
Häufigkeit
- absolute 253
- erwartete 118, 253, 279-284, 319f., 332, 366
- relative 243, 248-252, 268, 282, 318
Hermeneutik, objektive 31
Heterogenität 191
Homogenität 159, 190ff., 207, 313
Homomorphie 177, 179

Indikator 74, 119, 131, 180f.
Information, statistische 207, 209
Informationsfunktion 264, 268, 270f., 285, 368
Informationsmaß 323f., 331
Inhaltsanalyse
- qualitative 11, 31, 35, 38, 43, 45, 76-86, 96, 116, 120, 123-135
- quantitative 11, 31, 35, 38, 43, 75-86, 96, 112, 116, 119-135
Inhomogenität 192f., 207

Intelligenzalter 139f., 144f.
Intelligenzquotient (IQ) 140, 143-146
Intelligenz-Rückstand 140
Intelligenz-Vorsprung 139
Interaktion 15, 19f., 41, 45, 69
Interaktionsanalyse 44ff., 68ff.
Intercoder-Reliabilität 43, 79, 113, 116ff., 131
Interpretation, geometrische (des Korrelationskoeffizienten) 339
Interpretationsfolie 45f., 48, 81, 83
Intersubjektivität 34
Interview
- narratives 21ff., 25, 44f., 68, 70
- qualitatives 20, 26f., 70
Interviewleitfaden 21, 26f.
Interviewtechniken 20,27, 45
Items, identische 183-187, 202, 204f., 209-212, 220, 250ff., 258, 298f., 367f.
Itemcharakteristik 181-187, 209-213, 216, 218, 223, 227ff., 258, 261, 274f., 306ff., 367
- bedingte 213, 308, 367
Itemparameter 182-186, 211, 219f., 226f., 258-269, 274-277, 283, 286, 291, 297, 301-305, 311-314, 366ff.
Itemrandvektor 142
Itemunabhängigkeit 183, 186, 200, 204, 247f., 254, 366ff.

Kategorienbildung
- materialbezogene 86
- theoriebezogene 101
Kategorienschwierigkeit 216, 367
Kategorienzahl 247f., 250, 255, 321
Klassen
- latente 120, 189, 202, 206f., 314, 318-328
- manifeste 210f., 318
- qualitativ verschiedene 208
- quantitativ verschiedene 206f.
Klassifikationssystem 82
Kodiereinheit 82f., 97, 113, 115f.
Kodierregel siehe Prädikatorenregel
Kodierschema 79, 82f., 86, 96-101, 103, 107, 115ff., 124, 132
Koeffizient Alpha 167, 170-176, 368
Koeffizient Kappa 117, 370
Kommunalitätenproblem 347
Kommunikation, egalitäre 24
Konfidenzintervall 238f., 285
Konsistenz 34, 68
Konsistenzanalyse, interne 167, 175
Konstruktvalidität 334, 338
Kontext
- externer 34, 41f., 46, 49, 52f., 55, 57, 59, 61f., 69, 71, 79f., 83, 110, 130
- interner 34, 42, 53, 61, 69, 110
Kontingenzanalyse 74
Kontingenztafel 117f., 253f., 327
Konversationsanalyse 31
Konzentrationsleistungstest 217
Korrelation 119, 152-158, 163, 166f., 172, 191ff., 235, 238, 336-346
Kovarianz 155f., 159, 170, 238, 335, 342, 347
Kriteriumsvalidität 334f., 338
Kriteriumsvariable 152, 238, 327f., 335f., 366
Kuder-Richardson-Formel 168, 175

Ladungsmatrix 342ff., 347
Latent-Class-Analyse 11, 119f., 124, 126, 187, 189, 202, 205-212, 248, 317-328, 330ff., 368
Latent-Trait-Modell 12, 187, 193, 201, 205, 258-272, 318, 330
Lerneffekt 213f., 307
Lernprozess, erfolgskontingenter 193, 214, 305

Lesefehler 217
Likelihood
- bedingte 261ff., 267f., 277f., 282, 299, 304, 309, 367f.
- marginale 265ff., 312, 317, 320, 322, 324, 331, 367f.
- totale siehe unbedingte
- unbedingte 258, 260f., 367
Likelihood-Quotienten-Test 277f., 299f., 304, 331, 348
Likert-Skala 77
Lösungsstrategie 313f.
Lösungswahrscheinlichkeit 168, 181ff., 207ff., 221, 223, 366
- bedingte 213, 282, 366

Makroanalyse 49, 53, 60f.
Makrokontext 42
Maßzahltransformation 179, 181, 186
Maßzahlvariable 149ff., 171, 338, 366
Maßzahlvarianz 160, 204, 235, 237
Markov-Prozess 310
Maximum-Likelihood-Methode 243, 248f., 251f., 258-272, 275, 311, 318, 348
- unbedingte (unconditional); UML-Methode 261, 275
- bedingte (conditional); CML-Methode 261-269, 275f., 299, 305, 310
- gewichtete (weighted); WML-Methode 269f.
- marginale (marginal); MML-Methode 266-269, 275
Mediansplit 278
Membership-Wahrscheinlichkeit 168, 181ff., 207ff., 221, 223, 366
Merkmalsmuster 119f.
Messfehler 149-152, 155ff., 199, 206, 337, 346
Messgenauigkeit 143, 153ff., 159ff., 171f., 174
Messwert siehe Testscore
- wahrer siehe True-Score
Mikroanalyse 49, 53, 61
Mikrokontext 42
Mischverteilungs-Rasch-Modell 312f.
Missing Data 302, 308
Modell, saturiertes 190, 203-206, 243f., 321f., 331, 368
Modellannahmen von Novick 150f., 203, 234
Modellkontrolle 234, 243f., 247, 332
- graphische 276f.
- parameterfreie 279, 281
Modellparameter, unabhängige 244, 247f., 250, 253, 262, 267, 320ff., 326f., 331f., 367
Modellspezifikation 200, 203, 330
Momentenmethode 234
Motivzuschreibung 88f., 92f.

Neyman-Scott-Problem 261, 265, 269, 275, 318
Normal-Ogive-Modell 184
Normalverteilung 143ff., 184, 239, 285, 365, 374

Objektivität, spezifische 275f., 292, 305, 313f.
Objektivitätssicherung 32
Ordnungsrelation 177f., 181, 191, 206, 208

Parade aspect 38-43, 48, 52-60, 76, 82, 130f., 133
Parallelität 153, 166f., 234-238
Paralleltest-Methode 153ff., 166, 175f.
Parameter
- inzidentelle 261, 265f.
- strukturelle 261
Partial-Credit-Modell siehe ordinales Rasch-Modell
Pearson-χ^2 253ff., 278, 280f., 284, 300, 305, 332, 376

Personen, identische 190, 201, 204, 206, 248, 331, 368
Personenparameter 182, 187, 209, 211, 213, 218-229, 251, 258-262, 265-271, 274, 285f., 292, 304f., 307, 310ff., 319, 367f.
Personenrandvektor 142
Personenunabhängigkeit 253f.
Perspektive
- empfängerbezogene siehe parade aspect
- gegenstandsbezogene siehe report aspect
- senderbezogene siehe command aspect
Poisson-Modell 217f., 290
Poisson-Verteilung 217
Populationsabhängigkeit 161
Prädikatorenregel 77f., 87, 90, 92, 94, 103, 124
Prinzip, methodologisches 11, 16f., 20, 31, 34, 68
Prinzip der Fremdheit 31
Prinzip der Kommunikation 20
Prinzip der Offenheit 20
Prinzip der Reflexion 31
Profillinie 207f., 210, 212
Psychologie
- interpretative 18f.
- nomologische 18f.
Punktschätzung 238
Pure-Random-Modell 200, 204, 206, 247f.,368

Rangordnung 177, 192
Rasch-Modell 186, 210ff., 298f., 368
- dichotomes 215f., 218, 290
- mehrfaktorielles 290ff., 300
- ordinales 216
- polytomes 215, 290
Ratewahrscheinlichkeit 305f.
Realität, subjektive 21, 23, 25f.
Referenzpopulation 146, 149f., 159f., 176, 235, 254, 366
Rekonstruktion
- genetische 20, 44
- sozialpsychologische 10, 35, 37, 41f., 70f.
Rekontextualisierung 76
Reliabilität 143, 152-167, 170-176, 235, 335-338, 368
Report aspect 38-43, 48f., 52-60, 82, 130, 133
Reproduzierbarkeit
- der Itemlösungshäufigkeiten 279
- der Patternhäufigkeiten 332
Retest-Reliabilität 155, 166, 175
Rollenspiel 20, 35, 45, 68
Rollenverteilung 24
Rollenverwischung 24
Rückkoppelungsgespräch, argumentatives 72

Schätzfehler siehe Fehlerstreuung
Schätzfunktion
- erwartungstreue 234, 263, 304
- konsistente 261, 275
- suffiziente siehe Suffizienz
- wirksamste siehe Wirksamkeit
Schwellenparameter 216, 218, 366
Schwellenwahrscheinlichkeit 216, 218
Score siehe Testscore
Scoregruppe 278-281
Score-Variable siehe Maßzahlvariable
Scorevektor 215
Selbstkongruenz 25
Separierbarkeit (der Parameter) 274f., 292, 310ff.
Sequenzanalyse 48-60, 116, 134
Sinnkonstitution 20, 33, 70
Skala 96, 140f., 177-186, 202, 207, 214ff., 262, 297, 300
Skalentyp 179ff., 214
Spearman-Brown-Formel 162-166, 170-175
Split-Half Reliabilität 167, 173-176
Split-n-Reliabilität 167, 175
Sprachnormierung 125
Sprecherwechsel 69, 116

Standardmessfehler 160f., 234, 239, 368
Statistik, erschöpfende siehe Suffizienz
Stichprobe
- abhängige 236f.
- repräsentative 38, 113, 146
- systematische 113
- unabhängige 237, 304
Stichprobenunabhängigkeit 276
Stichprobenziehung 113
Störbedingung 151f.
Störeinfluss 151
Störvariable 365
Streuungsdiagramm 276
Subpopulation 234f., 254, 277
Subtest 166f., 170-176, 234f., 312f., 338, 344, 368
Suffizienz 209, 218-227, 252, 262, 265f., 274, 282, 307, 310ff.
Summenscore 141f., 145f., 203, 206, 209ff., 218-229, 252, 274f., 282f., 307, 309f., 366
- gewichteter 141, 212, 262, 310, 312, 366
- partieller 213, 366
Syndrom 187, 189
Symptom 141, 187ff., 334
Symptomcheckliste 141

Teilstichprobe 86, 116ff., 123, 192, 237, 276-280, 302f.
Test, zusammengesetzter 161-164, 171, 335
Test von McNemar 239f., 286
Testitem
- binäres 175, 218, 224
- dichotomes 140, 168, 199, 202, 221, 223, 227, 229, 251
- polytomes 140
- virtuelles 307ff.
Testkonstruktion 193, 330
Testscore 140-143, 149, 155f., 161, 366
Testteilungs-Methoden 166, 171-174
Testtheorie 141f.
- klassische 11f., 149-152, 161, 176, 199, 203, 234, 238, 285, 346, 365, 368
- probabilistische siehe stochastische
- stochastische 11, 182, 190, 193, 199
Testmodell
- dynamisches (DTM) 212ff., 224f., 227ff., 282f., 310, 365
- lineares logistisches (LLTM) 296-309, 365
Testtheoretisches Modell, allgemeines 11, 199-204
Textgenre 31, 113
Textinterpretation 31, 35, 38, 44-72, 80f., 120
Thematisierungsabwehr 39, 41, 66ff., 71
Themenverschiebung 68
Transferparameter 213f., 226f., 283, 367
Transformation, lineare 145f., 153, 179, 338, 343, 347
Transkript 28f.
Transsubjektivität 32, 79
Triangulation 10, 16f.
True-Score 149-153, 156-160, 166, 204f., 208, 234, 238, 285, 336, 367
True-Score-Variable 149f., 153f., 159, 171, 366
t-Test
- für abhängige Stichproben 236f.
- für unabhängige Stichproben 237
t-Verteilung 236f., 374
τ-Äquivalenz 154, 172, 234-237

Unabhängigkeit, lokale stochastische 124, 190-193, 199, 220, 274, 282

Validierung 16f., 28, 32, 72
Validität 70, 119, 143, 152-158, 334-339, 368

- theoretische 334f.
Variable
- binäre 97, 99f., 124
- dichotome 99f., 177, 181, 324
- inhaltsanalytische 77, 81
- latente 10, 180, 187, 201, 337, 346f.
- manifeste 180, 189, 343, 346f.
- polytome 99ff., 177
- quantitative 78, 177
- qualitative 10, 78, 119, 177, 181, 327
Varianzanalyse 291-295, 304
Varianzanteil 159, 343ff., 347
Verdrängung 41, 71
Verdünnungsformel siehe Correction for attenuation
Vergleichsnorm 143
Vergleichspopulation siehe Referenzpopulation
Versuchsbedingung 290-296, 300-305, 365
Versuchsplan 32, 291-296, 300-304
Verstehen
- einfühlendes 25
- logisches 39f.
- psycho-logisches 39f.
- sozio-logisches 39, 41
Verteilung, logistische 184
Verteilungsfunktion 145, 184, 240
Vertrauensbereich siehe Konfidenzintervall
Vierfeldertafel 191f., 240
Vorbegriff 34f., 45ff., 81, 86, 101

Wärme, emotionale 25
Wahrscheinlichkeitsdichte 199-202, 209, 215, 217, 264f., 267, 270f., 290, 305, 367
Wechselwirkung siehe Interaktion
Wertschätzung 25f.
Wirklichkeitskonstruktion, Regeln der 66f., 71
Wirksamkeit 234, 263

Zentraler Grenzwertsatz 145
Zirkel, hermeneutischer 33, 42, 61, 72, 80, 82f.
Z-Transformation 238
Zufallsfehler 151, 276f.
Zufallsvariation, reine siehe Pure-Random-Modell
Zusatzannahmen (der klassischen Testtheorie) 234ff.

regener
publishers off mainstream

verlag irena regener berlin Wir setzen Qualitätsstandards jenseits des Mainstreams

Wilhelm Kempf

Forschungsmethoden der Psychologie

Zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik

Band I: Theorie und Empirie

(Hochschullehrbücher, Bd. 2). 2006. 2., überarbeitete und erweiterte Auflage, 377 S., brosch., mit 71 Abbildungen und 34 Tabellen, € 42,90. ISBN 978-3-936014-04-9.

Herausgerissen aus dem Gesamtzusammenhang der Psychologie und reduziert auf Experimentiertechniken und Statistik kann der Stellenwert der Methodenlehre für die Konstitution der Psychologie als einer empirischen Wissenschaft nicht richtig erkannt werden.
Die vorliegende Einführung geht einen anderen Weg. In konsequenter Auffassung der Psychologie als Wissenschaft von der subjektiven Welt des Menschen, deren Gegenstandsbereich von den physiologischen Ursachen bis hin zu den Prozessen der sozialen Konstruktion von Wirklichkeit reicht, verfolgt der Autor eine wissenschaftstheoretische und didaktische Konzeption, welche die Methoden der Psychologie im Spannungsfeld zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik verortet und eine gegenstandsangemessene Umsetzung des galileischen Denkens in der Psychologie zum Ziel hat.

Aus dem Inhalt
Kapitel I: Psychologie als Erfahrungswissenschaft: Die Anfänge der wissenschaftlichen Psychologie / Wissenschaftliche Psychologie und psychologisches Alltagswissen / Propädeutische Grundlagen / Aufgabenverständnis, Gegenstandsverständnis und Wissensideal der Psychologie
Kapitel II: Hard Science – Das Erbe der Naturwissenschaften: Aristotelisches und galileisches Wissensideal / Deduktiv-nomologische und statistisch-induktive Erklärung / Messung und experimentelle Hypothesenprüfung
Kapitel III: Soft Science – Die Herausforderung an die Sozialwissenschaften: Intentionale und narrative Erklärung / Sinnrationalität, Systemtheorie und Informationsverarbeitungsmodelle / Strukturelle und empirische Theorieanteile / Handlungstheoretische Prozessanalyse / Qualitative Forschungsmethoden

verlag irena regener berlin
ostseestr. 109
d – 10409 berlin

fax: #49.30.42 85 67 62
www.regener-online.de
eMail: verlag@regener-online.de