Empirische Untersuchung zur Erfassung von Wissen durch deterministische und probabilistische Wissensstrukturen

Florian Wickelmaier

Diplomarbeit als Teil der Diplomprüfung für Psychologen

Universität Regensburg

5. Februar 2002

Inhaltsverzeichnis

1 Einleitung

1.1 Grundlagen der Theorie der Wissensräume

Das Ziel der Theorie der Wissensräume von Doignon & Falmagne (1985, 1999) ist, ein adaptives Diagnoseverfahren zur Verfügung zu stellen, das den Erfordernissen computergestützten Unterrichts gerecht wird. Durch die Auswahl weniger, effizient gestellter Fragen soll der Ablauf einer mündlichen Prüfung durch einen erfahrenen Prüfer automatisiert werden. Am Ende des Diagnoseprozesses steht nicht ein numerisches Konstrukt der Personenfähigkeit, sondern eine differenzierte Charakterisierung der einzelnen Stärken und Schwächen des Probanden in Teilgebieten des untersuchten Sachbereiches. Auf der Grundlage einer solchen Wissensdiagnose können intelligente tutorielle Systeme (vgl. Albert & Lukas, 1999) zum Einsatz kommen. Ferner findet die Theorie der Wissensräume Anwendung im Bereich des automatisierten Lernens (z. B. `http://www.aleks.com`). Voraussetzung für die formale Wissensdiagnose ist, daß sich der untersuchte Sachbereich als eine endliche Menge von Fragen oder Aufgaben beschreiben läßt, die der Proband lösen kann oder nicht.

Formale Konzepte

Die Theorie der Wissensräume von Doignon & Falmagne (1985, 1999) basiert auf elementarer Mengenlehre. Der zu untersuchende Wissensbereich, eine endliche, nichtleere Menge von Aufgaben oder Items, wird als Domain Q bezeichnet: $Q = \{a, b, c, \dots\}$. Der Wissenszustand K ist die Teilmenge von Items aus Q, die ein Proband lösen kann. Alle Teilmengen von Q sind mögliche Wissenszustände, d. h. es gibt $2^{|Q|}$ mögliche, wobei $|Q|$ die Anzahl der Elemente in Q angibt. Eine Stärke der Theorie der Wissensräume besteht darin, die Anzahl der potentiellen Zustände einzuschränken und eine Struktur zu etablieren. Dazu werden die Abhängigkeitsbeziehungen unter den Aufgaben herangezogen. Wenn zur Lösung einer bestimmten Aufgabe eine Menge anderer Aufgaben nötig ist, werden diejenigen Zustände ohne die Voraussetzung nicht vorkommen. Solche Abhängigkeitsbeziehungen können die Menge der möglichen Wissenszustände drastisch einschränken, und zwar umso stärker, je mehr Abhängigkeiten es unter den Items

gibt. Das Tupel $\langle Q, \mathcal{K} \rangle$ bildet die sogenannte Wissensstruktur, wobei $\mathcal{K}$ die Menge der Zustände ist.

Die Abhängigkeit zwischen den Items kann mit der sogenannten Surmise-Relation $\precsim$ beschrieben werden. Es gilt für zwei Items $p, q \in Q$

$$p \precsim q \qquad \Leftrightarrow \qquad \begin{array}{l} \text{von der Lösung des Items } q \text{ kann die Lösung} \\ \text{des Items } p \text{ erschlossen werden,} \end{array}$$

i. e. p ist eine Voraussetzung für q. Die Surmise-Relation $\precsim$ ist reflexiv und transitiv. Daher kann sie als Quasiordnung auf Q betrachtet werden. Jede Surmise-Relation auf einem Wissensbereich Q induziert eine zugehörige Wissensstruktur $\mathcal{K}$ auf Q. Dabei ist $K \subseteq Q$ genau dann ein Zustand in $\mathcal{K}$, wenn

$$(p \precsim q \ \wedge \ q \in K) \ \Rightarrow p \in K.$$

Wenn also das Item q in einem Zustand $K \in \mathcal{K}$ ist, so muß auch p in diesem Zustand sein, gesetzt p ist eine Voraussetzung für q. Der Zusammenhang zwischen der Surmise-Relation und der zugehörigen Wissensstruktur ist im Satz von Birkhoff (1937) zusammengefaßt. Dieser besagt, daß es eine eineindeutige Korrespondenz zwischen Quasiordnungen $\precsim$ auf einer Menge Q und den Familien von Teilmengen von Q gibt, die abgeschlossen sind bezüglich Mengenvereinigung und Mengendurchschnitt. Die zur Surmise-Relation korrespondierende Wissensstruktur $\mathcal{K}$ wird als quasiordinaler Wissensraum bezeichnet. Es gilt $\forall K, K' \in \mathcal{K}$

$$(K \cup K') \in \mathcal{K} \quad \text{und} \quad (K \cap K') \in \mathcal{K}.$$

Die Existenz einer Surmise-Relation bedeutet, daß es für jedes Item $q \in Q$ genau eine Menge von Items gibt, die Voraussetzung sind, um q lösen zu können. Es gibt jedoch viele Fälle, in denen nur eine einzige Voraussetzungsmenge nicht realistisch erscheint (Falmagne et al., 1990). Zwei Probanden mit völlig unterschiedlichem Vorwissen können in der Lage sein, eine schwierige Aufgabe zu lösen. Das Vorwissen jedes von beiden ist also eine hinreichende Voraussetzungsmenge. Naheliegend ist demnach eine Verallgemeinerung der Surmise-Relation. Um die Abhängigkeit zwischen den Items zu formalisieren, wird eine Funktion $\sigma : Q \rightarrow 2^{2^Q}$ definiert, die sogenannte Surmise-Funktion. Die Surmise-Funktion $\sigma(q)$ liefert die Menge aller Zustände, die mögliche

Voraussetzungen für eine Aufgabe $q \in Q$ bilden. Ein Proband, der Item q löst, muß alle Items aus mindestens einer Menge in $\sigma(q)$ lösen. Auch zur Surmise-Funktion korrespondiert eine Wissensstruktur, der Wissensraum. Von einem Wissensraum spricht man, wenn eine Wissensstruktur $\mathcal{K}$ die leere Menge und die Domain Q enthält. Außerdem muß gelten

$$(K \cup K') \in \mathcal{K} \quad \forall K, K' \in \mathcal{K},$$

d. h. $\mathcal{K}$ ist vereinigungsstabil – die Vereinigung beliebiger Zustände K, K' aus $\mathcal{K}$ ergibt wieder einen Zustand in $\mathcal{K}$, die Durchschnittsstabilität dagegen wird beim Wissensraum nicht gefordert.

Da die Anzahl der Zustände schon bei einer kleinen Domain beträchtlich sein kann, ist es sinnvoll, die Eigenschaften eines Wissensraumes für eine sparsamere Notation auszunutzen. Wegen der $\cup$-Stabilität müssen nur die Zustände angegeben werden, die nicht durch Vereinigung gebildet werden können, die sogenannten Atome. Alle Atome zusammen bilden die Basis des Wissensraumes. Zwischen der Surmise-Funktion und einem Wissensraum besteht nun ein eindeutiger Zusammenhang: jede Voraussetzung aus σ ist ein Atom im zugehörigen Wissensraum, bzw. alle Voraussetzungen bilden die Basis für den Wissensraum $\mathcal{K}$.

Folgendes Beispiel dient der Veranschaulichung der eingeführten Begriffe. In der Domain gäbe es vier Aufgaben: $Q = \{a, b, c, d\}$. Es gibt also $2^{|Q|} = 2^4 = 16$ mögliche Wissenszustände. Nicht alle sind plausibel: Wenn a eine Voraussetzung für d ist, kommt der Zustand $K = \{d\}$ nicht vor. Es gelte folgende Abhängigkeitsbeziehung zwischen den Items: Die Aufgaben a und b haben keine Voraussetzungen. Um aber c lösen zu können muß ein Proband bereits b wissen. Für d ist a oder b eine Voraussetzung. Daraus ergibt sich die Surmise-Funktion σ

$$\begin{aligned}
\sigma(a) &= \{\{a\}\} \\
\sigma(b) &= \{\{b\}\} \\
\sigma(c) &= \{\{b, c\}\} \\
\sigma(d) &= \{\{a, d\}, \{b, d\}\}\,.
\end{aligned}$$

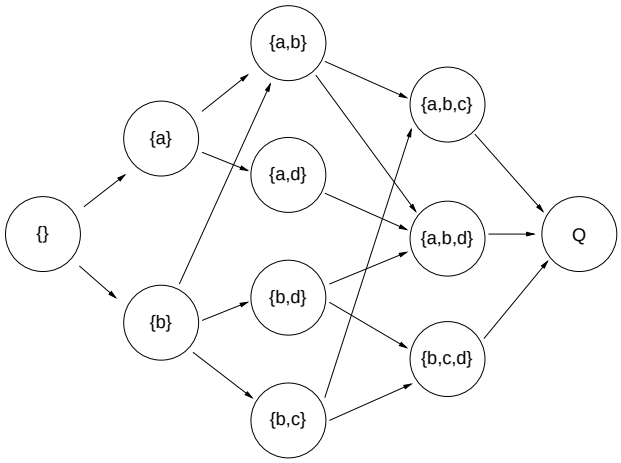

Abbildung 1: Hasse-Diagramm des Wissensraumes $\mathcal{K}$.

Definitionsgemäß ist q in jedem Zustand aus $\sigma(q)$ enthalten, d. h. jedes Item ist Voraussetzung für sich selbst. Löst ein Proband Aufgabe d, so kann er aus mindestens einer Menge in $\sigma(d)$ alle Aufgaben lösen. Alle Voraussetzungen bilden die Basis für einen Wissensraum:

$$\sigma = \{\{a\}, \{b\}, \{b, c\}, \{a, d\}, \{b, d\}\}\,.$$

Aus der Basis erhält man den Wissensraum durch Vereinigungsbildung:

$$\begin{aligned}\mathcal{K} \;=\; & \{\emptyset, \{a\}, \{b\}, \{a, b\}, \{a, d\}, \{b, c\}, \{b, d\},\\ & \{a, b, c\}, \{a, b, d\}, \{b, c, d\}, Q\}.\end{aligned}$$

Abbildung 1 zeigt die graphische Darstellung des Wissensraumes in Form eines Hasse-Diagramms. Kreise symbolisieren die Wissenszustände, durch Pfeile verbundene Zustände stehen in Teilmengenrelation.

Wissensräume können zusätzliche Eigenschaften besitzen, die für die späteren Anwendungen nützlich sind. Wissenszustände können auf einem sogenannten Lernpfad liegen. Ein Lernpfad $\mathcal{C}$ ist eine Kette von Zuständen aus $(\mathcal{K}, \subseteq)$ mit $\emptyset \subseteq C_1 \subseteq \cdots \subseteq C_n \subseteq Q$. Ein Lernpfad in $\mathcal{K}$ heißt Gradation, wenn es für alle $K \in \mathcal{C} \setminus \{Q\}$ ein Item $q \in Q \setminus K$ gibt, so daß $K \cup \{q\} \in \mathcal{C}$. Für je zwei Items gibt es einen Zustand in der Gradation, der das eine enthält, das andere aber nicht. Ist jeder Zustand in $\mathcal{K}$ in einer Gradation enthalten, so ist die zugehörige Wissensstruktur wellgraded. Bewegt

man sich entlang der Pfeile durch das Hasse-Diagramm eines Wissensraumes, der die Wellgradedness erfüllt, so enthält der nächste Zustand immer genau ein Item mehr als der gegenwärtige.

Dem theoretischen Wissensraum $\mathcal{K}$ steht der empirische Antwortraum $\mathcal{R}$ gegenüber. Dieser beinhaltet alle Antwortmuster R, die in der Untersuchungspopulation aufgetreten sind. Maximal können dies alle $2^{|Q|}$ möglichen Antwortmuster sein. Ein Antwortmuster R enthält alle Aufgaben, die ein Proband richtig beantwortet hat. Daher lassen sich Antwortmuster in Analogie zu den Wissenszuständen in Mengenschreibweise darstellen, z. B. $R = \{a, b, d\}$. Äquivalent dazu ist die Darstellung als binärer Antwortvektor $R = (1101)$. Dabei steht 1 für die richtige und 0 für die falsche Antwort bei einer Aufgabe.

1.2 Probabilistische Methoden zur Schätzung eines Wissensraumes

Wissensräume bieten die Möglichkeit zu einer sehr effizienten Wissensdiagnose. Dazu ist aber die vorherige Kenntnis der Abhängigkeitsbeziehungen zwischen den einzelnen Aufgaben nötig. Ist diese bekannt, kann aufgrund der Lösung bzw. Nichtlösung einer Menge von Aufgaben, die Lösung bzw. Nichtlösung einer anderen Aufgabe erschlossen werden. Somit muß bei der Wissensdiagnose nur ein geringer Teil der Gesamtmenge von Fragen tatsächlich gestellt werden. Ein sehr nützlicher Nebeneffekt bei diesem Vorgehen ist, daß die gestellten Fragen vom Vorwissen des Geprüften (d. h. den zuvor gelösten bzw. nicht gelösten Fragen) abhängen, und so individuell an die Prüfungssituation angepaßt sind. Dadurch kann eine Prüfung durch einen erfahrenen Prüfer automatisiert werden; die Vorteile der mündlichen Prüfung (adaptive Wissensdiagnose) bleiben erhalten, die Nachteile (Subjektivität, individuelle und soziale Störfaktoren) werden ausgeschaltet.

Voraussetzung für die Wissensdiagnose mit Wissensräumen ist also die Kenntnis des Wissensraumes für das zu testende Themengebiet. Besteht die Möglichkeit, den Raum auf induktive Art zu gewinnen? Dazu müßte man einer genügend großen Stichprobe

alle zum Themenbereich gehörenden Fragen vorlegen und die resultierenden Antworten untersuchen. Geht man von einem deterministischen Zusammenhang zwischen Antworten und Wissensraum aus, so sind der Antwortraum und der Wissensraum identisch. Das bedeutet, zu jedem Antwortmuster R aus dem Antwortraum $\mathcal{R}$ existiert genau ein identischer Wissenszustand $K \in \mathcal{K}$.

Beispiel. Gegeben sei die Domain $Q = \{a, b, c, d\}$. Eine Versuchsperson, die die Fragen a, b und d lösen kann, gibt die Antwort $R = (1101) \in \mathcal{R}$ (1 bezeichnet die richtige, 0 die falsche Antwort). Einen deterministischen Zusammenhang zwischen $\mathcal{R}$ und $\mathcal{K}$ vorausgesetzt muß sie zum Zeitpunkt der Antwort im Zustand $K = \{a, b, d\} \in \mathcal{K}$ gewesen sein.

Für den deterministischen Fall ist also eine induktive Generierung des Wissensraumes trivial: jede Antwort entspricht einem Zustand. Meist können jedoch deterministische Modelle menschliches Verhalten, das in vielerlei Hinsicht zufälligen Schwankungen unterworfen ist, nur suboptimal beschreiben. So kann bei der Wissensdiagnose nicht ausgeschlossen werden, daß ein Prüfling die Antwort nicht weiß, sondern erraten hat, bzw. trotz zugrundeliegenden Wissens einen Flüchtigkeitsfehler gemacht hat. Im vorangegangenen Beispiel könnte die richtige Antwort auf die Frage d geraten sein, während bei c ein Flüchtigkeitsfehler gemacht wurde. Dies hieße, der dahinterstehende Wissenszustand wäre $\{a, b, c\}$ und nicht $\{a, b, d\}$. Bei einer genügend großen Stichprobe sollte dies aber auch bedeuten, daß das Antwortmuster (1110) viel häufiger vorkommt als (1101), gegeben der Zustand $\{a, b, d\}$ existiert tatsächlich nicht.

Probabilistische Modelle, wie die im folgenden vorgestellten Markoffmodelle, beziehen solches Zufallsgeschehen mit ein und sind somit realistischer. Das Problem der induktiven Generierung von Wissensräumen wird dadurch allerdings erschwert, da nun nicht mehr jedem Antwortmuster ein Zustand entspricht. Eine Lösung des Problems wäre, zunächst vom allgemeineren Raum, der aus der deterministischen Beziehung folgt, auszugehen und diesen dann gegen einen kleineren Raum, der seltene Antwortmuster nicht enthält, zu testen. Grundlage für dieses Vorgehen sind sogenannte Nested Models.

Nested Models

Unter Nested Models versteht man verschachtelte Modelle. Für zwei verschachtelte Modelle gilt, daß das eine ein Spezialfall des anderen ist bzw. eines eine Verallgemeinerung des anderen darstellt. Der Vorteil des allgemeinen Modells liegt in seiner Universalität, d. h. es wird in der Lage sein, Daten besser zu beschreiben als das spezielle Modell. Dieses dagegen zeichnet sich durch eine geringere Anzahl von Parametern aus und ist dadurch sparsamer, allerdings auf Kosten der Universalität.

Hat man für einen bestimmten Datensatz ein Modell gefunden, so stellt sich oft die Frage nach einem sparsameren Spezialmodell. Kann dieses die Daten genauso gut beschreiben, wie das Universalmodell? Gibt es evtl. mehrere Spezialmodelle? Welches soll dann gewählt werden? Eine statistisch fundierte Antwort auf diese Fragen liefert der Likelihoodquotiententest, der es ermöglicht, Spezial- und Universalmodell gegeneinander zu testen und so das Spezialmodell zu verwerfen oder als passend beizubehalten.

Während in vielen Bereichen der Psychologie das Modellieren mit Nested Models zum Standard gehört (vgl. Wickens, 1982; Batchelder & Riefer, 1999), wurde die Frage, ob der Nested-Models-Ansatz auch für Wissensstrukturen geeignet ist, bisher noch nicht untersucht. Die folgenden beiden Markoffmodelle sind probabilistische Modelle für Wissensräume. Kann man mit ihnen nach dem Nested-Models-Ansatz Wissensräume induktiv aus empirischen Antwortmustern generieren?

1.3 Probabilistische Wissensstrukturen

Probabilistische Wissensstrukturen stellen eine Verfeinerung der bislang deterministisch formulierten Theorie der Wissensräume dar. Dabei wird versucht, zufällige Schwankungen im menschlichen Verhalten zu modellieren. In zweierlei Hinsicht wird diesem Zufallsgeschehen Rechnung getragen: Zum einen wird von einem deterministischen Zusammenhang zwischen Wissenszustand und Antwortmuster abgesehen. Antwortet ein Proband falsch auf eine Frage, die er gemäß seinem Zustand wissen sollte, so hat er einen Flüchtigkeitsfehler begangen, bzw. er hat eine Antwort richtig geraten, obwohl er sie nicht hätte wissen können. Zum anderen wird nicht davon ausgegangen,

daß alle Zustände in der zu untersuchenden Population gleich häufig vorkommen, manche Zustände werden häufiger sein als andere. Diese Information ist besonders für die Wissensdiagnose wertvoll.

Zur Modellierung dieser Prozesse werden zwei Funktionen benötigt. Die Antwortfunktion

$$r : 2^Q \times \mathcal{K} \to [0, 1].$$

Sie liefert $r(R|K)$, die Wahrscheinlichkeit, daß ein Proband das Antwortmuster R gibt, vorausgesetzt er befindet sich im Wissenszustand K. Außerdem muß die Wahrscheinlichkeitsverteilung

$$p : \mathcal{K} \to [0, 1]$$

spezifiziert werden, wobei $p(K)$ die Wahrscheinlichkeit des Zustands K in der Population der Probanden angibt. Modelle, die diese beiden Wahrscheinlichkeiten berechnen können, sind in der Lage, das Auftreten empirischer Antwortmuster vorherzusagen; denn mit dem Satz von der totalen Wahrscheinlichkeit ergibt sich

$$\varrho(R) = \sum_{K \in \mathcal{K}} r(R|K)p(K). \tag{1}$$

Dabei ist $\varrho(R)$ die theoretische Wahrscheinlichkeit für das Antwortmuster R. Durch die relativen Häufigkeiten kann $\varrho(R)$ direkt aus den beobachteten Daten geschätzt werden, während $r(R|K)$ und $p(K)$ zunächst unbekannte Parameter sind. Die bedingte Wahrscheinlichkeit $r(R|K)$ kann man berechnen, wenn man pro Aufgabe zwei Parameter β und η einführt. Bezeichnen β_q und η_q die Fehler- bzw. Ratewahrscheinlichkeit für die Aufgabe q, so gilt für alle $R \subseteq Q$ und $K \in \mathcal{K}$, daß

$$r(R|K) = \left(\prod_{q \in K \setminus R} \beta_q \right) \left(\prod_{q \in K \cap R} (1 - \beta_q) \right) \left(\prod_{q \in R \setminus K} \eta_q \right) \left(\prod_{q \in \overline{R \cup K}} (1 - \eta_q) \right) \tag{2}$$

unter der Annahme der lokalen stochastischen Unabhängigkeit: Bei gegebenem Zustand erfolgt die Lösung der Aufgaben stochastisch unabhängig von einander. Das bedeutet, daß β und η nur von der Aufgabe, nicht aber vom Zustand abhängen. Die

Wahrscheinlichkeit $r(R|K)$ ergibt sich also aus dem Produkt der Fehler- und Ratewahrscheinlichkeiten bzw. der jeweiligen Gegenwahrscheinlichkeiten (der Proband löst die Aufgabe korrekt mit Wahrscheinlichkeit $1-\beta$, bzw. antwortet falsch, ohne richtig zu raten, mit $1-\eta$, da die Aufgabe nicht in seinem Wissenszustand enthalten ist).

Beispiel. Seien $Q = \{a, b, c, d\}$, $K = \{a, b, c\}$ und $R = (1101)$. Unter der Annahme der lokalen Unabhängigkeit gilt dann

$$r(R|K) = \beta_c(1-\beta_a)(1-\beta_b)\eta_d.$$

Die Forderung der lokalen Unabhängigkeit ist sehr stark, da somit das Vorwissen eines Probanden für die Lösungswahrscheinlichkeit einer Aufgabe keine Rolle spielt, was in gewisser Weise dem Gedanken einer Struktur auf der Menge der Aufgaben widerspricht. Aus Gründen der Sparsamkeit wird sie dennoch aufrechterhalten.

Um die Wahrscheinlichkeitsverteilung über dem Wissensraum $\mathcal{K}$ zu bestimmen, kann nicht direkt auf die relativen Häufigkeiten der Antwortmuster zurückgegriffen werden, da im allgemeinen die Menge der Parameter $p(K)$ für eine unmittelbare Schätzung zu groß ist. Zur Lösung dieses Problems wird die Annahme eines Lernprozesses zugrunde gelegt. Man geht davon aus, daß Lernen in diskreten Schritten stattfindet und pro Lernschritt jeweils ein Item gelernt werden kann. Ferner wird angenommen, daß die ganze Population der Probanden den Lernprozeß vom absoluten Nichtwissen $\emptyset$ bis zur kompletten Beherrschung des Themengebietes Q durchläuft und zum Zeitpunkt der Diagnose eine feste Zahl von Lernschritten hinter sich hat. Voraussetzung für die Beschreibung eines solchen Lernprozesses mit Wissensräumen ist, daß diese die Wellgradedness erfüllen. Die folgenden beiden Markoffmodelle modellieren diesen Lernprozeß und erreichen dadurch die Berechnung der Wahrscheinlichkeitsverteilung über dem Wissensraum, wobei gleichzeitig die Zahl der zu schätzenden Parameter drastisch verringert wird. Eine ausführliche Erläuterung zum ersten Markoffmodell geben Doignon & Falmagne (1999), während sie das zweite nur in groben Zügen skizzieren.

Ein einfaches Markoffmodell (MM1)

Im allgemeinen muß man zwischen Markoffzuständen und Wissenszuständen unterscheiden. Beim MM1 fallen Wissenszustände und Markoffzustände aber zusammen, so daß der Übergang in einen Markoffzustand dem Übergang in einen Wissenszustand entspricht. Dieser Übergang wird durch eine stochastische Matrix M beschrieben. Zu den Annahmen des MM1 gehört, daß die Population vor dem eigentlichen Lernen nichts über die Domain Q weiß. Das bedeutet, die Anfangsverteilung der Zustände zum Zeitpunkt $n = 0$ beträgt $p_0(\emptyset) = 1$ und $p_0(K) = 0$ für $K \neq \emptyset$. Dies läßt sich im Startvektor

$$p_0(\mathcal{K}) = (1, 0, 0, \dots)$$

zusammenfassen. Bei jedem Lernschritt besteht nun die Möglichkeit, genau ein Item q zu lernen, dies geschieht mit Wahrscheinlichkeit g_q. Das Erlernen eines Items bedeutet den Übergang in einen neuen Zustand. Man kann nur in den nächsthöheren Zustand übergehen, Vergessen findet nicht statt. Das bedeutet, ein Übergang von einem Zustand K in einen anderen K' ist nur möglich, wenn $K' = K \cup \{q\}$. Dann gilt

$$p(K'_{n+1}|K_n) = g_q.$$

Die Wahrscheinlichkeit in K zu bleiben (also nicht zu lernen) ist

$$p(K_{n+1}|K_n) = 1 - \sum_{\substack{K' \in \mathcal{K}: \\ K' = K \cup \{q\}}} p(K'_{n+1}|K_n),$$

im Folgenden als $\bar{g}_{qr\dots}$ abgekürzt. Für alle anderen Zustände sind die Übergangswahrscheinlichkeiten Null. In der Übergangsmatrix M sind alle Übergangswahrscheinlichkeiten enthalten. Nach dem n-ten Lerndurchgang hat sich die Verteilung über den Zuständen verändert zu

$$p_n(\mathcal{K}) = p_0(\mathcal{K})M^n. \tag{3}$$

Damit ist die unbekannte Wahrscheinlichkeitsverteilung $p(\mathcal{K})$ auf nur einen Parameter pro Frage g_q zurückgeführt. Diese Lernwahrscheinlichkeit g_q hängt nicht vom Zustand ab, sondern ausschließlich von der Aufgabe.

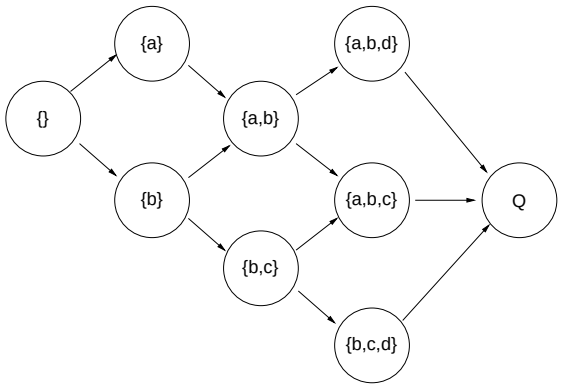

Abbildung 2: Hasse-Diagramm des Wissensraumes $\mathcal{H}$ und Zustandsübergangsdiagramm des MM1.

Beispiel. Auf der Domain $Q = \{a, b, c, d\}$ sei folgender Wissensraum gegeben:

$$\mathcal{H} = \{\emptyset, \{a\}, \{b\}, \{a, b\}, \{b, c\}, \{a, b, c\}, \{a, b, d\}, \{b, c, d\}, Q\}\,.$$

Das Hasse-Diagramm des Wissensraumes entspricht beim MM1 dem Zustandsübergangsdiagramm, da $\mathcal{H}$ der Zustandsraum des Markoffmodells ist (vgl. Abbildung 2). Dem Übergangsdiagramm entspricht die Übergangsmatrix M (Tabelle 1). Dabei sind g_a, g_b, g_c und g_d die Lernwahrscheinlichkeiten der entsprechenden Items. Die Wahrscheinlichkeit beispielsweise im Zustand $\emptyset$ zu verweilen und nicht zu lernen beträgt $\bar{g}_{ab}$ oder ausführlicher $1 - g_a - g_b$.

Das MM1 ermöglicht die Bestimmung der Wahrscheinlichkeitsverteilung $p(K)$ für alle $K \in \mathcal{K}$ und damit die Prognose der Wahrscheinlichkeiten der Antwortmuster. Dazu müssen die Aufgabenparameter g_q bestimmt werden sowie die Anzahl der absolvierten Lernschritte n. Für die Anzahl der zu schätzenden Parameter ergibt sich

$$3 \cdot |Q| + 1,$$

da für jede Aufgabe $q \in Q$ drei Parameter zu bestimmen sind, nämlich g_q, β_q und η_q und außerdem die Lernschrittzahl n. Im Vergleich zur Menge der möglichen Zustände ist diese Zahl sehr gering, allerdings um den Preis einer inhaltlichen Einschränkung: Die Annahme, daß die Lernwahrscheinlichkeit g_q nicht vom Vorwissen der Versuchsperson abhängt erscheint unrealistisch. Das nächste Markoffmodell versucht diesen Einwand auszuräumen.

	$\emptyset$	a	b	ab	bc	abc	abd	bcd	Q
$\emptyset$	$\bar{g}_{ab}$	g_a	g_b	0	0	0	0	0	0
a		$\bar{g}_b$	g_b	0	0	0	0	0	0
b			$\bar{g}_{ac}$	g_a	g_c	0	0	0	0
ab				$\bar{g}_{cd}$	0	g_c	g_d	0	0
bc					$\bar{g}_{ad}$	g_a	0	g_d	0
abc						$\bar{g}_d$	0	0	g_d
abd							$\bar{g}_c$	0	g_c
bcd								$\bar{g}_a$	g_a
Q									1

Tabelle 1: Übergangsmatrix M.

Ein komplexeres Markoffmodell (MM2)

Beim MM2 wird die zusätzliche Annahme gemacht, daß das zuletzt gelernte Item die Lernwahrscheinlichkeit für das nächste Item beeinflußt. Die Wahrscheinlichkeit, eine Aufgabe r zu lernen, hängt somit von dieser Aufgabe selbst und der zuletzt gelernten Aufgabe q ab. Diese Annahme wird folgendermaßen in die Modellierung einbezogen: g_{qr} bezeichnet die Wahrscheinlichkeit, Item r zu lernen, wenn als letztes Item q gelernt wurde. War die Versuchsperson im Zustand $\emptyset$ und lernt q als erstes, so geschieht dies mit Wahrscheinlichkeit g_{0q}. Damit fallen Wissenszustände und Markoffzustände nicht mehr zusammen. (K, q) bezeichnet den Markoffzustand, für den gilt, daß sich der Proband im Wissenszustand K befindet und als letztes Item q gelernt hat. Für einen Wissenszustand gibt es nun evtl. mehrere Markoffzustände (im Folgenden m-states genannt), höchstens aber soviele, wie der Wissenszustand Items hat. Lediglich für $\emptyset$ gibt es auch nur einen m-state, nämlich $(\emptyset)$. Ansonsten gelten die Annahmen aus dem MM1.

Beispiel. Auf der Domain $Q = \{a, b, c\}$ sei folgender Wissensraum gegeben:

$$\mathcal{P} = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}, \{b, c\}, Q\}\,.$$

Dazu gehören die Markoffzustände des MM2

$$\begin{aligned}\text{m-states}(\mathcal{P}) = \{&(\emptyset), (\{a\}, a), (\{b\}, b), (\{c\}, c),\\ &(\{a, b\}, a), (\{a, b\}, b), (\{a, c\}, a), (\{a, c\}, c), (\{b, c\}, b), (\{b, c\}, b),\\ &(Q, a), (Q, b), (Q, c)\}.\end{aligned}$$

Weil es sich bei $\mathcal{P}$ um die Potenzmenge von Q handelt, gibt es in diesem Fall pro $K \in \mathcal{P}$ soviele m-states wie Items des zugehörigen Wissenszustands. Das Zustandsübergangsdiagramm (Abbildung 3) spezifiziert die Übergangswahrscheinlichkeiten, wobei aus Gründen der Übersichtlichkeit die Gegenwahrscheinlichkeiten nicht eingezeichnet sind. Beispielsweise beträgt die Wahrscheinlichkeit, in $(\{a\}, a)$ zu verweilen und nicht zu lernen $1 - g_{ab} - g_{ac}$ oder abgekürzt $\bar{g}_{abac}$.

Auch das MM2 ermöglicht die Bestimmung der Wahrscheinlichkeitsverteilung $p(K)$ für alle $K \in \mathcal{K}$ und damit die Prognose der Wahrscheinlichkeiten der Antwortmuster. Dazu müssen die Aufgabenparameter g_{qr} bestimmt werden sowie die Anzahl der Lernschritte n. Im Vergleich zur Zahl möglicher Zustände ist die Zahl möglicher Parameter immer noch gering (sie wächst mit zunehmender Itemzahl wesentlich langsamer als die Zahl der möglichen Zustände), aber sie ist deutlich höher als beim MM1. Die mögliche

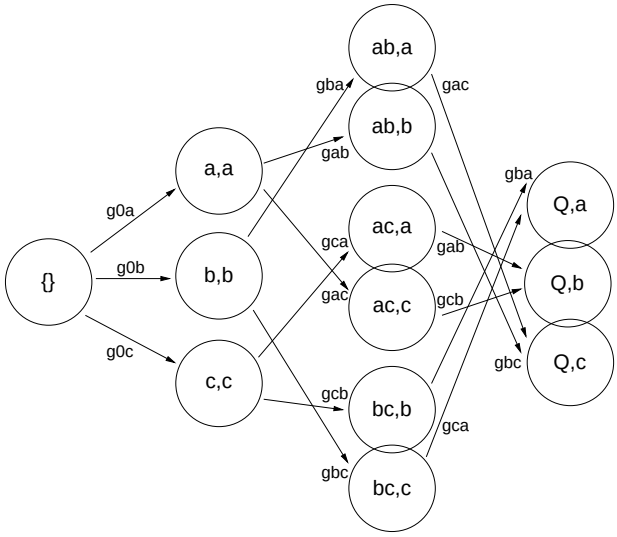

Abbildung 3: Zustandsübergangsdiagramm des MM2. Zur besseren Übersichtlichkeit wurde auf Klammern verzichtete, z. B. steht ac, a für $(\{a, c\}, a)$.

Anzahl der zu schätzenden Parameter ist

$$|Q|^2 + 2 \cdot |Q| + 1,$$

denn die Lernparameter g_{qr} sind für alle vorkommenden Paare von Fragen, deren Anzahl höchstens $|Q| \cdot (|Q| - 1)$ beträgt, sowie für alle q_{0q} zu berechnet, insgesamt also höchstens $|Q|^2$. Hinzu kommen pro Frage je ein β_q und η_q, außerdem die Lernschrittzahl n. Die Einschränkung der unrealistischen Annahme, die Lernwahrscheinlichkeit sei vom Vorwissen unabhängig (MM1), wurde also teuer mit vielen zusätzlichen Parametern erkauft. Wie eignen sich nun die beiden Markoffmodelle zur induktiven Gewinnung eines Wissensraumes aus empirischen Antwortmustern nach dem Nested-Models-Ansatz?

Probleme bei der Schätzung von Wissensräumen

Beim Versuch, einen Wissensraum aus Daten zu schätzen ohne zusätzliche strukturelle Informationen werden in vielen Fällen unterschiedliche Räume in Frage kommen, die für die Daten adäquat sein könnten. Da in der probabilistischen Fassung der Theorie der Wissensräume die Möglichkeit besteht, Raten und Flüchtigkeitsfehler von Probanden in die Diagnose miteinzubeziehen, muß entschieden werden, ob ein Antwortmuster einen dahinterliegenden Wissenszustand widerspiegelt oder eher zufällig durch Raten oder Fehler zustandegekommen ist. Mathematische Modelle sollten also die Möglichkeit bieten, auf Zustände zu verzichten, die durch empirische Antwortmuster nur wenig gestützt werden, und unterschiedliche Wissensräume bzgl. der Anpassung der Daten vergleichbar zu machen.

Der Nested-Models-Ansatz stellt ganz andere Anforderungen an Modelle. Ein Spezialmodell S eines allgemeineren Basismodells B erhält man durch Restriktionen auf dem Parameterraum von B. Grundsätzlich darf S zwar weniger Parameter enthalten als B, der Parameterraum von S muß aber Teilmenge des Parameterraumes von B sein. Eine Teilmenge erhält man durch Nullsetzen bestimmter Parameter von B. S ist dadurch nested in B. Bietet das Nullsetzen von Parametern beim MM1 oder MM2 die Möglichkeit, Zustände mit geringer empirischer Häufigkeit auszuschließen?

Das einfache Markoffmodell (MM1) bietet diese Möglichkeiten nicht. Zwar können die einzelnen Lernparameter Null gesetzt werden, die Auswirkungen für den Wissensraum sind jedoch so gravierend ($g_a = 0$ bedeutet beispielsweise, daß die Frage a in keinem Zustand mehr vorkommen kann), daß eine Anwendung für das Schätzen und Vergleichen unterschiedlicher Wissensräume für gegebenes Datenmaterial ausgeschlossen ist.

Auch das Markoffmodell 2 (MM2) scheint für diese Anwendung nur sehr bedingt geeignet zu sein. Weil nun die Lernwahrscheinlichkeit einer Frage immer von der zuletzt gelernten abhängt, beeinflußt das Nullsetzen von Lernparametern die ursprüngliche Struktur weniger als beim MM1. Leider können aber – im Gegensatz zum MM1 – die strukturellen Anforderungen von Wissensräumen im allgemeinen nicht mehr erfüllt werden. So führt das Nullsetzen von Lernparametern dazu, daß mindestens ein transienter Zustand absorbierend wird, und somit ein Lernpfad eine Sackgasse bildet, über den die Gesamtmenge Q nicht mehr erreicht werden kann. Lediglich für die eindeutigen Lernparameter g_{0a}, g_{0b}, ... bewirkt das Nullsetzen, daß die Elementarzustände ({a},a), ({b},b), ... aus dem Zustandsraum und damit $\{a\}$, $\{b\}$, ... aus dem Wissensraum verschwinden, der nun mit der ursprünglichen Struktur mittels eines Likelihoodquotiententests verglichen werden kann.

An folgendem Beispiel kann man diese Überlegungen leicht nachvollziehen: Als Wissensraum für die Domain $Q = \{a, b, c\}$ diene die Potenzmenge

$$\mathcal{P} = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}, \{b, c\}, Q\}\,.$$

Wissenszustände und Markoffzustände fallen beim MM1 zusammen. Es gibt drei Lernparameter g_a, g_b und g_c, deren Nullsetzung jeweils die Verkleinerung des Raums auf vier Zustände zur Folge hat. Der ursprüngliche Wissensraum wird dadurch völlig verändert, und es ist äußerst unwahrscheinlich, daß der resultierende Raum zu denselben Daten paßt wie der ursprüngliche. Das zugehörige MM2 hat die Markoffzustände

$$\begin{aligned}\text{m-states}(\mathcal{P}) = \{&(\emptyset), (\{a\}, a), (\{b\}, b), (\{c\}, c),\\ &(\{a, b\}, a), (\{a, b\}, b), (\{a, c\}, a), (\{a, c\}, c), (\{b, c\}, b), (\{b, c\}, b),\\ &(Q, a), (Q, b), (Q, c)\}.\end{aligned}$$

Möchte man nun beispielsweise auf den Wissenszustand $\{a, b\}$ verzichten, etwa weil die Daten kaum Hinweise auf seine Existenz geben, muß man die Parameter g_{ab} und g_{ba} nullsetzen, um ein Nested Model zu erhalten. Dies reduziert allerdings die Wahrscheinlichkeit, die Markoffzustände $(\{a, c\}, a)$ und $(\{b, c\}, b)$ zu verlassen, auf 0; sie werden absorbierend (vgl. Abbildung 3, S. 16). Lediglich Nullsetzen von z. B. g_{0a} führt zum gewünschten Effekt: Nur der Zustand $\{a\}$ verschwindet, die restliche Struktur bleibt erhalten und kann gegen die ursprüngliche getestet werden. Für eine Selektion von Zuständen auf empirischer Grundlage ist dies allerdings nicht ausreichend.

Wie die obigen Ausführungen gezeigt haben, bietet weder das MM1 noch das MM2 die Möglichkeit, mit Hilfe des Nested-Models-Ansatzes Wissensräume aus empirischen Antwortmustern zu schätzen. Es scheint nicht möglich zu sein, durch Verkleinerung einer ursprünglichen Wissensstruktur schließlich zum adäquaten Wissensraum zu gelangen; zumindest würde ein solcher Ansatz Modelle erfordern, die eine bei weitem größere Parameterzahl hätten als das MM1 oder das MM2, was zu Problemen bei der Parameterschätzung führen würde. Welche anderen Strategien gibt es, für empirisch erhobene Antwortmuster den geeigneten Wissensraum zu finden? In den klassischen Lösungsansätzen für dieses Problem wird postuliert, daß vor der Modellierung bereits Hypothesen über die Datenstruktur bestehen müssen. Möglichkeiten, zu solchen strukturellen Hypothesen zu gelangen, sind die Item-Tree-Analyse (Leeuwe, 1974; Schrepp, 1999) und die Expertenbefragung (Doignon & Falmagne, 1999). Beide Methoden sind nicht unproblematisch. Die aus der Befragung von verschiedenen Experten resultierenden Wissensräume sind keineswegs identisch (vgl. Cosyn & Thiéry, 2000). Welcher Wissensraum ist am besten geeignet, die Datenstruktur zu beschreiben? Um diese Frage zu beantworten, werden Anpassungsmaße berechnet. Mit diesen können im Nachhinein die A-priori-Hypothesen über die Struktur in den Daten evaluiert werden und evtl. – im Falle der Expertenbefragung – als fehlerhaft entlarvt werden. Ferner ermöglichen solche Anpassungsmaße die Selektion des Wissensraumes mit der besten Anpassung, d. h. wenn mehrere Strukturen für Daten in Frage kommen, wird diejenige mit der besten Anpassung gewählt.

Im Folgenden werden zwei völlig unterschiedliche Methoden vorgestellt, wie die Anpassung von Wissensräumen an Daten berechnet werden kann: probabilistische und deterministische.

1.4 Probabilistische Methoden zur Beurteilung der Anpassungsgüte von Wissensräumen

Die beiden Markoffmodelle mögen sich zwar für den Nested-Models-Ansatz als ungeeignet erwiesen haben, um die Anpassungsgüte eines Wissensraumes an Daten zu beurteilen, sind sie dennoch wertvoll. Allerdings konnte bereits gezeigt werden, daß die Verwendung des MM2 einen starken Anstieg der Parameterzahl bedeutet, so daß das sparsame MM1 das Modell der Wahl für die folgende Methode ist.

Der χ^2-Anpassungstest

Zu den gängigsten Methoden, die Anpassung eines Modells an empirische Phänomene zu beurteilen, zählt der χ^2-Anpassungstest (vgl. Bortz, 1999). Voraussetzung für die Durchführung des Tests ist, daß die Wahrscheinlichkeiten der Beobachtungskategorien durch das Modell vorhergesagt werden können. Somit können die Modellvorhersagen mit den Beobachtungen verglichen werden. In dieser konkreten Anwendung sind die Beobachtungen die empirischen absoluten Häufigkeiten der Antwortmuster. Die empirische Häufigkeit des Antwortmusters $R \in \mathcal{R}$ in der Untersuchungspopulation mit N Untersuchungseinheiten wird mit $N(R)$ bezeichnet. Die theoretische Antworthäufigkeit bezeichnet $\varrho_\vartheta(R)N$, wobei $N = \sum_{R\in\mathcal{R}} N(R)$ die Gesamtzahl der Versuchspersonen ist. Probabilistische Wissensstrukturen spezifizieren $\varrho_\vartheta(R)$ durch

$$\varrho_\vartheta(R) = \sum_{K\in\mathcal{K}} r(R|K)p(K)$$

(vgl. Gleichung 1, S. 11), wobei sich die bedingte Wahrscheinlichkeit $r(R|K)$ auf die Parameter β_q und η_q zurückführen läßt, und die Wahrscheinlichkeitsverteilung $p(\mathcal{K})$ durch das MM1 vorhergesagt wird, das lediglich einen Parameter g_q pro Frage sowie

die Anzahl der Lernschritte n als Parameter benötigt. Im Parametervektor

$$\vartheta = (g_a, g_b, \ldots, \beta_a, \beta_b, \ldots, \eta_a, \eta_b, \ldots, n)$$

sind alle Parameter enthalten, die zur Berechnung der theoretischen Antworthäufigkeiten benötigt werden.

Als Maß der Anpassungsgüte des probabilistischen Modells an die Daten werden für alle Antwortmuster $R \in \mathcal{R}$ die Differenzen zwischen empirischen und theoretischen Antworthäufigkeiten gebildet. Diese werden quadriert, an den theoretischen Häufigkeiten normiert und aufsummiert. Für die daraus resultierende Größe $\chi^2(\vartheta; N(\mathcal{R}))$ gilt

$$\chi^2(\vartheta; N(\mathcal{R})) = \sum_{R \subseteq Q} \frac{(N(R) - \varrho_\vartheta(R)N)^2}{\varrho_\vartheta(R)N} \tag{4}$$

mit dem Parametervektor ϑ und dem Datenvektor $N(\mathcal{R})$, der alle Antworthäufigkeiten $N(R)$ enthält. Diese Größe ist asymptotisch χ^2-verteilt. Die Freiheitsgrade (df) errechnen sich aus der Differenz zwischen der Anzahl der möglichen verschiedenen Antwortmuster minus eins und der Anzahl der zu schätzenden Parameter. Beim MM1 beträgt die Parameterzahl $3 \cdot |Q| + 1$, und so ergibt sich für die χ^2-Statistik eine Zahl von

$$\nu = (2^{|Q|} - 1) - (3 \cdot |Q| + 1) \tag{5}$$

Freiheitsgraden. Die Nullhypothese behauptet, das Modell sei für die Daten adäquat. Nach Doignon & Falmagne (1999) kann das Modell bei einem p-value kleiner als 5% verworfen werden. Die χ^2-Methode bietet eine Möglichkeit, unterschiedliche Modelle bezüglich der Anpassung zu vergleichen. Bei gleicher Zahl der Freiheitsgrade gilt, daß das Modell mit der kleineren χ^2-Größe die Daten besser beschreibt.

1.5 Deterministische Methoden zur Beurteilung der Anpassungsgüte von Wissensräumen

Bei den im vorangegangenen Abschnitt erörterten probabilistischen Wissensstrukturen wurde auf zweierlei Wegen versucht, dem Zufallsgeschehen bei der Modellierung

Rechnung zu tragen. Erstens wird die Möglichkeit berücksichtigt, daß Flüchtigkeitsfehler oder das Erraten der Lösung die Daten verrauschen können. Zweitens wird nicht angenommen, daß alle Wissenszustände gleich häufig in der Untersuchungspopulation vorkommen, sondern eine Verteilung über diesen postuliert, die sich in Abhängigkeit von der Anzahl der Lernschritte verändert.

Die folgende deterministische Methode ist frei von solchen probabilistischen Verfeinerungen der Theorie der Wissensräume. Es wird ein deterministischer Zusammenhang zwischen Antwortmustern und Wissensraum angenommen. Diskrepanzen zwischen diesen gehen in das Anpassungsgütemaß ein.

Die Diskrepanzstatistik

Eine deterministische Methode zur Beurteilung der Anpassung eines Wissensraumes an Daten ist die sogenannte Diskrepanzstatistik, die von Kambouri et al. (1994) ursprünglich eingeführt wurde, um die Ähnlichkeit zweier Wissensräume zu bestimmen. Aber auch für einen gegebenen Datensatz und einen dazugehörigen Wissensraum läßt sich die Diskrepanzstatistik berechnen.

Für jedes Antwortmuster $R \in \mathcal{R}$ gilt, daß es entweder einen Zustand $K \in \mathcal{K}$ gibt, der optimal zu diesem Antwortmuster paßt (d. h. identisch mit dem Antwortmuster ist), oder einen der ähnlich dem Antwortmuster ist. Als Maß der Ähnlichkeit wird die symmetrische Mengendifferenz herangezogen. Es gilt

$$R \Delta K = R \setminus K \cup K \setminus R.$$

Die Distanz zwischen Antwort R und Zustand K ist die Kardinalität der symmetrischen Mengendifferenz

$$d(R, K) = |R \Delta K|.$$

Für eine Antwort R läßt sich die Distanz zu jedem Zustand $K \in \mathcal{K}$ bilden. Das Minimum

$$d(R, \mathcal{K}) = \min_{K \in \mathcal{K}} \{d(R, K)\}$$

heißt Distanz zwischen dem Antwortmuster R und dem Wissensraum $\mathcal{K}$. Bildet man nun für jede Antwort aus dem Antwortraum $\mathcal{R}$ die Distanz zum Wissensraum $\mathcal{K}$, erhält man die sogenannte Diskrepanzfunktion, d. h. eine Verteilungsfunktion der minimalen Distanz von $\mathcal{R}$ und $\mathcal{K}$. Formal ist diese definiert durch

$$f_{\mathcal{R},\mathcal{K}}(d) = \frac{1}{N} \sum_{\substack{R\in\mathcal{R}:\\ d(R,\mathcal{K})=d}} N(R),$$

wobei $N(R)$ die Häufigkeit des Antwortmusters R bezeichnet, und $N = \sum_{R\in\mathcal{R}} N(R)$ ist, also die Anzahl der Versuchspersonen. Der Erwartungswert der Diskrepanzfunktion heißt Diskrepanzindex und läßt sich mit folgender Formel berechnen:

$$di(\mathcal{R},\mathcal{K}) = \sum_{d=0}^{|Q|} d \cdot f_{\mathcal{R},\mathcal{K}}(d). \tag{6}$$

Der Diskrepanzindex drückt aus, wie stark sich im Durchschnitt die Antworten von den Wissenszuständen unterscheiden. Je kleiner er ist, desto geringer ist der durchschnittliche Unterschied und desto besser ist die Anpassung des Wissensraumes an die Daten.

Beispiel. Es seien $Q = \{a, b\}$, $\mathcal{K} = \{\emptyset, \{a\}, Q\}$ und $\mathcal{R} = \{(00), (10), (01), (11)\}$. Ferner seien die Antworthäufigkeiten gegeben durch $N(00) = N(10) = N(11) = 3$ und $N(01) = 1$. Es gäbe also $N = 10$ Versuchspersonen. Die Distanzen zwischen den Antworten und dem Wissensraum betragen

$$\begin{aligned} d((00),\mathcal{K}) &= 0 \\ d((10),\mathcal{K}) &= 0 \\ d((01),\mathcal{K}) &= 1 \\ d((11),\mathcal{K}) &= 0. \end{aligned}$$

Damit ergibt sich für die Diskrepanzfunktion

$$\begin{aligned} f_{\mathcal{R},\mathcal{K}}(0) &= \frac{1}{10} \cdot (3+3+3) = .9 \\ f_{\mathcal{R},\mathcal{K}}(1) &= \frac{1}{10} \cdot 1 = .1 \end{aligned}$$

und ein dazugehöriger Diskrepanzindex von $di(\mathcal{R},\mathcal{K}) = .1$.

Probabilistische Konzepte mit der Diskrepanzstatistik

Obwohl die Diskrepanzstatistik als deterministische Methode konzipiert ist, besteht dennoch die Möglichkeit, die Wahrscheinlichkeitsverteilung über dem Wissensraum $p(\mathcal{K})$ sowie die Fehler- und Ratewahrscheinlichkeiten (β_q, η_q) mit ihrer Hilfe zu berechnen. Die Verteilung $p(\mathcal{K})$ erhält man, indem man für jedes Antwortmuster R überprüft, welcher Zustand die minimale Distanz hat. Die Häufigkeit $N(R)$ aller Antwortmuster, für die derselbe Zustand gewählt wurde, werden addiert. Wenn allerdings mehrere Zustände die gleiche minimale Distanz haben, muß die Antworthäufigkeit des Antwortmusters, für das mehrere Zustände in Frage kommen, durch die Anzahl der zugeordneten Zustände geteilt werden, bevor sie als Summand in die Summe eingeht, also $N(R)/$(Anzahl der zugeordneten Zustände). Teilt man die resultierende Summe durch die Anzahl der Versuchspersonen, so erhält man die geschätzte Wahrscheinlichkeit des Zustands in der Population. Diese Prozedur ist für alle $K \in \mathcal{K}$ zu wiederholen. Formal ausgedrückt wird die Wahrscheinlichkeit für einen Wissenszustand $K \in \mathcal{K}$ geschätzt durch

$$\hat{p}(K) = \frac{1}{N} \sum_{\substack{R \in \mathcal{R}: \\ d(R,K)=d(R,\mathcal{K})}} N(R)/\alpha(R). \tag{7}$$

Dabei gibt die Funktion $\alpha : \mathcal{R} \to \mathbb{N}$ an, zu wievielen Zuständen aus $\mathcal{K}$ das Antwortmuster R minimale Distanz aufweist:

$$\alpha(R) = |\{K \in \mathcal{K} : d(R,K) = d(R,\mathcal{K})\}|.$$

Die Fehlerwahrscheinlichkeit β_q ist die bedingte Wahrscheinlichkeit[1], q falsch zu beantworten, obwohl die Versuchsperson q im Wissenszustand hat, d. h.

$$\beta_q = p(R_{\bar{q}}|K_q), \tag{8}$$

wobei $R_{\bar{q}} := \{R \in \mathcal{R} : q \notin R\}$ alle erdenklichen Antwortmuster bezeichnet, bei denen q falsch beantwortet wird, während $K_q := \{K \in \mathcal{K} : q \in K\}$ für alle Zustände

[1]Ich verdanke Thomas Augustin wertvolle Hinweise zur Herleitung der bedingten Wahrscheinlichkeiten.

steht, die q enthalten, z. B. auch die Domain Q. Nach der Definition der bedingten Wahrscheinlichkeit ergibt sich

$$\beta_q = \frac{p(R_{\bar{q}} \wedge K_q)}{p(K_q)}.$$

Die Wahrscheinlichkeit $p(K_q)$ ergibt sich – wegen Disjunktheit – als Summe der Einzelwahrscheinlichkeiten der Zustände, die die Bedingung erfüllen. Diese können der Verteilung $p(\mathcal{K})$ entnommen werden: $p(K_q) = \sum_{K \in \mathcal{K}: q \in K} p(K)$. Ein Ereignis aus $R_{\bar{q}} \wedge K_q$ tritt ein, wenn einem Antwortmuster aus $R_{\bar{q}}$ ein Zustand aus K_q zugeordnet wurde. Die geschätzte Wahrscheinlichkeit für ein solches Ereignis ist Antworthäufigkeit geteilt durch Anzahl der zugeordneten Zustände, relativiert an der Versuchspersonenzahl. Alle diese Wahrscheinlichkeiten addieren sich zu

$$\hat{p}(R_{\bar{q}} \wedge K_q) = \frac{1}{N} \sum_{K \in \mathcal{K}: q \in K} \sum_{\substack{R \in \mathcal{R}: q \notin R \\ d(R,K) = d(R,\mathcal{K})}} N(R)/\alpha(R).$$

Beispiel. Es seien $Q = \{a, b\}$, $\mathcal{K} = \{\emptyset, \{a\}, Q\}$ und $\mathcal{R} = \{(00), (10), (01), (11)\}$ mit den Antworthäufigkeiten $N(00) = N(10) = N(11) = 3$ und $N(01) = 1$ ($N = 10$). Den Antwortmustern werden die Zustände mit minimaler Distanz zugeordnet:

$$\begin{aligned} (00) &\rightarrow \emptyset \\ (10) &\rightarrow \{a\} \\ (01) &\rightarrow \emptyset, Q \\ (11) &\rightarrow Q. \end{aligned}$$

Also sind $\alpha(00) = \alpha(10) = \alpha(11) = 1$ und $\alpha(01) = 2$. Als Schätzung der Zustandswahrscheinlichkeiten gilt dann nach Gleichung 7

$$\begin{aligned} \hat{p}(\emptyset) &= \frac{1}{10}(3/1 + 1/2) = 35/100 \\ \hat{p}(\{a\}) &= 30/100 \\ \hat{p}(Q) &= 35/100. \end{aligned}$$

Die bedingte Wahrscheinlichkeit β_a bezeichnet die Wahrscheinlichkeit, die Frage a falsch zu beantworten, obwohl man sie eigentlich wissen sollte, also

$$\begin{aligned}\beta_a &= p(R_{\bar{a}}|K_a) \\ &= \frac{p(R_{\bar{a}} \wedge K_a)}{p(K_a)} \\ &= \frac{p((00) \wedge \{a\}) + p((01) \wedge \{a\}) + p((00) \wedge Q) + p((01) \wedge Q)}{p(\{a\}) + p(Q)}.\end{aligned}$$

Durch Einsetzen der Antworthäufigkeiten ergibt sich als Schätzung

$$\hat{\beta}_a = \frac{0 + 0 + 0 + (1/10 \cdot 1/2)}{30/100 + 35/100} = 1/13.$$

Die bedingten Wahrscheinlichkeiten $1 - \beta_q$, η_a und $1 - \eta_q$ werden auf analoge Weise berechnet. Für die Ratewahrscheinlichkeit gilt

$$\eta_q = p(R_q|K_{\bar{q}}), \tag{9}$$

wobei $p(K_{\bar{q}}) = \sum_{K \in \mathcal{K}: q \notin K} p(K)$ ist. Die Menge R_q steht für alle Antwortmuster, bei denen q richtig beantwortet wird, während $K_{\bar{q}}$ für alle Zustände steht, die q nicht enthalten. Somit liefert die ursprünglich deterministische Methode der Diskrepanzstatistik Wahrscheinlichkeiten, die mit denen der probabilistischen Wissensstrukturen vergleichbar sind.

1.6 Fragestellung

Anpassungsmaße ermöglichen die Selektion der für empirische Antwortmuster adäquaten Wissensstruktur. Probabilistische und deterministische Anpassungsmaße kommen auf völlig unterschiedliche Weise zustande. In dieser Arbeit soll die Beziehung zwischen beiden Methoden geklärt werden, um eine bessere Grundlage für Entscheidungen zu schaffen, die aufgrund eines Anpassungsmaßes zugunsten einer Wissensstruktur oder gegen sie getroffen werden. Folgende vier Fragen sollen anhand von computersimulierten sowie von empirisch erhobenen Antwortmustern untersucht werden:

Lassen sich die Modellparameter des MM1 aus simulierten Antwortmustern zurückschätzen?

Computersimulationen bieten die Möglichkeit, die Identifizierbarbeit der Parameter des Markoffmodells zu überprüfen. Nur wenn sich die zur Simulation verwendeten Parameter zurückschätzen lassen, ist eine empirische Anwendung sinnvoll.
Hypothese: Für große Stichproben gelingt die Parameterschätzung hinreichend exakt.

Beeinflussen unterschiedliche Wissensstrukturen die Anpassungsmaße auf unterschiedliche Weise?

Wie verhalten sich die Anpassungsmaße, wenn die Form des Wissensraumes bzw. die Anzahl der Items systematisch variiert werden?
Hypothese: Beide Anpassungsmaße reagieren gleichsinnig auf Veränderungen des Wissensraumes, d. h. es besteht eine monotone Beziehung zwischen beiden.

Beurteilt die Diskrepanzstatistik den Fit zu optimistisch?

Die χ^2-Methode kompensiert die Anzahl der Wissenszustände durch die Zahl der Freiheitsgrade, während ein solcher Kompensationsmechanismus bei der Diskrepanzstatistik fehlt. Ferner wird von der Diskrepanzstatistik stets der nächstliegende Zustand ausgewählt und nicht der möglicherweise wahre Zustand, der aber eine größere Distanz aufweist.

Hypothese: Die Diskrepanzstatistik beurteilt die Anpassungsgüte besser als die χ^2-Methode. Dies zeigt sich insbesondere in einer Unterschätzung von Fehler- und Ratewahrscheinlichkeiten im Vergleich zur χ^2-Statistik.

Wie beurteilen die Anpassungsmaße die Anpassung eines Wissensraumes an empirisch erhobene Antwortmuster?

Die Anpassung von Wissensräumen an empirische Antwortmuster ist meist schwierig, da eine Struktur aufgrund von A-priori-Hypothesen zugrundegelegt werden muß.
Hypothese: Mit der Diskrepanzstatistik wird leichter eine zufriedenstellende Anpassung erreicht als mit der χ^2-Methode.

2 Methoden

2.1 Computersimulation und Parameterschätzung mit dem MM1

Ein grundsätzliches Problem bei der Modellierung von Beobachtungen durch latente Prozesse ist die Identifizierbarkeit von Modellen bzw. der Parameter (Wickens, 1982). Zwei Modelle, die exakt dieselben Vorhersagen bezüglich der Beobachtungskategorien machen, können anhand ihrer Anpassung nicht unterschieden werden. Sie sind nicht identifizierbar verschieden. Sind Modelle nicht identifizierbar verschieden, so lassen sich die sie unterscheidenden Parameter nicht schätzen. Nicht schätzbare Parameter heißen nicht identifizierbar. Die Überprüfung der Identifizierbarkeit ist meistens nicht ganz unproblematisch. Wenn für die Parameterschätzer analytische Lösungen existieren, ist zwar offensichtlich, ob die Identifizierbarkeit gegeben ist: Enthält ein Modell nicht identifizierbare Parameter, so ergeben sich weniger unterschiedliche Gleichungen als zu schätzende Parameter. Ein Gleichungssystem mit mehr Unbekannten als Gleichungen ist nicht eindeutig lösbar. Im allgemeinen aber werden die Schätzungen durch numerische Optimierung ermittelt. Sind die Parameter nicht identifizierbar, so sind die Schätzungen nicht eindeutig. Je nach Schätzmethode kann das gefundene Minimum oder Maximum auch mit anderen Lösungen erreicht werden. Eine Möglichkeit, die Identifizierbarkeit zu prüfen, bieten Computersimulationen. Die Parameter sind identifizierbar, wenn die Schätzungen den für die Simulation festgesetzten Werten entsprechen.

Erster Schritt bei der Datenmodellierung sollte also immer die Beantwortung der Frage sein, ob sich die Modellparameter wieder aus den Daten zurückschätzen lassen. Dazu sind Computersimulationen unumgänglich; denn sie erlauben es, mit vorher festgelegten Parametern Daten zu generieren. Nur wenn die aus den simulierten Daten geschätzten Parameter die ursprünglich festgesetzten exakt genug annähern, ist eine Anwendung des Modells auf empirisch erhobene Antwortmuster überhaupt sinnvoll.

Computersimulation

Wenn man Daten – also Antworten virtueller Versuchspersonen – im Bereich der Theorie der Wissensräume simulieren will, muß man zunächst einen Wissensraum annehmen, der der simulierten Population zugrunde liegen soll. Für die Modellierung mit dem MM1 müssen ferner die Modellparameter g_q (die Lernwahrscheinlichkeiten für jedes Item q) sowie die Anzahl der Lernschritte n, die die Population durchlaufen hat, spezifiziert werden. Für eine reale Prüfungssituation müssen schließlich die Fehler- bzw. Ratewahrscheinlichkeit bei jeder Frage (β_q bzw. η_q) festgesetzt werden. Mit den Parametern g_q und n steht auch die Wahrscheinlichkeitsverteilung der Zustände $p_n(\mathcal{K})$ fest, sie wird mit Gleichung 3 (S. 13) berechnet.

Zwei Zufallsmechanismen spielen bei der Simulation von Daten eine Rolle: Zunächst wird per Zufallsgenerator bestimmt, in welchem Wissenszustand sich die virtuelle Versuchsperson befindet. Dabei wird die Zustandsverteilung $p(\mathcal{K})$ verwendet. Dann wird für jede Frage überprüft, ob die Versuchsperson darauf gemäß ihrem Zustand richtig antworten kann oder nicht. Falls das Item im Zustand der Vp ist, wird mit einem zweiten Zufallsgenerator bestimmt, ob sie einen Flüchtigkeitsfehler mit Wahrscheinlichkeit β begeht oder nicht; falls das Item nicht im Zustand ist, rät die Vp richtig mit Wahrscheinlichkeit η oder gibt die falsche Antwort.

Die Software zur Simulation von Antworten wurde in C programmiert und kann im Anhang A.1 eingesehen werden. Um einen systematischen Bias in den Daten zu verhindern, wurden verbesserte Zufallsgeneratoren implementiert (Press et al., 1988). Der Output des Programms ist für eine vorher festzulegende Zahl von Probanden je ein binärer Antwortvektor. Mit 1 wird die richtige, mit 0 die falsche Antwort codiert, also beispielsweise `11010` ... Die Länge des Vektors entspricht der Mächtigkeit der Domain Q.

Parameterschätzung

Die in Abschnitt 1.4 erläuterte χ^2-Methode kann auch zur Parameterschätzung verwendet werden. Gesucht ist dann der Parametervektor ϑ, der bei gegebenen Antworthäufig-

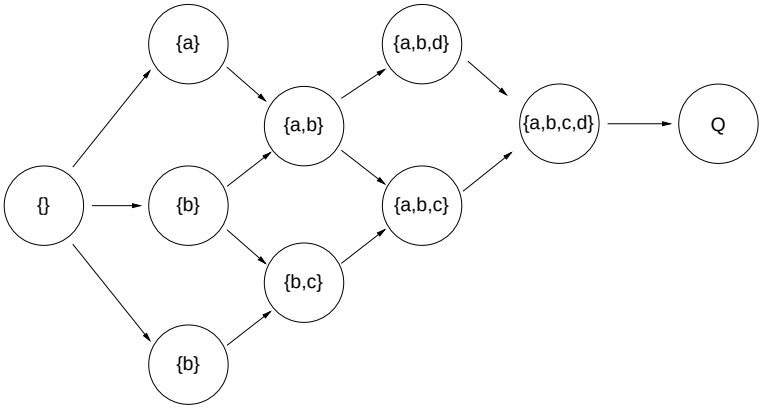

Abbildung 4: Hasse-Diagramm des Wissensraumes $\mathcal{M}$.

keiten $N(\mathcal{R})$ die χ^2-Größe aus Gleichung 4 (S. 21) minimiert. Beim MM1 besteht der Parametervektor aus den drei Parametern g, β und η pro Item und der Lernschrittzahl n. Da keine eindeutige analytische Lösung dieses Minimierungsproblems existiert, wird ein Minimum mittels iterativer Verfahren durch numerische Optimierung erreicht. Die Software zur Parameterschätzung wurde von Jürgen Heller in C programmiert. Sie implementiert den Minimierungsalgorithmus PRAXIS (Gegenfurtner, 1992; Brent, 1973). Dieses Programm verarbeitet die mit dem Simulationsprogramm generierten Antwortmuster und liefert als Output Parameterschätzer für alle g_q, β_q, η_q und n. Allerdings konnte bereits in anderen Untersuchungen (vgl. Fries, 1997) gezeigt werden, daß sich die Lernschrittzahl n nicht zuverlässig aus den Daten schätzen läßt. Da dieser Parameter ohnehin nicht eindeutig zu interpretieren ist, wurde er konstant auf den Wert der Simulation gesetzt. Dadurch gewinnt man einen Freiheitsgrad bei der χ^2-Statistik. Dies beeinträchtigt aber nicht die Möglichkeit, das MM1 auf seine Anwendbarkeit bzw. die Identifizierbarkeit der Parameter zu prüfen.

Das Zusammenspiel von Computersimulation und Schätzprogramm wurde an unterschiedlichen Wissensstrukturen überprüft. An folgendem Wissensraum läßt sich das Vorgehen exemplarisch darstellen. Der Wissensraum

$$\mathcal{M} = \{\emptyset, \{a\}, \{b\}, \{c\}\{a, b\}, \{b, c\}, \{a, b, c\}, \{a, b, d\}, \{a, b, c, d\}, Q\}$$

(vgl. Abbildung 4) mit der Domain $Q = \{a, b, c, d, e\}$ wurde einer Population von $N = 100000$ virtuellen Versuchspersonen zugrunde gelegt. Diese Population habe $n =$

10 Lernschritte durchlaufen. Ferner wurden die Parameter g, β und η für alle Items a, b, c, d und e festgesetzt. Mit diesen Vorgaben wurden mit der Simulationsroutine 100000 Antworten generiert und dem Schätzprogramm als Input übergeben. Nach der Festlegung von Startwerten für jeden Parameter und der konstanten Lernschrittzahl auf $n = 10$ wurde versucht, die zur Simulation verwendeten Parameter zurückzuschätzen.

2.2 Vergleich der Anpassungsmaße bei unterschiedlichen Wissensräumen

Um die Beziehung zwischen dem probabilistischen und dem deterministischen Anpassungsmaß näher zu untersuchen, wurden Wissensräume systematisch verändert und die Reaktion der Anpassungsmaße getestet. Zunächst wurde die Form der Wissensstruktur variiert, um den Einfluß auf die Anpassung zu untersuchen. Die vier verwendeten Formen sind in Abbildung 5 für den Fall $Q = \{a, b, c\}$ graphisch veranschaulicht. Sie weisen zunehmende Komplexität bzw. Strukturiertheit auf. Die einfachste mögliche Form ist die Form einer Kette (chain), bei der nur ein einziger Lernpfad zum Zustand Q führt. Die komplexeste Form stellt die Potenzmenge (power) dar, die alle maximal möglichen Lernpfade enthält. Die Formen ring und helix liegen bezüglich der Komplexität zwischen chain und power. Als zweiter Parameter wurde die Größe des Raumes variiert, operationalisiert durch die Anzahl der Items der zugrunde liegenden Domain Q. Es wurden Wissensräume mit drei, vier, fünf und sechs Items gebildet. Aus der Kombination der beiden Parameter Form und Größe resultieren sowohl sehr einfache (chain mit 3 Items $\rightarrow$ 4 Wissenszustände) als auch recht komplexe Strukturen (power mit 6 Items $\rightarrow$ 64 Wissenszustände). Alle 16 Form-Größe-Kombinationen wurden untersucht.

Für den Vergleich der Anpassungsmaße wurde folgendes Vorgehen gewählt: Erst wurde ein Wissensraum $\mathcal{K}$ festgelegt durch die Kombination von Form und Größe. Dann wurden mit diesem Raum und einer festen Zahl virtueller Versuchspersonen zweimal Antworten generiert, d. h. es entstanden zwei Antworträume $\mathcal{A}$ und $\mathcal{B}$. Wegen der bei der Simulation verwendeten Zufallsgeneratoren unterscheiden sich die Antworten in $\mathcal{A}$ und $\mathcal{B}$. Auch der Wissensraum $\mathcal{K}$ paßt besser zu einem von beiden Antworträu-

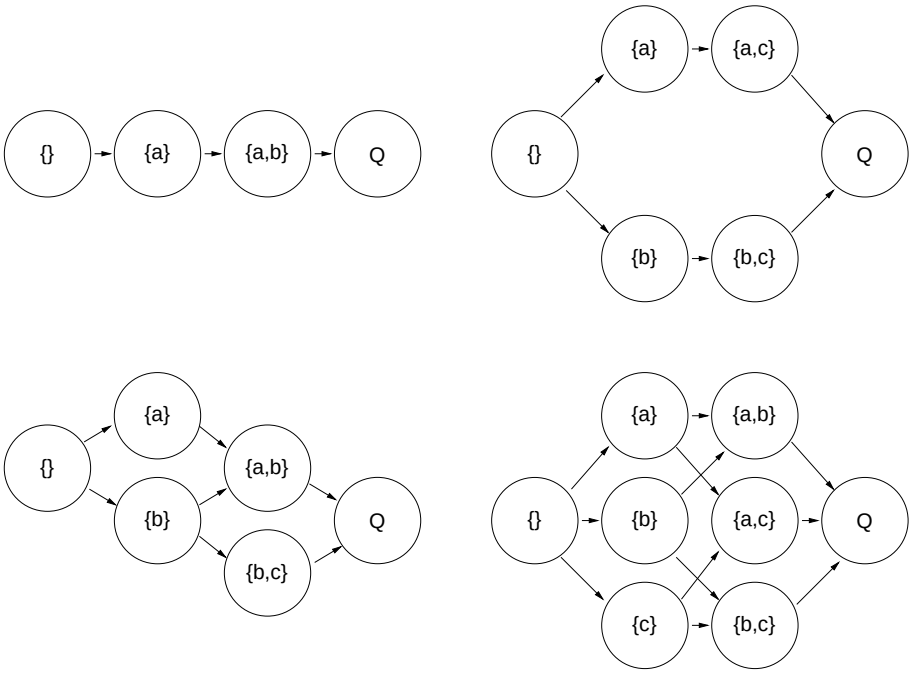

Abbildung 5: Vier verschiedene Formen von Wissensstrukturen (v. li. oben n. re. unten: chain, ring, helix und power).

men. Es ist sogar denkbar, daß $\mathcal{K}$ als Modell für einen Antwortraum abgelehnt werden muß, nämlich dann, wenn die Daten durch Flüchtigkeitsfehler und Raten stark verrauscht sind. Im nächsten Schritt wurden für jeden Antwortraum die χ^2-Statistik und der Diskrepanzindex berechnet, die jeweils die Anpassung an denselben Wissensraum $\mathcal{K}$ beurteilen (Die Software zur Berechnung der Diskrepanzstatistik, sowie der Rate- und Fehlerwahrscheinlichkeiten kann im Anhang A.2 eingesehen werden.). Die vier resultierenden Werte $\chi^2(\vartheta; N(\mathcal{A}))$ und $\chi^2(\vartheta; N(\mathcal{B}))$ bzw. $di(\mathcal{A}, \mathcal{K})$ und $di(\mathcal{B}, \mathcal{K})$ konnten paarweise verglichen werden, wobei der jeweils kleinere Wert für einen besseren Fit spricht. Unter Voraussetzung einer monotonen Beziehung zwischen der χ^2-Statistik und dem Diskrepanzindex sollte einem kleineren (größeren) $\chi^2(\vartheta; N(\mathcal{A}))$ auch ein kleinerer (größerer) $di(\mathcal{A}, \mathcal{K})$ entsprechen, d. h. beide Methoden sollten bei der Beurteilung der Anpassung zum gleichen Ergebnis kommen.

Bei diesem Vorgehen ergibt sich allerdings ein grundsätzliches Interpretationsproblem. Die zu erwartenden Unterschiede der χ^2-Statistiken bzw. Diskrepanzindizes sind rein deskriptiver Natur, und es stellt sich die Frage nach der statistischen Absicherung

und der Beurteilung der statistischen Bedeutsamkeit der Unterschiede. Die approximative Verteilungsform der χ^2-Größe ist bekannt; somit sind zumindest dichotome Entscheidungen über die Ablehnung bzw. Beibehaltung eines Modells für den Datensatz statistisch vertretbar. Voraussetzung für die Approximation durch die χ^2-Verteilung ist aber, daß die theoretische Häufigkeit jedes Antwortmusters mindestens 5 beträgt (vgl. Bortz, 1999), was bei größeren Wissensräumen unrealistisch erscheint, da eine immense Datenmenge von Nöten wäre. Die wahre Verteilung der χ^2-Größe kann sich also bei Verletzung dieser Voraussetzung von einer χ^2-Verteilung unterscheiden. Die Verteilung des Diskrepanzindexes dagegen ist völlig unbekannt; große Werte deuten zwar auf schlechte Anpassung hin, wann ein Modell abgelehnt werden muß, ist aber an einem einzigen *di*-Wert nicht ersichtlich.

Die Lösung dieses Interpretationsproblems liegt in der Generierung der unbekannten Verteilungen. Dazu wurden jeweils 1000 Datensätze für einen Wissensraum mit fester Form und Größe simuliert und 1000 χ^2- und 1000 *di*-Werte berechnet. Somit entstanden je eine Verteilung für die χ^2-Statistik und für den *di*. Das 95%-Perzentil markiert jeweils den Beginn des Ablehnungsbereiches. Größere Werte treten in weniger als 5% der Fälle auf und sprechen somit für die Ablehnung des Modells. Dichotomisiert man beide Verteilungen am 95%-Perzentil, so entsteht ein Vierfelderschema, und man erhält eine gemeinsame Verteilung. Diese gibt an wie häufig beide Methoden gleichzeitig zur Beibehaltung bzw. Ablehnung des Modells raten, bzw. wie häufig sie im Widerspruch zueinander stehen. Mit einem χ^2-Unabhängigkeitstest läßt sich nun die Beziehung der beiden Maße statistisch überprüfen. Unter Voraussetzung einer engen Beziehung zwischen χ^2-Statistik und Diskrepanzindex ist eine deutliche Ablehnung der Unabhängikeitshypothese zu erwarten.

2.3 Vergleich der Anpassungsmaße bezüglich der berechneten Fehlerwahrscheinlichkeiten

Da sowohl bei der Berechnung der χ^2-Statistik als auch des Diskrepanzindexes Schätzungen für die Fehler- und Ratewahrscheinlichkeiten abgegeben werden, erscheint

es sinnvoll, diese Wahrscheinlichkeiten miteinander zu vergleichen. Dabei kann die Frage beantwortet werden, welche der beiden Methoden konservativere und welche liberalere Schätzungen für die β- und η-Parameter liefert. Die Unterschätzung von Fehler- und Ratewahrscheinlichkeiten geht einher mit einer zu optimistischen Beurteilung der Modellanpassung an die Daten bzw. mit einer Überschätzung des Fit. Da die Diskrepanzstatistik die Modellanpassung umso besser beurteilen muß, je mehr Zustände der Wissensraum umfaßt, scheint es plausibel anzunehmen, daß diese die Modellanpassung besser einschätzt als das χ^2-Maß, bei dem die Zahl der Wissenszustände durch die Zahl der Freiheitsgrade kompensiert wird.

Um die Frage zu klären, ob die Diskrepanzstatistik die Modellanpassung zu optimistisch beurteilt, indem β- und η-Parameter unterschätzt werden, wurden mit unterschiedlichen Wissensräumen Computersimulationen und Parameterschätzungen durchgeführt. Für einen Raum von bestimmter Form und Größe wurden Antworten generiert und anschließend sowohl mit der χ^2-Methode als auch mit der Diskrepanzstatistik die Fehler- und Rateparameter geschätzt. Zur Erleichterung des Vergleichs der beiden Methoden wurden die mittleren Parameter gebildet, wobei der mittlere β-Parameter

$$\hat{\beta}_{\chi^2} = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \hat{\beta}_{q,\chi^2}$$

dem arithmetischen Mittel aus allen mit der χ^2-Statistik geschätzten β-Parametern entspricht. Auf analoge Weise wurden die mittleren Parameter $\hat{\eta}_{\chi^2}$, $\hat{\beta}_{di}$ und $\hat{\eta}_{di}$ berechnet. Ist $\hat{\beta}_{\chi^2} > \hat{\beta}_{di}$ und $\hat{\eta}_{\chi^2} > \hat{\eta}_{di}$, so gilt für diese Stichprobe, daß die Diskrepanzstatistik die Fehlerwahrscheinlichkeiten im Vergleich zur χ^2-Methode unterschätzt hat und somit die Modellanpassung günstiger einschätzt.

Zur statistischen Überprüfung eines solchen Befundes wurden mit demselben Wissensraum 1000mal Daten simuliert und die Parameter mit beiden Methoden geschätzt. Die resultierenden 1000 $\hat{\beta}_{\chi^2}$ und $\hat{\beta}_{di}$ bzw. $\hat{\eta}_{\chi^2}$ und $\hat{\eta}_{di}$ wurden paarweise verglichen. Da die Verteilung dieser Größen a priori nicht bekannt ist, wurde zur statistischen Beurteilung der Unterschiede ein parameterfreies Verfahren, der Wilcoxon-Rangsummen-Test für verbundene Stichproben (vgl. Bortz, 1999), angewandt.

2.4 Anpassung von Wissensräumen an empirisch erhobene Antwortmuster

Der Datensatz wurde von Mag. Gudrun Wesiak von der Universität Graz zur Verfügung gestellt. Es handelt sich dabei um eine Untersuchung zum induktiven Denken (Wesiak & Albert, 2001). Der Test besteht aus 20 Items, die sich in vier Gruppen zu je fünf Items unterteilen lassen: verbale Analogien, figurale Analogien, verbale Reihenergänzung und figurale Reihenergänzung. Unter verbale Analogien fallen Items der folgenden Art:

> **Herz** zu **Pumpe** wie **Gehirn** zu ?
> a) denken b) Zentrale c) Verstand d) Kopf e) Nerven.

Dabei soll die Versuchsperson erkennen, daß für das Organ Herz eine Analogie im technischen Bereich (Pumpe) vorgegeben wurde. Somit muß sie eine entsprechende Analogie für das Organ Gehirn finden – in diesem Fall Zentrale. Bei der verbalen Reihenergänzung ist es die Aufgabe der Versuchsperson, Zahlenreihen fortzusetzen, wie zum Beispiel

> 25 23 20 18 15 ?
> a) 12 b) 17 c) 13 d) 14 e) 11.

Dabei muß die Versuchsperson die Regelmäßigkeit in der Reihe erkennen, nämlich daß zunächst 2, dann 3 abgezogen wird. Wenn sie diese Regel gefunden hat, sollte es ihr nicht schwer fallen, als richtige Lösung 13 auszuwählen. Abbildung 6 zeigt Beispiele für Items aus den Bereichen figurale Analogien bzw. figurale Reihenergänzung. Von der Versuchsperson werden ähnliche Fähigkeiten verlangt wie im verbalen Bereich. Im Falle der Analogien muß die Vp in diesem Beispiel erkennen, daß ein Rechteck eine Verallgemeinerung eines Quadrats ist. Überträgt sie dieses Prinzip auf die Ellipse, so sollte sie den Kreis (Antwort b) als richtige Antwort wählen. Bei der Reihenergänzung müssen mehrere Prinzipien der Reihenkonstruktion erkannt werden. Durch die richtige Kombination dieser Prinzipien, sollte man in diesem Beispiel zur Antwort d gelangen.

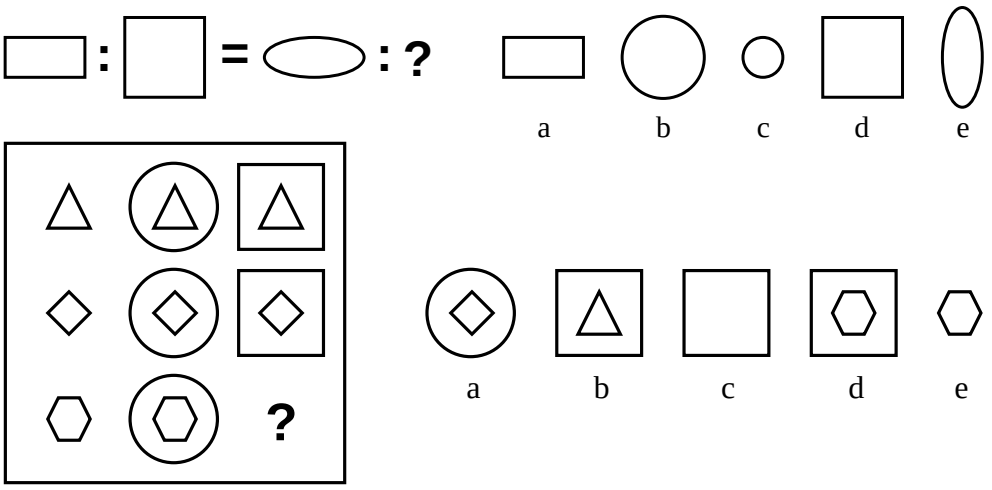

Abbildung 6: Beispielitems für figurale Analogien (oben) und figurale Reihenergänzung (unten). Die richtigen Lösungen sind Antwort b (oben) bzw. Antwort d (unten).

Mit diesem Test wurden 121 Probanden untersucht. Durch Item-Tree-Analyse (Leeuwe, 1974; Schrepp, 1999) wurde von Mag. Gudrun Wesiak ein Wissensraum gefunden, der den Abhängigkeitsbeziehungen zwischen den Items Rechnung trägt. Dieser besteht aus 438 Wissenszuständen, was in Anbetracht der $2^{20} = 1048576$ möglichen Antwortmuster eine beträchtliche Strukturierung darstellt. Die Basis dieses Wissensraumes hat 17 Elemente.

Zur Beurteilung der Anpassungsgüte des ITA-Wissensraumes an die Antworten der Versuchspersonen wurde versucht, eine probabilistische Wissensstruktur anzupassen und die χ^2-Statistik (Gleichung 4) zu ermitteln. Außerdem wurde der Diskrepanzindex (Gleichung 6) berechnet. Um die Aussage des Diskrepanzindexes statistisch zu untermauern, wurden Monte-Carlo-Studien (vgl. Lindgren, 1976; Fishman, 1973; Greene, 1990) durchgeführt. Dabei wurden für den angepaßten Wissensraum mit den Gleichungen 7, 8 und 9 die Zustands-, Fehler- und Ratewahrscheinlichkeiten berechnet. Dann wurden mit derselben Stichprobengröße von je 121 Versuchspersonen 1000mal Antworten simuliert. Anschließend wurde jeweils der Diskrepanzindex *di* ermittelt. Anhand der resultierenden Verteilung von *di* wurde überprüft, ob der empirisch ermittelte Diskrepanzindex oberhalb oder unterhalb des 95%-Perzentils lag. Ist der empirische Diskrepanzindex größer als das 95%-Perzentil, spricht dies für eine schlechte Anpassung

des Wissensraumes an die Daten – also für die Ablehnung des Modells.

In einem zweiten Schritt wurde versucht, für Teilmengen der 20 Items adäquate Wissensstrukturen zu finden; dabei dienten die fünf Items aus dem Bereich verbale Analogien als Untermenge. Sie bilden die Domain $Q = \{a, b, c, d, e\}$. Um einen geeigneten Wissensraum zu finden, wurde ein empirisches Vorgehen gewählt: Antwortmuster mit großer absoluter Häufigkeit deuten auf potentielle Wissenszustände hin. Ist die Häufigkeit größer als ein festgesetztes Kriterium, so wird das Antwortmuster als Wissenszustand übernommen:

$$K = R \Leftrightarrow N(R) \geq crit.$$

Die auf diese Weise gefundenen Zustände müssen noch keinen (für die Modellierung geeigneten) Wissensraum ergeben, der die Wellgradedness erfüllt. Dementsprechend müssen gegebenenfalls Zustände ergänzt werden. Untersucht man nur die ersten fünf Items des Datensatzes, so ergeben sich 28 verschiedene Antwortmuster von 32 möglichen. Acht dieser Muster hatten eine Häufigkeit größer oder gleich sechs. Setzt man das Kriterium $N(R) \geq 6$, so ergibt sich bereits ein Wissensraum $\mathcal{K}_{\geq 6}$, der die Wellgradedness erfüllt:

$$\mathcal{K}_{\geq 6} = \{\emptyset, \{a\}, \{b\}, \{a, b\}, \{a, c\}, \{a, b, c\}, \{a, b, c, e\}, Q\},$$

mit $|\mathcal{K}_{\geq 6}| = 8$. Zu allen acht Zuständen wurden mindestens sechs exakt passende Antworten gegeben. Setzt man das Kriterium $N(R) \geq 5$, so ergeben sich zehn Zustände, die noch keinen Wissensraum bilden, der wellgraded ist. Erst durch die Hinzunahme eines weiteren Zustands $\{b, e\}$ wird die Struktur ein Wissensraum $\mathcal{K}_{\geq 5}$, der die Wellgradedness erfüllt:

$$\mathcal{K}_{\geq 5} = \{\emptyset, \{a\}, \{b\}, \{a, b\}, \{a, c\}, [\{b, e\}], \{a, b, c\}, \{a, b, e\}, \{b, c, e\}, \{a, b, c, e\}, Q\},$$

mit $|\mathcal{K}_{\geq 5}| = 11$, wobei auf zehn der elf Zustände Antwortmuster mit einer empirischen Häufigkeit von mindestens fünf hindeuten. Die Abbildung 7 und 8 zeigen die Hassediagramme der beiden Wissensräume $\mathcal{K}_{\geq 6}$ und $\mathcal{K}_{\geq 5}$.

Um die Anpassung der beiden Wissensräume $\mathcal{K}_{\geq 6}$ und $\mathcal{K}_{\geq 5}$ an die Daten zu untersuchen, wurde für beide Räume das MM1 angepaßt und anschließend jeweils die

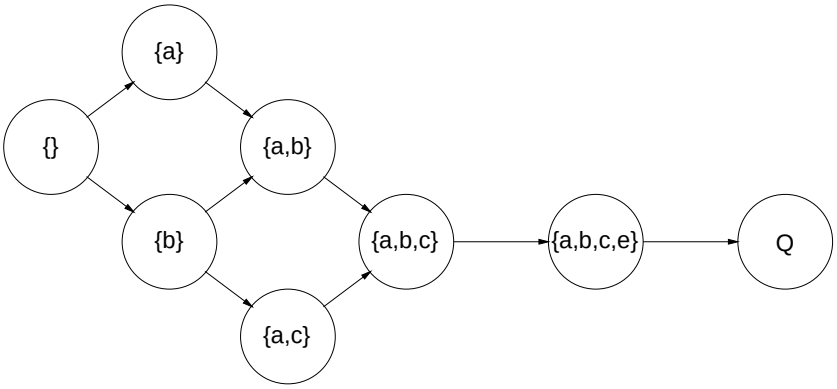

Abbildung 7: Hasse-Diagramm des Wissensraumes $\mathcal{K}_{\geq 6}$ für den Subtest Verbale Analogien. Jedem Zustand entspricht ein Antwortmuster mit einer empirischen Häufigkeit von mindestens 6.

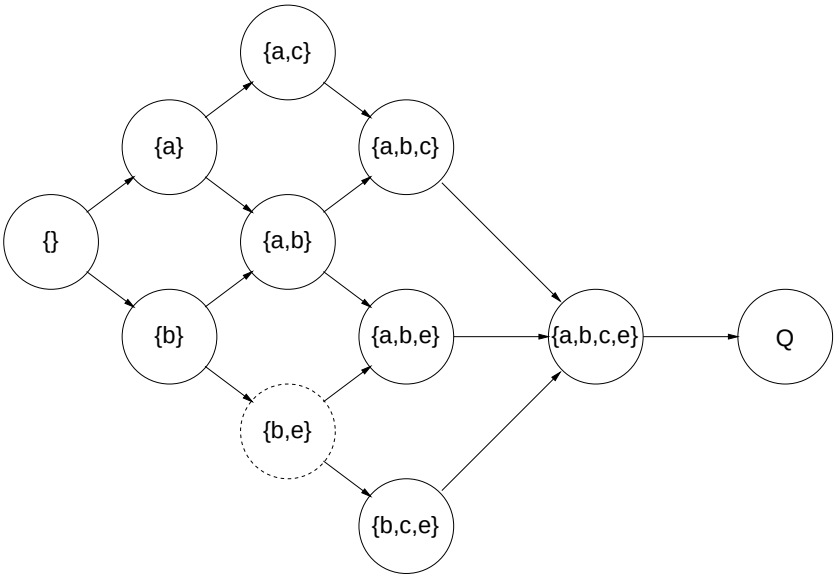

Abbildung 8: Hasse-Diagramm des Wissensraumes $\mathcal{K}_{\geq 5}$ für den Subtest Verbale Analogien. Jedem Zustand außer $\{b, e\}$ entspricht ein Antwortmuster mit einer empirischen Häufigkeit von mindestens 5.

χ^2-Größe (Gleichung 4) berechnet. Anhand der χ^2-Werte konnte entschieden werden, welcher Wissensraum die Daten besser beschreibt, nämlich der mit dem kleineren χ^2. Da die Verteilung der χ^2-Größe bekannt ist, konnte darüberhinaus eine statistische Entscheidung darüber getroffen werden, ob das jeweilige Modell (und damit der Wissensraum) abgelehnt werden muß oder beibehalten werden kann. Ferner wurden für $\mathcal{K}_{\geq 6}$ und $\mathcal{K}_{\geq 5}$ nach Gleichung 6 die Diskrepanzindizes ermittelt. Auch diese treffen eine Aussage darüber, welcher Wissensraum der adäquatere ist, derjenige mit dem kleineren *di*. Da $\mathcal{K}_{\geq 6}$ und $\mathcal{K}_{\geq 6}$ in Teilmengenrelation stehen und deshalb mit dem größeren Wissensraum $\mathcal{K}_{\geq 5}$ ein kleinerer *di* einhergeht, wurden die beiden Diskrepanzindizes an der Diskrepanz zwischen der Potenzmenge und dem jeweiligen Wissensraum normiert. Für den normierten Diskrepanzindex gilt

$$di_{norm}(\mathcal{K}) = \frac{di(\mathcal{R}, \mathcal{K})}{di(2^Q, \mathcal{K})}. \tag{10}$$

Wegen der Unkenntnis der Verteilung des Diskrepanzindexes wurden – wie schon beim ITA-Wissensraum – Computersimulationen als Hilfsmittel verwendet: Erst wurden Zustands-, Fehler- und Ratewahrscheinlichkeiten (Gleichungen 7, 8 und 9) berechnet. Mit diesen Werten wurden dann mit einem Stichprobenumfang von 121 Versuchspersonen 1000mal Antworten generiert. Schließlich wurde pro Simulationslauf ein *di* berechnet. Dieses Vorgehen wurde sowohl für $\mathcal{K}_{\geq 6}$ als auch für $\mathcal{K}_{\geq 5}$ durchgeführt. In beiden Fällen erhielt man eine Verteilung des Diskrepanzindexes. Durch Dichotomisierung dieser Verteilungen am 95%-Perzentil entstanden die jeweiligen Ablehnungsbereiche. Somit konnte auf diesem Wege eine statistische Entscheidung über die Ablehnung oder Beibehaltung der Wissensräume ermöglicht werden.

Dieses Vorgehen erlaubte einen Vergleich der probabilistischen und der deterministischen Methode in zweierlei Hinsicht: Erstens konnte überprüft werden, ob beide Methoden übereinstimmend denselben Wissensraum als den passenderen beurteilen, d. h. ob sie das jeweils kleinere Anpassungsmaß zuordnen. Zweitens konnte untersucht werden, ob beide Methoden auch bei der statistischen Entscheidung über Ablehnung oder Beibehaltung der Wissensräume zum gleichen Ergebnis kommen.

3 Ergebnisse

3.1 Ergebnisse der Parameterschätzung aus simulierten Antwortmustern

Die Parameterschätzer aus der Schätzroutine konnten die zur Simulation verwendeten Parameter des MM1 gut annähern. Zumindest bei großen Stichproben wurden meist sehr gute Resultate erzielt. Die Parameter des MM1 können damit als identifizierbar gelten.

Die am Wissensraum $\mathcal{M}$ (vgl. Abbildung 4, S. 31) gefundenen Ergebnisse sind in Tabelle 2 zusammengefaßt und gelten exemplarisch für Datensätze mit vergleichbarer Größe (hier $N = 100000$): Die erste Spalte der Tabelle gibt an, welcher Parameter (g, β oder η) geschätzt wurde. Der zweiten Spalte ist der wahre Wert zu entnehmen, auf den der entsprechende Parameter bei der Simulation gesetzt wurde. In der dritten Spalte steht der Startwert, mit dem das iterative Verfahren begann, ein Minimum für die χ^2-Größe zu finden. Die vierte Spalte führt schließlich die Parameterschätzungen auf. Für die Lernparameter g_a bis g_e konnten vom Schätzprogramm sehr gute Näherungen gefunden werden, die Abweichungen vom wahren Wert liegen im Tausendstelbereich. Auch die Fehler- und Rateparameter β_q und η_q wurden im allgemeinen gut angenähert. Allerdings kann nicht von einer perfekten Lösung gesprochen werden, da es bei jeder Parameterschätzung Ausreißer gab, die sich deutlich vom wahren Wert unterschieden – in diesem Fall bei β_d und bei β_e. Die Anpassung des Wissensraumes $\mathcal{W}$ an die simulierten Antwortmuster war zufriedenstellend ($\chi^2 = 24.584, df = 16, p = .079$).

3.2 Reaktion der Gütemaße auf Veränderungen des Wissensraumes

Tabelle 3 zeigt, wie sich unterschiedliche Formen der Wissensstruktur auf die χ^2-Größe bzw. den Diskrepanzindex auswirken. Die Datensätze $\mathcal{A}$ und $\mathcal{B}$ wurden jeweils mit demselben Wissensraum generiert, d. h. jeder Zeile von Tabelle 3 liegt derselbe Wissensraum

Parameter	Simulation	Startwert	Schätzung
g_a	.15	.1	.1493
g_b	.15	.1	.1489
g_c	.10	.1	.1018
g_d	.05	.1	.0452
g_e	.05	.1	.0537
β_a	.05	.01	.0466
β_b	.05	.01	.0469
β_c	.05	.01	.0586
β_d	.10	.01	.0000
β_e	.10	.01	.0175
η_a	.10	.01	.1057
η_b	.10	.01	.1029
η_c	.10	.01	.0960
η_d	.05	.01	.0484
η_e	.05	.01	.0503

Tabelle 2: Schätzungen der Übergangsparameter und der Fehler- und Ratewahrscheinlichkeiten beim Wissensraum $\mathcal{M}$ (vgl. Abbildung 4). Für die meisten Parameter gelang eine gute Annäherung. Ausreißer sind β_d und β_e.

zugrunde. Wegen der Zufallsmechanismen bei der Antwortgenerierung fallen die Anpassungsmaße für die Datensätze unterschiedlich aus. Erwartungsgemäß sollte dabei keine Rolle spielen, ob die Anpassung auf probabilistische (χ^2) oder deterministische (di) Weise berechnet wurde. Entgegen den Erwartungen konnten aber widersprüchliche Ergebnisse gefunden werden, d. h. einem kleineren $\chi^2(\vartheta; N(\mathcal{A}))$ entspricht ein größerer $di(\mathcal{A}, \mathcal{K})$ und umgekehrt. Die fettgedruckten Werte in Tabelle 3 zeigen diese Widersprüche an. Beispielsweise wurde bei der Wissensstruktur der Form helix mit 5 Items ein kleinerer χ^2-Wert (18.183) für den Datensatz $\mathcal{A}$ gefunden als für den Datensatz $\mathcal{B}$. Der Diskrepanzindex ist dagegen für den Datensatz $\mathcal{B}$ kleiner (0.110) als bei Datensatz $\mathcal{A}$. Diese Resultate scheinen weder systematisch von der Größe noch von der Form der zugrundeliegenden Wissensstruktur abzuhängen. In Tabelle 4 wird deutlich, daß auch bei großen Stichproben ($N = 100000$) die Frage nach der besseren Anpassung von beiden Maßen widersprüchlich beantwortet wird (siehe die fettgedruckten Werte). Beispielsweise kommt die χ^2-Methode beim Datensatz $\mathcal{B}$ zum besseren Ergebnis, während die Diskrepanzmethode beim Datensatz $\mathcal{A}$ besser ausfällt.

Um die Interpretation dieser deskriptiven Befunde zu erleichtern, wurden pro Wissensraum nicht nur zwei Datensätze ($\mathcal{A}$ und $\mathcal{B}$) generiert, sondern 1000 Datensätze, und jeweils das zugehörige χ^2-Maß und der Diskrepanzindex berechnet. Der Vergleich der Verteilung der χ^2-Maße mit der theoretischen χ^2-Verteilung (vgl. Csörgö & Faraway, 1996) zeigte, daß die Verteilungsannahme nur für große Stichprobenumfänge ($N = 100000$) gerechtfertigt ist. Bei den χ^2-Maßen, die aus Datensätzen mit kleinen Stichprobenumfängen berechnet wurden ($N = 1000$), kam es zu Verletzungen der Verteilungsannahmen.

Abbildung 9 zeigt die resultierenden Verteilungen exemplarisch für einen Wissensraum der Form ring mit 5 Items bei einem Stichprobenumfang von 100000 Versuchspersonenantworten; dies bedeutet, daß je einem Paar von χ^2- und Diskrepanzstatistik ein gemeinsamer Datensatz mit $N = 100000$ zugrunde liegt. Beide Verteilungen wurden am 95%-Perzentil dichotomisiert, größere Werte sprechen für die Ablehnung des Modells, kleinere für die Beibehaltung. Im vorliegenden Beispiel ergab sich bei der χ^2-Verteilung ein 95%-Perzentil von 28.53052, bei der di-Verteilung ein 95%-Perzentil von 0.14495.

Form	Größe	$\chi^2(\vartheta; N(\mathcal{A}))$	$\chi^2(\vartheta; N(\mathcal{B}))$	$di(\mathcal{A}, \mathcal{K})$	$di(\mathcal{B}, \mathcal{K})$
chain	3 Items	**0.046268**	2.129324	0.068	**0.057**
ring	3 Items	**0.074161**	1.662044	0.050	**0.045**
helix	3 Items	0.000000	0.198646	0.031	0.037
power	3 Items	0.157925	0.000000	0	0
chain	4 Items	5.202843	1.287320	0.119	0.113
ring	4 Items	2.399760	4.364908	0.078	0.084
helix	4 Items	1.298520	0.541226	0.071	0.055
power	4 Items	2.704338	7.553963	0	0
chain	5 Items	**11.931243**	13.208470	0.165	**0.156**
ring	5 Items	24.137315	15.544560	0.140	0.136
helix	5 Items	**18.183022**	21.979168	0.122	**0.110**
power	5 Items	23.423666	11.551135	0	0
chain	6 Items	31.749052	47.802238	0.197	0.213
ring	6 Items	36.308564	29.915716	0.202	0.200
helix	6 Items	51.194651	25.907621	0.163	0.160
power	6 Items	44.709842	46.798288	0	0

Tabelle 3: χ^2-Statistik und Diskrepanzindex in Abhängigkeit von Form und Größe der Wissensstruktur ($N = 1000$). Fettgedruckte Werte markieren einen Widerspruch der Anpassungsmaße.

Form	Größe	$\chi^2(\vartheta; N(\mathcal{A}))$	$\chi^2(\vartheta; N(\mathcal{B}))$	$di(\mathcal{A}, \mathcal{K})$	$di(\mathcal{B}, \mathcal{K})$
chain	5 Items	**14.758099**	15.354685	0.155960	**0.153980**
ring	5 Items	27.154207	10.003423	0.142430	0.141840
helix	5 Items	**14.906978**	19.821348	0.107420	**0.105330**
power	5 Items	24.693598	20.033786	0	0
chain	6 Items	54.684653	**53.841095**	**0.204420**	0.207050
ring	6 Items	52.671282	52.436355	0.191840	0.189920
helix	6 Items	**44.256936**	52.404960	0.156430	**0.154620**
power	6 Items	50.512498	21.893879	0	0

Tabelle 4: χ^2-Statistik und Diskrepanzindex in Abhängigkeit von Form und Größe der Wissensstruktur ($N = 100000$). Fettgedruckte Werte markieren einen Widerspruch der Anpassungsmaße.

Die gemeinsame Verteilung läßt sich in folgender Vierfeldertafel zusammenfassen:

		di		
		$\leq .14495$	$> .14495$	
χ^2	≤ 28.53052	904	46	950
	< 28.53052	46	4	50
		950	50	1000

Die beiden 95%-Perzentile teilen die Randverteilungen in je 950 Fälle im Annahme- und je 50 Fälle im Ablehnungsbereich. Bei 904 Datensätzen kamen beide Methoden zum Ergebnis, daß das Modell zu den Daten paßt, bei 4 Datensätzen lehnten beide Methoden übereinstimmend das Modell ab. In den verbleibenden 92 Fällen widersprachen sich χ^2 und di. Eine statistisch begründete Aussage über die Beziehung zwischen der χ^2-Statistik und dem Diskrepanzindex liefert der χ^2-Unabhängigkeitstest, der die bei Unabhängigkeit erwarteten Häufigkeiten mit den tatsächlich gefundenen vergleicht. Im vorliegenden Beispiel unterschieden sich die tatsächlich gefundenen Häufigkeiten kaum von den erwarteten. Somit konnte die Unabhängigkeitshypothese nicht abgelehnt werden ($\chi^2 = 0.997$, $df = 1$, $p = 0.318$).

Auch für die Formen chain und helix bzw. für Wissensräume mit fünf und sechs Items und für kleinere ($N = 1000$) sowie größere ($N = 100000$) Stichprobenumfänge wurden jeweils 1000 Datensätze simuliert und somit die Verteilungen der χ^2-Statistik

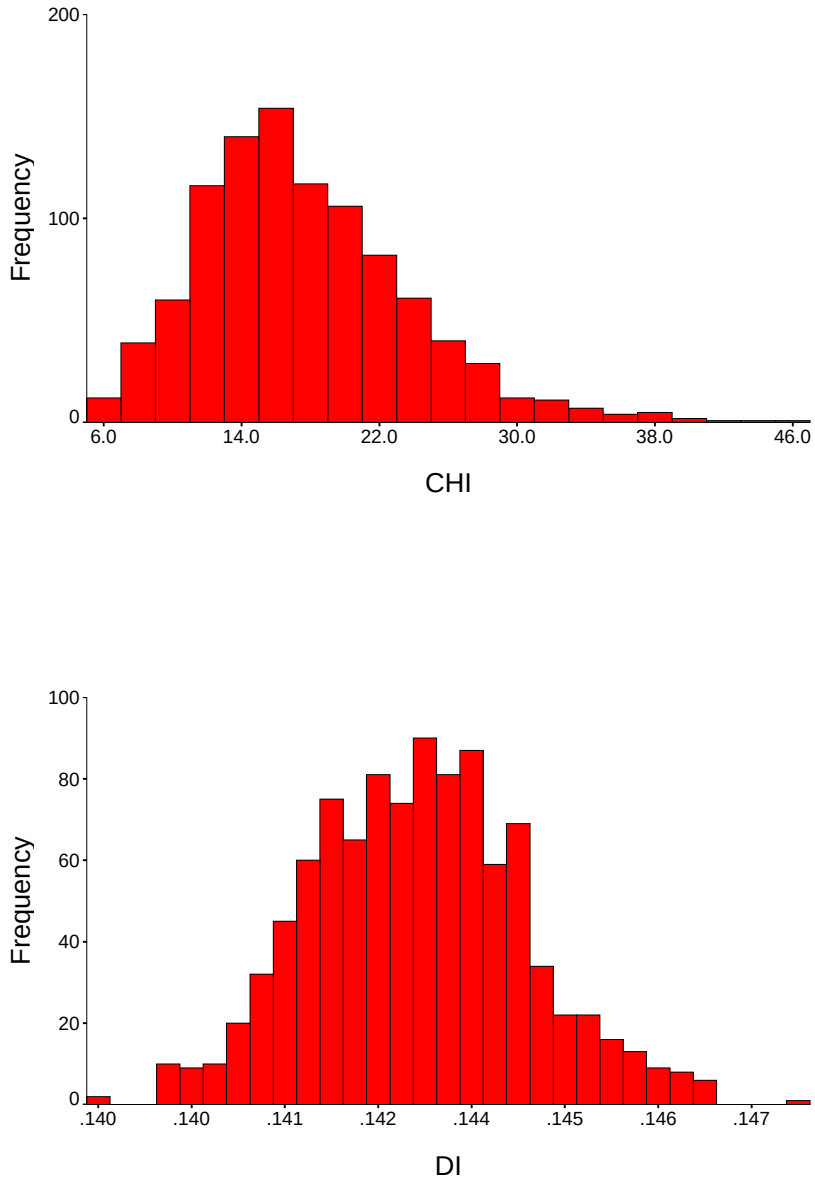

Abbildung 9: Verteilungen der χ^2-Größe (oben) und des Diskrepanzindexes (unten) bei einem Raum der Form ring mit 5 Items und SP-Umfang 100000; je 1000 χ^2-Maße und Diskrepanzindizes wurden berechnet.

Form	Größe	SP-Umfang	χ^2	p
chain	5 Items	1000	1.085	0.297
ring	5 Items	1000	0.165	0.684
helix	5 Items	1000	0.721	0.396
chain	6 Items	1000	0.997	0.318
ring	6 Items	1000	0.424	0.515
helix	6 Items	1000	1.498	0.221
chain	5 Items	100000	0.903	0.342
ring	5 Items	100000	0.997	0.318
helix	5 Items	100000	2.770	0.096
chain	6 Items	100000	5.694	0.017^*
ring	6 Items	100000	2.712	0.100
helix	6 Items	100000	0.137	0.712

Tabelle 5: Ergebnisse des χ^2-Unabhängigkeitstests bei Wissensstrukturen unterschiedlicher Größe und Form bei $N = 1000$ und $N = 100000$. Nur in einem Fall $(*)$ konnte die Unabhängigkeitshypothese verworfen werden.

und des Diskrepanzindexes ermittelt. Auch diese Daten wurden in Vierfeldertafeln zusammengefaßt, und die gemeinsame Verteilung wurde mit dem χ^2-Unabhängigkeitstest untersucht. Die Ergebnisse sind in Tabelle 5 zusammengefaßt. Lediglich in einem von zwölf Tests konnte die Unabhängigkeitshypothese verworfen werden. Dies entspricht bei einem α-Niveau von 5% in etwa der Menge der erwarteten zufälligen Signifikanzen.

3.3 Fehler- und Ratewahrscheinlichkeiten

Die Ergebnisse des Vergleichs der Schätzungen von Fehler- und Ratewahrscheinlichkeiten durch die beiden Methoden sind in den Tabellen 6 und 7 dargestellt. Jeder Tabellenzeile lag ein anderer Wissensraum zugrunde, als Resultat der Kombination von Form und Größe. Mit einem festen Wissensraum wurden pro Zeile 1000 Simulationen und Schätzungen mit beiden Methoden durchgeführt. Die Tabellen 6 und 7 zeigen

Raum	$\hat{\beta}_{\chi^2}$	$\hat{\beta}_{di}$	Z	$\hat{\eta}_{\chi^2}$	$\hat{\eta}_{di}$	Z
chain5	0.182	0.041	-26.461^*	0.090	0.025	-27.375^*
ring5	0.102	0.040	-24.635^*	0.057	0.029	-27.224^*
helix5	0.124	0.026	-26.066^*	0.086	0.016	-27.308^*
chain6	0.257	0.054	-27.281^*	0.092	0.029	-27.388^*
ring6	0.175	0.051	-26.120^*	0.054	0.033	-27.014^*
helix6	0.185	0.039	-27.019^*	0.082	0.021	-27.391^*

Tabelle 6: Mittlere Schätzungen der Fehler- und Rateparameter β und η mittels χ^2- bzw. *di*-Methode und Ergebnisse des Wilcoxontests für verbundene Stichproben ($N = 1000$). Die wahren Werte betrugen 0.05. Mit $*$ gekennzeichnete Werte haben p-values < 0.0001.

jeweils die Mittelwerte der geschätzten Parameter. Beim Wissensraum ring5 (Tabelle 6) betrug der Mittelwert $\hat{\beta}_{\chi^2}$ von 1000 mit der χ^2-Methode geschätzten Fehlerparametern 0.102. Der Mittelwert $\hat{\beta}_{di}$ der Fehlerparameter, die mit der Diskrepanzstatistik berechnet wurden, betrug lediglich 0.040. Nach der Teststatistik des Wilcoxontests entspricht diesem Unterschied der beiden Mittelwerte ein Z-Wert von -24.635. Damit ist der Unterschied statistisch bedeutsam ($p < 0.0001$). Für die Mittelwerte der geschätzten Rateparameter ergaben sich $\hat{\eta}_{\chi^2} = 0.057$ bei der χ^2-Methode bzw. $\hat{\eta}_{di} = 0.029$ bei der Diskrepanzstatistik. Auch dieser Mittelwertsunterschied ist statistisch bedeutsam ($Z = -27.224$, $p < 0.0001$). Jeder Simulation in Tabelle 6 lag ein Stichprobenumfang von $N = 1000$ zugrunde, jeder Simulation in Tabelle 7 ein Stichprobenumfang von $N = 100000$.

Für alle Paare von Parametern $\hat{\beta}_{\chi^2}$ und $\hat{\beta}_{di}$ bzw. $\hat{\eta}_{\chi^2}$ und $\hat{\eta}_{di}$ konnten signifikante Unterschiede gefunden werden. Dabei war der Gruppenmittelwert der mit der Diskrepanzstatistik berechneten Parameter stets kleiner als der Mittelwert der mit der χ^2-Methode geschätzten Parameter. Auch beim deutlich erhöhten Stichprobenumfang in Tabelle 7 blieben diese Unterschiede signifikant. Aufgrund dieser Ergebnisse kann man von einer Unterschätzung der Fehler- und Ratewahrscheinlichkeiten durch die Diskrepanzstatistik sprechen. Dies läßt sich vor allem dadurch begründen, daß die Dis-

Raum	$\hat{\beta}_{\chi^2}$	$\hat{\beta}_{di}$	Z	$\hat{\eta}_{\chi^2}$	$\hat{\eta}_{di}$	Z
chain5	0.092	0.041	-23.634^*	0.082	0.025	-27.393^*
ring5	0.051	0.039	-20.953^*	0.050	0.029	-27.393^*
helix5	0.082	0.026	-27.197^*	0.059	0.016	-27.393^*
chain6	0.131	0.053	-23.776^*	0.082	0.029	-27.393^*
ring6	0.074	0.050	-14.749^*	0.049	0.033	-27.393^*
helix6	0.094	0.038	-23.438^*	0.057	0.021	-27.393^*

Tabelle 7: Mittlere Schätzungen der Fehler- und Rateparameter β und η mittels χ^2- bzw. *di*-Methode und Ergebnisse des Wilcoxontests für verbundene Stichproben ($N = 100000$). Die wahren Werte betrugen 0.05. Mit $*$ gekennzeichnete Werte haben p-values < 0.0001.

krepanzstatistik nicht in der Lage ist, den einem Antwortmuster zugrundeliegenden wahren Zustand zu finden, der durch Fehler und Raten überlagert ist. Dagegen wird stets der nächstliegende Zustand ausgewählt; somit werden die Fehler- und Ratewahrscheinlichkeiten minimiert.

3.4 Ergebnisse der empirischen Anwendung

Die Ergebnisse der Untersuchung des von Mag. Gudrun Wesiak durchgeführten Tests zum induktiven Denken lassen sich in zwei Gruppen unterteilen. Zum einen wurde die Anpassung des von Frau Mag. Wesiak vorgeschlagenen ITA-Wissensraumes an den kompletten Datensatz mit 20 Items analysiert. Zum anderen wurden die ersten fünf Items des Tests für den Vergleich zweier Wissensräume verwendet, die aus empirischen Überlegungen heraus konstruiert wurden.

Analyse aller zwanzig Items

Alle zwanzig Items des Tests wurden von 121 Versuchspersonen bearbeitet. Von den insgesamt $2^{20} = 1048576$ Antwortmustern traten 120 verschiedene Muster empirisch auf. Mittels Item-Tree-Analyse entstand ein Wissensraum mit 438 Wissenszuständen. Für

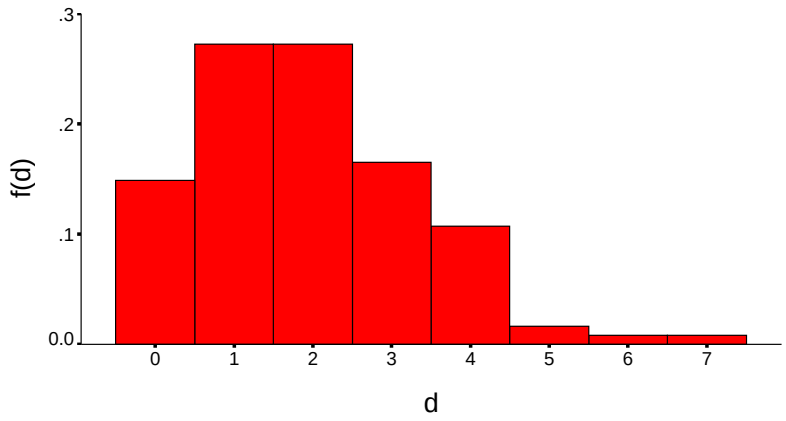

Abbildung 10: Diskrepanzfunktion des ITA-Wissensraumes und der 121 Versuchspersonenantworten. Der Erwartungswert *di* beträgt 1.934.

jedes der 120 Antwortmuster wurde die minimale Diskrepanz zum ITA-Raum berechnet. Die daraus resultierende Verteilung der minimalen Distanzen ist in Abbildung 10 dargestellt. Der Erwartungswert dieser Verteilung entspricht dem Diskrepanzindex des ITA-Raumes zu den empirisch gefundenen Antwortmustern. Er beträgt 1.934.

Um eine statistisch begründete Aussage über die Anpassung des ITA-Wissensraumes an die Daten treffen zu können, wurden wiederholt Computersimulationen durchgeführt. Für diese Monte-Carlo-Studien wurden die Zustands-, Rate- und Fehlerwahrscheinlichkeiten verwendet, die nach den Gleichungen 7, 8 und 9 berechnet wurden. Tabelle 8 faßt die Schätzungen für die je zwanzig β- und η-Parameter zusammen. Mit diesen Wahrscheinlichkeiten wurden 1000mal Antwortmuster simuliert und 1000 Diskrepanzindizes geschätzt. Abbildung 11 zeigt die resultierende Verteilung der Diskrepanzindizes. Minimum und Maximum betragen 1.231 bzw. 1.967. Der Mittelwert liegt bei 1.621. Damit liegt der empirisch gefundene Diskrepanzindex von 1.934 innerhalb der simulierten Verteilung. Da das 95%-Perzentil allerdings 1.794 beträgt, fällt der empirische Diskrepanzindex in den Ablehnungsbereich. Aus statistischer Sicht scheint demnach der ITA-Wissensraum nicht die geeignete Struktur zur Beschreibung der empirischen Antwortmuster zu sein.

Die Berechnung der χ^2-Statistik (Gleichung 4) erwies sich als unmöglich. Die

Parameter	Schätzung	Parameter	Schätzung
β_a	0.0965	η_a	0.0634
β_b	0.0309	η_b	0.0162
β_c	0.0087	η_c	0.1267
β_d	0.0525	η_d	0.1863
β_e	0.0485	η_e	0.2102
β_f	0.0158	η_f	0.0295
β_g	0.2003	η_g	0.0000
β_h	0.0996	η_h	1.0000
β_i	0.0472	η_i	0.0482
β_j	0.0238	η_j	0.0000
β_k	0.0622	η_k	0.0000
β_l	0.1197	η_l	0.0733
β_m	0.1983	η_m	0.0000
β_n	0.1512	η_n	0.0000
β_o	0.0331	η_o	0.1376
β_p	0.0606	η_p	0.0000
β_q	0.0000	η_q	1.0000
β_r	0.1203	η_r	0.0000
β_s	0.1068	η_s	0.1938
β_t	0.0000	η_t	0.1324

Tabelle 8: Schätzungen der Fehler- und Ratewahrscheinlichkeiten beim ITA-Wissensraum mit der Diskrepanzmethode.

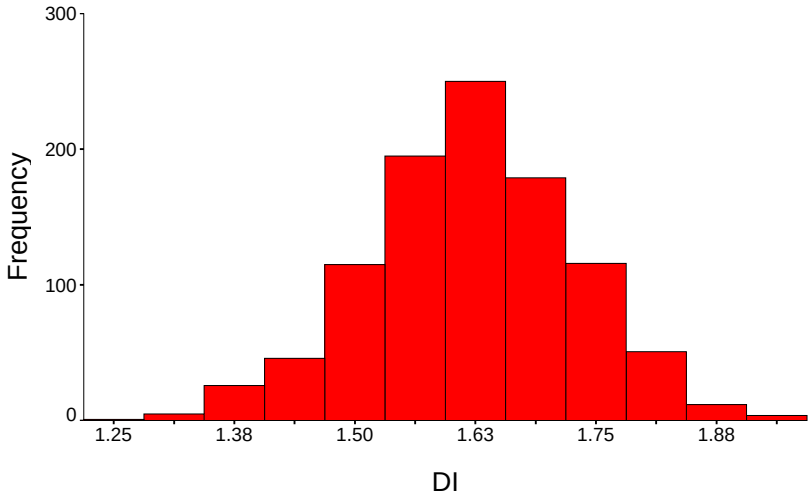

Abbildung 11: Verteilung der 1000 Diskrepanzindizes. Jedem Wert liegen virtuelle Antworten zugrunde, die mit denselben Parametern simuliert wurden. Das 95%-Perzentil beträgt 1.794.

Schätzroutine lieferte auch nach mehreren Tagen Rechenzeit kein Ergebnis. Dies ist zum einen auf die immense Zahl möglicher Antwortmuster (1048576) zurückzuführen, zum anderen auf den hochdimensionierten Parameterraum (31 Dimensionen).

Analyse der ersten fünf Items

Für den Subtest verbale Analogien, der aus fünf Items besteht, wurden zwei Wissensräume $\mathcal{K}_{\geq 6}$ und $\mathcal{K}_{\geq 5}$ untersucht. Diese entstanden aus empirischen Überlegungen heraus: Die Zustände in $\mathcal{K}_{\geq 6}$ werden durch Antwortmuster gestützt, deren absolute Häufigkeit mindestens 6 beträgt. Den Zuständen in $\mathcal{K}_{\geq 5}$ entsprechen Antwortmuster mit einer empirischen Häufigkeit von mindestens 5. Einzige Ausnahme ist der Zustand $\{b, e\}$, der in die Struktur aufgenommen wurde, um die Wellgradedness zu erfüllen. Für beide Wissensräume, die in Abbildung 7 bzw. 8 dargestellt sind, wurde das MM1 angepaßt und die χ^2-Größe berechnet. Die Parameter, die beide Modelle beschreiben, sind in den Parametervektoren $\vartheta_{\geq 6}$ bzw. $\vartheta_{\geq 5}$ zusammengefaßt. Das Anpassungsmaß $\chi^2(\vartheta_{\geq 6}; N(\mathcal{R}))$ betrug 28.038, während $\chi^2(\vartheta_{\geq 5}; N(\mathcal{R}))$ bei 22.809 lag. Dies bedeutet, daß das χ^2-Maß bezüglich der Anpassung eine Entscheidung zugunsten von $\mathcal{K}_{\geq 5}$ na-

helegt. Auch die Diskrepanzindizes wurden für beide Wissensräume berechnet. Beim Wissensraum $\mathcal{K}_{\geq 6}$ ergab sich ein Diskrepanzindex von $di(\mathcal{R}, \mathcal{K}_{\geq 6}) = 0.463$. Beim Raum $\mathcal{K}_{\geq 5}$ wurde ein Diskrepanzindex von $di(\mathcal{R}, \mathcal{K}_{\geq 5}) = 0.322$ ermittelt. Somit spricht auch die Diskrepanzmethode dem Wissensraum $\mathcal{K}_{\geq 5}$ eine bessere Anpassung an die Daten zu als dem Raum $\mathcal{K}_{\geq 6}$ (vgl. Tabelle 9). Die normierten Diskrepanzindizes (Gleichung 10) betrugen $di_{norm}(\mathcal{K}_{\geq 6}) = .478$ und $di_{norm}(\mathcal{K}_{\geq 5}) = .397$. Die Normierung veränderte also die Reihenfolge nicht.

Sowohl das deterministische als auch das probabilistische Anpassungsmaß bieten die Möglichkeit, zu einer statistischen Aussage über die Anpassung des Modells an die Daten zu gelangen. Im deterministischen Fall ist allerdings die Verteilung des Diskrepanzindexes a priori nicht bekannt. Daher wurden, um das 95%-Perzentil zu finden, Monte-Carlo-Studien durchgeführt. Abbildung 12 zeigt die Verteilungen der Diskrepanzindizes nach je 1000maliger Computersimulation für die beiden Wissensräume $\mathcal{K}_{\geq 6}$ und $\mathcal{K}_{\geq 5}$. Die Ablehnungsbereiche der Verteilungen werden durch die 95%-Perzentile markiert. Das 95%-Perzentil beim Raum $\mathcal{K}_{\geq 6}$ beträgt 0.413, und beim Raum $\mathcal{K}_{\geq 5}$ 0.281. Die beiden empirisch gefundenen Diskrepanzindizes liegen jenseits ihrer 95%-Perzentile, und somit im Ablehnungsbereich der Verteilungen. Die Monte-Carlo-Studien legen also den Schluß nahe, daß keiner der beiden Wissensräume die Daten hinreichend exakt beschreiben kann. Das probabilistische Anpassungsmaß hat den Vorteil, daß seine Verteilung a priori bekannt ist. D. h. das 95%-Perzentil kann der Tabelle der χ^2-Verteilung entnommen werden. Die Anzahl der Freiheitsgrade beträgt $\nu = (2^5 - 1) - (3 \cdot 5 + 1) = 15$ (Gleichung 5). Der kritische Wert, der von der χ^2-Verteilung mit 15 Freiheitsgraden 5% abschneidet beträgt 24.996. Die berechneten χ^2-Größen $\chi^2(\vartheta_{\geq 6}; N(\mathcal{R}))$ für $\mathcal{K}_{\geq 6}$ und $\chi^2(\vartheta_{\geq 5}; N(\mathcal{R}))$ für $\mathcal{K}_{\geq 5}$ liegen bei 28.038 bzw. 22.809. Dies bedeutet, daß aus statistischer Sicht das Modell $\mathcal{K}_{\geq 6}$ abgelehnt werden muß, während das Modell $\mathcal{K}_{\geq 5}$ beibehalten werden kann. Diese Aussage enthält mehr Information, da bei zwei Modellen nicht nur (wie oben) entschieden werden kann, welches die Daten besser beschreibt, sondern zusätzlich, ob die Modelle überhaupt adäquat für die Daten sind.

Tabelle 9 faßt die Ergebnisse der Untersuchung des Subtests verbale Analogien zusammen. Interpretiert man lediglich die Absolutbeträge der Anpassungsmaße, so

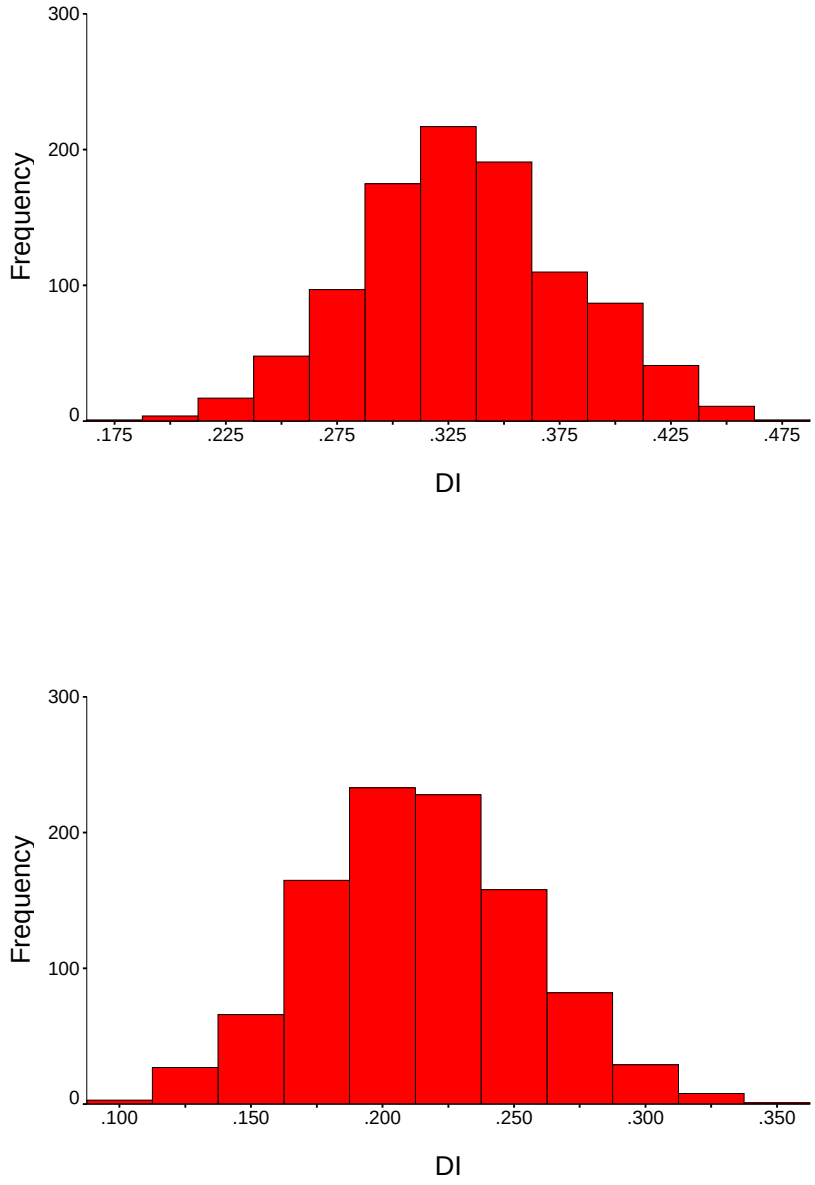

Abbildung 12: Verteilungen der 1000 Diskrepanzindizes, die mit den Parametern des Raumes $\mathcal{K}_{\geq 6}$ (oben) bzw. $\mathcal{K}_{\geq 5}$ (unten) berechnet wurden. Die 95%-Perzentile liegen bei 0.413 (oben) bzw. 0.281 (unten).

	$\mathcal{K}_{\geq 6}$	$\mathcal{K}_{\geq 5}$
$\lvert\mathcal{K}\rvert$	8	11
95%-Perzentil χ^2	24.996	24.996
$\chi^2(\vartheta; N(\mathcal{R}))$	28.038	22.809
95%-Perzentil di	0.413	0.281
$di(\mathcal{R}, \mathcal{K})$	0.463	0.322

Tabelle 9: Zusammenfassung der Ergebnisse der Analyse des Subtests verbale Analogien. Das probabilistische und das deterministische Anpassungsmaß weisen in dieselbe Richtung. Bei der statistischen Ablehnung widersprechen sich die Methoden.

kommen die deterministische und die probabilistische Methode zum gleichen Ergebnis. Beide bescheinigen $\mathcal{K}_{\geq 5}$ die bessere Anpassung an die Daten. In der statistischen Aussage unterscheiden sie sich jedoch: Die Diskrepanzmethode schlägt vor, beide Modelle als inadäquat abzulehnen. Nach der χ^2-Methode dagegen muß nur der Wissensraum $\mathcal{K}_{\geq 6}$ abgelehnt werden.

4 Diskussion

4.1 Interpretation der gefundenen Ergebnisse

Das MM1 als Modell für Wissensräume

Als erstes zentrales Ergebnis dieser Untersuchung bleibt festzuhalten, daß die Identifizierbarkeit der Parameter des einfachen Markoffmodells (MM1) von Doignon & Falmagne (1999) gegeben ist. Somit ist es geeignet ist für die Modellierung probabilistischer Wissensstrukturen. Es bietet die Möglichkeit einer drastischen Parameterreduktion. Allerdings wird diese Sparsamkeit mit mehreren Einschränkungen erkauft, wie die Ergebnisse in den Abschnitten 3.1 und 3.4 zeigen. Während besonders die Übergangsparameter g_q sehr gut geschätzt werden konnten, ergaben sich bereits bei den computersimulierten Antwortmustern für die Fehler- und Ratewahrscheinlichkeiten β_q und η_q mitunter deutliche Abweichungen von den wahren Werten (siehe Tabelle 2). Dies scheint ein allgemeines Problem des sehr sparsamen MM1 zu sein (vgl. Fries, 1997). Besonders schwierig zu interpretieren sind extreme Parameterwerte von 0 oder 1, da es sich um bedingte Wahrscheinlichkeiten handelt. So bedeutet $\beta_q = 0$, daß es unmöglich sei, bei Frage q einen Flüchtigkeitsfehler zu machen. Ein falsche Antwort bei q kann dann nur aufgrund von Nichtwissen der Lösung zustande kommen. Dagegen heißt $\eta_q = 1$, daß auch bei Nichtwissen der Antwort bei Item q immer die richtige Lösung geraten werde.

Die Interpretation der einzelnen Parameterschätzungen beim MM1 wird weiterhin erschwert durch die Abhängigkeit von der Lernschrittzahl n. Für unterschiedliche Werte von n sind verschiedene Parametervektoren ϑ denkbar, die die χ^2-Größe (Gleichung 4) minimieren. Die Parameterschätzungen bei Computersimulationen liegen nicht zuletzt deshalb so nahe am wahren Wert, weil die Lernschrittzahl n bekannt ist und auf den bekannten Wert festgesetzt wird. Bei empirischen Anwendungen ist die Lernschrittzahl a priori meist nicht bekannt. Daher ist es nur schwer möglich, sie auf einen festen Wert zu setzen; es sei denn, es gäbe begründete Hypothesen über den Lernprozeß in der Untersuchungspopulation. Alle Parameterschätzungen dürfen daher nur im Bezug auf

eine spezielle Lernschrittzahl interpretiert werden. Bei den empirischen Anwendungen hat sich gezeigt, daß der Minimierungsalgorithmus dazu tendiert, die Lernschrittzahl sukzessive höher zu setzen. Ein minimales χ^2 geht meist mit maximaler Lernschrittzahl einher.

Abgesehen von den Interpretationsschwierigkeiten, die von der Abhängigkeit der Parameter herrühren, machen es natürlich auch diverse Unabhängigkeitsannahmen schwer, die einzelnen Parameter empirisch zu deuten. Die Annahme der lokalen stochastischen Unabhängigkeit (Gleichung 2) fordert, daß die Parameter nicht vom Vorwissen einer Versuchsperson abhängen. Außerdem gelten die Parameter nur „im Durchschnitt“ für alle Mitglieder der Population. Zur Abschwächung der Unabhängigkeitsannahmen schlagen Doignon & Falmagne (1999) das MM2 vor. Zwar erscheint dieses realistischer, da das Vorwissen eines Probanden in gewisser Weise berücksichtigt wird, der Parameterzuwachs im Vergleich zum MM1 läßt sich aber kaum rechtfertigen. Es ist also angezeigt, die einzelnen Parameter nur sehr vorsichtig zu interpretieren.

Einfacher dagegen – und empirisch von größerer Relevanz – ist die Interpretation der prognostizierten Zustandswahrscheinlichkeiten. Für die individuelle Wissensdiagnose ist die Information über die A-priori-Wahrscheinlichkeiten von großer Bedeutung. Wie sich in den Computersimulationen zeigte, konnten diese durch das Markoffmodell recht gut vorhergesagt werden. Auch die Prognose der absoluten Häufigkeiten der einzelnen Antwortmuster konnte das Modell leisten. Insgesamt kann man also sagen, daß das MM1 trotz der Schwierigkeiten bei der Interpretation der Parameter, für empirische Zwecke geeignet ist.

Die Übereinstimmung des deterministischen und des probabilistischen Anpassungsmaßes

Das zweite Ergebnis dieser Untersuchung ist der Befund, daß Variationen von Form und Größe eines Wissensraumes keinen eindeutigen Schluß darüber zulassen, wie sich das Anpassungsmaß der einen oder anderen Art verändern wird. Die deskriptiven Resultate in den Tabellen 3 und 4 deuten bereits an, daß Widerspruch oder Übereinstimmung zwischen dem deterministischen und dem probabilistischen Anpassungsmaß

eher vom Zufall als von Merkmalen der Wissensstruktur abhängen. Lediglich bei Potenzmengenstrukturen (vgl. die Form power in Abbildung 5) zeigt sich ein vorhersagbarer Unterschied: während das χ^2-Maß in der Lage ist, zwischen unterschiedlichen Datensätzen bezüglich der Anpassung deutlich zu differenzieren, kann dies der Diskrepanzindex nicht leisten. Egal welche Antworten von Versuchspersonen gegeben wurden, ein Wissensraum mit Potenzmengenstruktur paßt immer perfekt darauf ($di = 0$). Dabei ist allerdings zu beachten, daß die Diskrepanzstatistik lediglich den Abstand zwischen Antwortmuster und Zustand minimiert. Somit wird stets der nächstliegende Zustand ausgewählt und nicht der wahre Zustand, der der Antwort zugrunde liegt, aber eine größere Distanz aufweist. Freilich ist die Potenzmenge in keinem Fall eine wünschenswerte Wissensstruktur, da sie keine Annahmen über Abhängigkeitsbeziehungen zwischen Items macht. Aber in vorsichtiger Verallgemeinerung der Ergebnisse kann man doch sagen, daß der Diskrepanzindex umso schlechter zwischen Wissensräumen unterscheiden kann, je mehr Zustände diese haben.

Diese deskriptiven Ergebnisse regten die Frage an, wie eng denn überhaupt die Beziehung zwischen dem deterministischen und dem probabilistischen Anpassungsmaß ist. Dies führte zum wohl kontraintuitivsten Resultat dieser Analyse: Es konnte die Hypothese nicht verworfen werden, daß beide Maße über die Anpassung von Wissensräumen Aussagen machen, die unabhängig voneinander sind. Dies galt abermals für Wissensstrukturen von unterschiedlicher Form und Größe. Dies bedeutet, daß es nicht gerechtfertigt ist, davon auszugehen, daß beide Methoden dasselbe Ergebnis liefern. Statistisch argumentiert, bringt es keinen Informationsgewinn, zu wissen, ob sich eine Methode für oder gegen den Wissensraum ausgesprochen hat (vgl. dazu die Vierfeldertafel auf Seite 45). Bei den computersimulierten Antwortmustern konnte trotzdem eine hohe Übereinstimmung in den Vierfeldertafeln von über 90% gefunden werden. Wichtig ist dabei aber zu beachten, daß diese Übereinstimmung auch bei Unabhängigkeit zu erwarten ist. Bei empirisch erhobenen Antwortmustern muß die Übereinstimmung keineswegs dieselbe Größenordnung besitzen. Gelangt ein Untersucher mittels des einen Anpassungsmaßes zum Ergebnis, daß sein Modell für die Daten inadäquat sei, muß dies für die andere Methode nicht gelten. Als wesentliches Fazit bleibt festzuhalten, daß eine Übereinstimmung zwischen der Aussage der χ^2-Methode und der Diskrepanzmethode zufällig ist.

Welche Methode ist die bessere?

Da die Computersimulationen zeigen konnten, daß man keineswegs von der Aussage der einen Anpassungsmethode auf die der anderen schließen kann, stellt sich die Frage nach dem Grund ihrer Unterschiedlichkeit und danach, welche Methode vorzuziehen ist. Wie die Ergebnisse der Tabellen 6 und 7 nahelegen, kann ein Teil der Unterschiedlichkeit in den Aussagen der Methoden darauf zurückgeführt werden, daß sie sich deutlich in den Schätzungen für die β- und η-Parameter unterscheiden. Die Analyse der computersimulierten Antwortmuster konnte offenlegen, daß diese Unterschiede systematisch sind: Die deterministische Methode schätzte konsequent bei allen untersuchten Wissensräumen die Fehler- und Ratewahrscheinlichkeiten geringer ein als die probabilistische Methode. Die Unterschiede sind hoch signifikant und teilweise so deutlich, daß die Grenzen der Unterscheidungsfähigkeit der Teststatistik erreicht werden und Deckeneffekte auftreten. So scheint es zunächst verwunderlich, daß in Tabelle 7 für alle Differenzen der η-Parameter dieselbe Teststatistik berechnet wurde. Dies ist allerdings auf die Methode zurückzuführen, durch die die Teststatistik zustande kommt (vgl. Bortz, 1999). Beim Wilcoxontest für verbundene Stichproben werden, wie beim t-Test Differenzen zwischen den zu vergleichenden Größen berechnet. Diese Differenzen werden in eine Rangreihe gebracht. Die Rangsumme der Differenzen mit dem selteneren Vorzeichen wird mit T bezeichnet. Die Rangsumme der Differenzen mit dem häufigeren Vorzeichen heißt T'. Ferner gilt die Beziehung

$$T + T' = \frac{n(n+1)}{2}, \tag{11}$$

wobei n die Stichprobengröße bezeichnet, in diesem Fall 1000. Bei den η-Parametern in Tabelle 7 sind die mit der Diskrepanzmethode berechneten $\hat{\eta}_{di}$ stets kleiner als die mit der χ^2-Methode berechneten $\hat{\eta}_{\chi^2}$. Demzufolge kommt die Differenz mit dem selteneren Vorzeichen nie vor ($T = 0$). Also wird T' aus Gleichung 11 bestimmt: $T' = (1000 \cdot 1001)/2 = 500500$. Unter Annahme der H_0, daß sich *di*- und χ^2-Parameter nicht unterscheiden, beträgt der Erwartungswert für T

$$\mu_T = \frac{n(n+1)}{4},$$

hier beträgt $\mu_T = 250250$. Für Werte von $n < 25$ wird die Verteilung der T-Statistik durch die Standardnormalverteilung approximiert, d. h. die T-Werte müssen mittels der Transformation

$$Z = \frac{T - \mu_T}{\sigma_T}$$

umgerechnet werden. Falls keine Rangbindungen auftreten, gilt für die Varianz von Z

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

In unserem Fall ergibt sich demnach ein Z-Wert von $Z = (0 - 250250)/9135.555 = -27.393$. Man kann sehen, daß dieser Wert immer dann zustande kommen muß, wenn alle Werte der einen Gruppe kleiner sind als die der anderen, woraus $T = 0$ folgt. Außerdem dürfen keine Rangbindungen auftreten. Ist dies gegeben, so hängt die Teststatistik nur noch von der Stichprobengröße n ab. Für $n = 1000$ ergibt sich ein Minimum von -27.939, das nicht unterschritten werden kann. Bei den η-Parametern in Tabelle 7 wird dieses Minimum für alle untersuchten Wissensräume erreicht. Alle anderen Z-Werte in den Tabellen 6 und 7 liegen mehr oder weniger nahe am Minimum. Die inhaltliche Interpretation dieses Verhaltens ist, daß sich die Fehler- und Ratewahrscheinlichkeiten der beiden Methoden sehr stark unterscheiden. Die Diskrepanzmethode schätzt diese Wahrscheinlichkeiten wesentlich geringer als die χ^2-Statistik. Demzufolge wird auch die Anpassung des Modells von der Diskrepanzstatistik günstiger eingeschätzt als von der χ^2-Statistik.

Weniger eindeutig zu beantworten ist die Frage, welche Methode die besseren Parameterschätzungen liefert. Die wahren Werte der geschätzten Parameter in den Tabellen 6 und 7 lagen immer bei 0.05. Betrachtet man nur die absoluten Abweichungen von diesem Wert, stellt man fest, daß bei kleinem Stichprobenumfang ($N = 1000$; Tabelle 6) die Fehlerparameter $\hat{\beta}_{\chi^2}$ deutlich weiter von 0.05 entfernt sind als die $\hat{\beta}_{di}$ der Diskrepanzmethode. Bei den Rateparametern dagegen ist der Betrag der Abweichungen in etwa gleich: so stark, wie von der Diskrepanzmethode die Ratewahrscheinlichkeit unterschätzt wird, wird sie von der χ^2-Methode überschätzt. Bei großem Stichprobenumfang ($N = 100000$; Tabelle 7) verändert sich das Verhältnis zugunsten der χ^2-Methode. Die Fehlerparameter $\hat{\beta}_{\chi^2}$ werden nicht mehr so stark überschätzt wie bei den

kleineren Stichproben. Bei der Schätzung der Rateparameter ist die χ^2-Methode der Diskrepanzmethode sogar überlegen.

Aus dem Vergleich der Ergebnisse in den Tabellen 6 und 7 lassen sich zwei Einflußfaktoren identifizieren, die auf die Güte der Parameterschätzungen einwirken. Die Stichprobengröße wirkt besonders auf die Parameterschätzungen der χ^2-Methode: für größeres N liegen die Schätzungen näher am wahren Wert, bei kleinem N ist die absolute Abweichung relativ groß. Die Schätzungen der Diskrepanzmethode werden allerdings durch die Stichprobengröße kaum beeinflußt. Die Veränderungen der Parameter sind meist so geringfügig, daß sie sich erst in der vierten Nachkommastelle manifestieren. Der zweite Einflußfaktor ist der Quotient aus der Anzahl der Zustände im Wissensraum und der Anzahl potentiell möglichen Zustände $|\mathcal{K}|/2^{|Q|}$. Je kleiner dieser Quotient ist, desto besser fallen die Schätzungen der Diskrepanzmethode aus. Bei Potenzmengenstrukturen gilt $|\mathcal{K}|/2^{|Q|} = 1 = max$, und die Fehlerwahrscheinlichkeiten werden maximal unterschätzt ($\hat{\beta}_{di} = \hat{\eta}_q = 0$). Die Ergebnisse in Abschnitt 3.3 deuten aber auch darauf hin, daß die Parameterschätzungen der χ^2-Methode besser werden, je größer der Quotient $|\mathcal{K}|/2^{|Q|}$ ist.

Eine andere Bewertung der deterministischen und der probabilistischen Methode ergibt sich, wenn man nicht nur den Betrag der Abweichung der Schätzung vom wahren Wert, sondern auch die Richtung der Abweichung in die Beurteilung einbezieht. Es hat sich fast ausschließlich gezeigt, daß die χ^2-Methode ein sehr konservatives Verfahren der Parameterschätzung ist, das Fehler- und Rateparameter tendenziell zu hoch schätzt. Die Diskrepanzmethode dagegen erwies sich als sehr liberal, was in der Unterschätzung der Fehlerwahrscheinlichkeiten deutlich wurde. Dies bedeutet, daß man tendenziell mehr Fehler und Raten prognostiziert, verläßt man sich auf die Parameterschätzungen der χ^2-Methode; die Diskrepanzmethode zeichnet dagegen ein bezüglich der Fehler- und Ratewahrscheinlichkeiten eher idealisiertes Bild.

Wissensstrukturen für empirische Antwortmuster

Die Untersuchung von empirischen Antwortmustern konnte zeigen, daß es keine Schwierigkeiten bereitet, deterministische Wissensstrukturen auch bei größeren Wissensberei-

chen anzuwenden. Freilich ist die Diskrepanz zwischen Daten und Wissensraum deutlich höher als bei computersimulierten Antwortmustern. Die Diskrepanzfunktion (vgl. Abbildung 10) ordnet dementsprechend Werten von $d > 0$ größere Häufigkeiten zu als $d = 0$. Das bedeutet, daß die perfekte Übereinstimmung zwischen Antwortmuster und Wissenszustand im Vergleich zu den Computersimulationen seltener geworden ist. Der Erwartungswert der Diskrepanzfunktion *di* liegt mit 1.934 über den Diskrepanzindizes der simulierten Antwortmuster. Die Schätzungen für Fehler- und Rateparameter in Tabelle 8 sollten vorsichtig interpretiert werden. Insgesamt sollte man eher von einer Unterschätzung dieser Wahrscheinlichkeiten ausgehen. Nichtsdestoweniger ergaben sich bei zwei geschätzten η-Parametern Werte von 1, die zusammen mit den 0-Werten die Interpretation erschweren. Als Hinweis darauf, in welchem Ausmaß in der Population mit Flüchtigkeitsfehlern bzw. Raten gerechnet werden muß, können die Parameterschätzungen dennoch nützlich sein.

Die unterschätzten Fehlerwahrscheinlichkeiten sollten zum Teil erklären, warum der empirische Diskrepanzindex von 1.934 im Ablehnungsbereich der simulierten Verteilung (Abbildung 11) lag. Je kleiner Fehler- und Rateparameter sind, desto besser sollte die Anpassung beurteilt werden, d. h. desto kleiner wird der Diskrepanzindex. Der Grund dafür liegt im Mechanismus der Monte-Carlo-Methode: Zunächst liegen empirisch erhobene Antwortmuster vor, und ein Modell wird spezifiziert und der empirische Diskrepanzindex berechnet. Dann werden die Parameter des Modells geschätzt. Mit dem Modell und den geschätzten Parametern werden wiederholt Daten generiert und die Diskrepanz zum Modell berechnet, so daß eine ganze Verteilung von Diskrepanzindizes entsteht. Dabei kommt es allerdings zur Konfundierung der Validität der Struktur und der Güte des Parameterschätzvervahrens. Liegt nun der empirische *di* in den Extrembereichen der Verteilung, bedeutet das, daß das Modell zusammen mit den Parameterschätzungen nicht in der Lage ist, Antwortmuster zu erzeugen, die den empirisch Antwortmustern ähnlich sind. Die eine Interpretationsmöglichkeit dieses Befundes ist, daß das Modell inadäquat für die Daten ist; es muß verworfen werden. Die andere Interpretationsmöglichkeit ist, daß die Parameterschätzungen zu ungenau sind und eine Verzerrung bei der Ermittelung der Diskrepanzindizes bewirken. Im Falle des ITA-Wissensraumes scheint letzteres der Fall zu sein: Die Unterschätzung der

Fehlerwahrscheinlichkeiten bewirkt, daß Antwortmuster generiert werden, die besser zum Modell passen als die empirischen Antwortmuster; deshalb sind die „simulierten“ Diskrepanzindizes kleiner. Die Verteilung ist bezüglich des empirischen *di* nach links verschoben. Leider läßt sich aber der Einfluß der Parameterschätzungen auf die Verschiebung der Verteilung nicht vom Einfluß des Modells isolieren. Man sollte daher nicht vorschnell zum Schluß gelangen, das gefundene Modell müsse abgelehnt werden.

Bei der empirischen Anwendung wurde ferner deutlich, daß die Modellierung mit probabilistischen Wissensstrukturen noch mit rein technischen Problemen zu kämpfen hat. Mit anwachsendem Antwort- und Parameterraum wird es zunehmend schwierig für einen Suchalgorithmus, ein Minimum für die χ^2-Größe (Gleichung 4) zu finden. Dies bedeutet, daß die Größe der Domain Q festlegt, ob probabilistische Wissensstrukturen überhaupt für die Modellierung in Frage kommen.

Bei kleineren Wissensbereichen dagegen stehen sowohl die deterministische als auch die probabilistische Methode zur Verfügung. Stehen mehrere Wissensräume zur Auswahl, die zur Modellierung der Daten in Frage kommen, so treffen beide Methoden zwei Aussagen über die Anpassung: Die erste Aussage ist rein deskriptiver Natur und ermöglicht nur die relative Beurteilung eines Modells. Von allen potentiellen Modellen ist dasjenige am besten geeignet, das zum kleinsten Anpassungsmaß führt, sei die Wissensstruktur deterministisch oder probabilistisch. Die zweite Aussage ist statistischer Natur und erlaubt für jedes einzelne Modell eine absolute Beurteilung. Ein Modell muß verworfen werden, wenn seine theoretischen Vorhersagen zu stark von den empirischen Befunden divergieren. Bei der χ^2-Methode läßt sich diese Divergenz durch eine Teststatistik ermitteln, deren Verteilung bekannt ist. Bei der Diskrepanzmethode kann die Divergenz nur mittels Computersimulationen abgeschätzt werden. Tabelle 9 zeigt, daß in der deskriptiven, relativen Aussage eine Übereinstimmung der beiden Methoden bei empirischen Antwortmustern gefunden wurde. Den Ergebnissen an computersimulierten Antwortmustern in Abschnitt 3.2 zufolge, ist dies kein zwangsläufiges Resultat, das etwa auf eine monotone Beziehung zwischen deterministischem und probabilistischem Anpassungsmaß hindeutet. Hinsichtlich der statistischen Aussage kommen die beiden Methoden nicht zum selben Ergebnis. Dennoch können die statistischen Resultate zei-

gen, daß beide Methoden in ihren Urteilen konsistent sind. Den berechneten χ^2-Werten von 28.038 und 22.809 entsprechen p-values von .023 bzw. .088. Dies bedeutet, daß die statistische Bewertung in dieselbe Richtung tendiert wie die deskriptive und überdies noch zur Ablehnung des Modells $\mathcal{K}_{\geq 6}$ rät. Auch die Diskrepanzmethode ist intern konsistent. Der empirische Diskrepanzindex des Raumes $\mathcal{K}_{\geq 6}$ von 0.463 entspricht exakt dem Maximum (also dem 1000sten Wert) der simulierten *di*-Verteilung; i. e. einem p-value von .001. Demgegenüber gehört zum *di* von 0.322 bei $\mathcal{K}_{\geq 5}$ ein p-value von .005. D. h. auch statistisch wird $\mathcal{K}_{\geq 5}$ von der Diskrepanzmethode als das bessere Modell beurteilt. Bei einem Signifikanzniveau von $p = .05$ fallen diese subtilen Unterschiede allerdings nicht auf.

Eine Schlußfolgerung, die aus der empirischen Anwendung gezogen werden kann ist, daß tatsächlich mit der Diskrepanzmethode leichter eine zufriedenstellende Anpassung erreicht werden kann; nämlich dann, wenn aufgrund der Größe des Wissensbereiches eine probabilistische Modellierung nicht in Frage kommt. Zieht man dagegen Monte-Carlo-Studien heran, um die Anpassung der deterministischen Struktur statistisch zu überprüfen, so erwies sich dieser Test als extrem konservativ. Keines der empirischen Modelle konnte beibehalten werden. Die zu niedrigen Parameterschätzungen können eine Erklärung dafür sein.

4.2 Erkenntnisse für die praktische Anwendung

In Zusammenfassung der Erkenntnisse, die in dieser Untersuchung aus der Analyse sowohl computersimulierter als auch empirisch gefundener Antwortmuster gewonnen werden konnten, lassen sich auch einige anwendungsorientierte Aspekte herausarbeiten. Der Ausgangspunkt einer Diagnose auf der Basis von Wissensräumen ist immer das Datenmaterial. Der Anwender steht vor der Frage, welcher Wissensraum für die erhobenen Daten geeignet ist. Der erste Hinweis auf die Datenstruktur sind die empirischen Häufigkeiten der Antwortmuster. Große Häufigkeiten deuten darauf hin, daß dem Antwortmuster ein Wissenszustand zugrunde liegen könnte, während kleine Häufigkeiten eher ein Indiz für Flüchtigkeitsfehler und Raten sind. Mit den Antwortmustern mit großen Häufigkeiten und zusätzlichen Zuständen, die die Wellgradedness sicherstellen, sollten für den Modellierer bereits potentielle Wissensräume entstehen. Diese müssen in einem zweiten Schritt darauf geprüft werden, wie gut sie zu den Daten passen.

Der nächste entscheidende Faktor ist die Mächtigkeit der Domain Q. Diese bestimmt, ob eine probabilistische Modellierung (aus technischen Gründen) möglich ist. Wenn ja, sollte sie in jedem Fall durchgeführt werden, wenn nein muß eine deterministische Wissensstruktur angepaßt werden. Bei beiden Methoden steht an nächster Stelle die Prüfung der Anpassung der Modelle. Dabei kann aus den in Frage kommenden Wissensräumen der beste ausgewählt werden. Außerdem muß darüber entschieden werden, welches Modell statistisch zu verwerfen ist. Bei probabilistischen Modellen kann dabei das von Doignon & Falmagne (1999) vorgeschlagene Signifikanzniveau von 5% verwendet werden. Bei deterministischen Modellen und den damit verbundenen Monte-Carlo-Techniken zur statistischen Prüfung ist dies sicher zu konservativ. Verbunden mit der Prüfung der Modelle ist die Schätzung der Fehler- und Rateparameter. Dabei sind zwei Werte besonders kritisch: die Stichprobengröße und der Quotient $|\mathcal{K}|/2^{|Q|}$. Die Computersimulationen konnten offenlegen, daß die Parameterschätzungen der χ^2-Methode für kleine Stichproben deutlich von den wahren Werten divergieren. Auf die Diskrepanzmethode hat die Stichprobengröße einen geringeren Einfluß. Nahm der Quotient $|\mathcal{K}|/2^{|Q|}$ kleine Werte an, so verbesserten sich die Schätzungen der Diskrepanzmethode. Die Schätzer der χ^2-Methode wurden dagegen bei größerem $|\mathcal{K}|/2^{|Q|}$ exakter. Auch

diese beiden Größen sollten also die Wahl der Modellierungsart mitbestimmen.

Das optimale Vorgehen wäre, sofern es die Ressourcen zulassen, sowohl die deterministische also auch die probabilistische Modellierung durchzuführen. Dies konnte an unterschiedlichen Stellen gezeigt werden. Die Übereinstimmung zwischen beiden Anpassungsmaßen ist keineswegs so gut, daß man etwa vom einen auf das andere schließen könnte (vgl. Abschnitt 3.2). Die Parameterschätzer reagieren unterschiedlich auf exogene Bedingungen (vgl. Abschnitt 3.3). Die statistische Ablehnung durch die Diskrepanzmethode ist auf dem 5%-Niveau kaum aussagekräftig (vgl. Abschnitt 3.4). Bei gleichzeitiger deterministischer und probabilistischer Modellierung kann eine weitaus bessere Bestätigung der Adäquatheit eines Modells gefunden werden. Ferner wird dadurch die Zuverlässigkeit der Parameterschätzungen gesteigert. Die wahren Werte befinden sich in einem Konfidenzintervall. Die Schätzungen der beiden Methoden liegen nahe an dessen Grenzen.

Zusammenfassung

Wissensstrukturen (Doignon & Falmagne, 1985, 1999) bieten einen formalen Rahmen zur Erfassung und effizienten Diagnose des Wissens von Probanden in einem bestimmten Wissensbereich. Eine Wissensstruktur beschreibt dabei die für eine Menge von Aufgaben möglichen Antwortmuster. In der vorliegenden Arbeit werden die mit einer Anwendung dieser Theorie verbundenen Probleme erörtert. Für eine wichtige Klasse probabilistischer Wissensstrukturen wird gezeigt, daß sich eine die gegebenen Daten adäquat beschreibende Wissensstruktur nicht über die Restriktion einer allgemeineren Struktur im Rahmen eines Nested-Models-Ansatzes bestimmen läßt. Für deterministische und probabilistische Wissensstrukturen werden die in der Literatur zur Überprüfung ihrer empirischen Validität vorgeschlagenen Anpassungsmaße systematisch untersucht. Zur Güte des häufig verwendeten Diskrepanzindexes, eines deterministischen Anpassungsmaßes, liegen bislang keine weitergehenden Studien vor. In dieser Arbeit werden die der Berechnung des Diskrepanzindexes unterliegenden Annahmen explizit formuliert. Konkrete Hypothesen zu dessen Eigenschaften werden sowohl an computersimulierten, wie auch an empirisch erhobenen Antwortmustern überprüft.

Die Ergebnisse legen einen differenzierten Umgang mit den untersuchten Anpassungsmaßen nahe. Es zeigte sich, daß zwischen der χ^2-Statistik des χ^2-Anpassungstests und dem Diskrepanzindex keine monotone Beziehung besteht. Es konnte nachgewiesen werden, daß die Verwendung des Diskrepanzindexes zu einer systematischen Überschätzung der Güte der Anpassung einer Wissensstruktur führt. Eine Methode zur statistischen Überprüfung des Diskrepanzindexes wurde vorgeschlagen. Die erhaltenen Erkenntnisse sind für die praktische Anwendung der Theorie der Wissensstrukturen unmittelbar nutzbar.

Literatur

Albert, D. & Lukas, J. (Hrsg.) (1999). *Knowledge Spaces: Theories, Empirical Research, Applications*. Mahwah: Lawrence Erlbaum Associates.

Batchelder, W. H. & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, **6**, 57-86.

Birkhoff, G. (1937). Rings of sets. *Duke Mathematical Journal*, **3**, 443-454.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler*. Berlin: Springer.

Brent, R. P. (1973). *Algorithms for Function Minimization without Derivatives*. Englewood Cliffs, N. J.: Prentice-Hall.

Cosyn, E. & Thiéry, N. (2000). A Practical Procedure to Build a Knowledge Structure. *Journal of Mathematical Psychology*, **44**, 383-407.

Csörgö, S. & Faraway, J. (1996). The exact and asymptotic distributions of Cramer-von Mises statistics. *Journal of the Royal Statistical Society*, Series B, **58**, 221-234.

Doignon, J.-P. & Falmagne, J.-C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, **23**, 175-196.

Doignon, J.-P. & Falmagne, J.-C. (1999). *Knowledge Spaces*. Berlin: Springer.

Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J.-P. & Johannesen, L. (1990). Introduction to knowledge spaces: how to build, test and search them. *Psychological Review*, **97**, 201-224.

Fishman, G. S. (1973). *Concepts and methods in discrete event digital simulation*. New York: John Wiley & Sons.

Fries, S. (1997). Empirical validation of a markovian learning model for knowledge structures. *Journal of Mathematical Psychology*, **41**, 65-70.

Gegenfurtner, K. (1992). Praxis: Brent's algorithm for function minimization. *Behavior Research Methods Instruments and Computers*, **24**, 560-564.

Greene, W. H. (1990). *Econometric analyis*. New York: Macmillan Publishing Company.

Kambouri, M., Koppen, M., Villano, M. & Falmagne, J.-C. (1994). Knowledge assessment: tapping human expertise by the QUERY routine. *International Journal of Human-Computer Studies*, **40**, 119-151.

Leeuwe, J. F. J. v. (1974). Item tree analysis. *Nederlands Tijdschrift voor de Psychologie*, **29**, 475-484.

Lindgren, B. W. (1976). *Statistical theory*. New York: Macmillan Publishing Company.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1988). *Numerical Recipes in C. The Art of Scientific Computing*. New York: Camebridge University Press.

Schrepp, M. (1999). On the empirical construction of implications between bi-valued test items. *Mathematical Social Sciences*, **38**, 361-375.

Wesiak, G. & Albert, D. (2001). *Ordering Inductive Reasoning Tests based on Knowledge Space Theory*. Poster presented at the 5th Meeting of the German Cognitive Science Society, Leipzig, Germany, September 2001.

Wickens, T. D. (1982). *Models for behavior: Stochastic processes in psychology*. San Francisco: Freeman and Company.

A Verwendete Software

A.1 Simulationsroutine

```
/*
 *  simulate.c: simulates responses of subjects in a
 *              specified knowledge space assuming a
 *              Markovian model (improved by ran3)
 *  input: number_of_items
 *         structure.dat (knowledge structure)
 *         parameters.ini (model parameters)
 *  output: response.out (response patterns)
 *  author: Florian Wickelmaier (wickelmaier@web.de)
 *  last modified: 08/OCT/2001
 */

#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <time.h>

#define NITEMS  10        /* number of items */
#define NSTATES 128       /* number of knowledge states */
#define NPAR (3*NITEMS+2) /* number of parameters */

#define MBIG 1000000000   /* needed by the random    */
#define MSEED 161803398   /* numbers generator ran3  */
#define MZ 0              /* cf. numerical recipes   */
#define FAC (1.0/MBIG)    /* p.212                   */

int nitems,nstates,state[NSTATES],learnstep,nsubjects;
double g[NITEMS],beta[NITEMS],eta[NITEMS],init_par[NPAR],
  matrix[NSTATES][NSTATES],prob_K[NSTATES],cprob_K[NSTATES];

/*************** input routines ***************/

int ReadItemNumber(char *number){
   nitems = atoi(number);
   if(nitems > NITEMS){
     fprintf(stderr, "maximal number of items is %i\n", NITEMS);
     return 1;
   }
   return 0;
}

int ReadKnowledgeStructure(char *structfile){
  int i,code;
  char pattern[NITEMS];
  FILE *fp,*fopen();
```

```
  if((fp = fopen(structfile,"r")) == NULL){
    fprintf(stderr, "i/o error: cannot open file %s!\n", structfile);
    return 1;
  }
  nstates = 0;
  while(fscanf( fp, " %s", pattern) != EOF){
    if(nitems != (int)strlen(pattern)){
      fprintf(stderr, "number of items doesn‘t match file %s!\n", structfile);
      return 1;
    }
    code = 0;
    for(i=0; i<nitems; i++){
      if (pattern[i] == '1'){
        code += 1 << (nitems-i-1);
      }
    }
    state[nstates++] = code;
  }
  fclose(fp);
  return 0;
}

int ReadParameterValues(char *inifile){
  int i,npar;
  FILE *fp,*fopen();

  if((fp = fopen(inifile,"r")) == NULL){
    fprintf(stderr, "i/o error: cannot open file %s!\n", inifile);
    return 1;
  }
  npar = 0;
  while(fscanf(fp, " %lf", &init_par[npar]) != EOF){
    npar++;
  }
  close(fp);
  if(npar != (3*nitems+2)){
    printf("error: %d parameter values requested in input file %s\n",
           3*nitems+2, inifile);
    return 1;
  }
  for (i=0; i<nitems; i++){
    g[i] = init_par[i];
    beta[i] = init_par[i + nitems];
    eta[i] = init_par[i + 2 * nitems];
  }
  learnstep = (int)init_par[npar - 2];
  nsubjects = (int)init_par[npar - 1];
  return 0;
}

/*************** general routines **************/
```

```
int Adjacent(int state1, int state2){
  int i;

  if((state1 & state2) == state1){
    for(i=0; i<nitems; i++){
      if((state1 | (1 << i)) == state2){
        return nitems-i-1;
      }
    }
  }
  return -1;
}

void SetupMatrix(){
  int i, j, item;

  for(i=0; i<nstates; i++){
    matrix[i][i] = 1.0;
    for(j=0; j<nstates; j++){
      if(i != j){
        matrix[i][j] = 0.0;
        if((item = Adjacent(state[i], state[j])) != -1){
          matrix[i][j] = g[item];
          matrix[i][i] -= g[item];
        }
      }
    }
  }
}

void Transition(){
  int i,j;
  double new_prob[NSTATES];

  for(i=0; i<nstates; i++){
    new_prob [i] = 0.0;
    for(j=0; j<nstates; j++){
      new_prob[i] += prob_K[j] * matrix[j][i];
    }
  }
  for(i=0; i<nstates; i++){
    prob_K[i] = new_prob[i];
  }
}

void ComputeStateProbabilities(){
  int i,j;

  prob_K[0] = 1.0;
  for(i=1; i<nstates; i++){
    prob_K[i] = 0.0;
  }
```

```
  for(j=0; j<learnstep; j++){
    Transition();
  }
}

void ComputeCumulativeProbabilities(void){
  int i;
  double c = .0;

  for(i=0; i<nstates; i++){
    c += prob_K[i];
    cprob_K[i] = c;
  }
}

double ran3(int *idum){ /* generates random number r in (0;1) */
  static int inext,inextp;
  static long ma[56];
  static int iff=0;
  long mj,mk;
  int i,ii,k;

  if(*idum < 0 || iff == 0){
    iff = 1;
    mj = MSEED-(*idum < 0 ? -*idum : *idum);
    mj %= MBIG;
    ma[55] = mj;
    mk = 1;
    for(i=1;i<=54;i++){
      ii = (21*i) % 55;
      ma[ii] = mk;
      mk = mj - mk;
      if(mk < MZ) mk += MBIG;
      mj = ma[ii];
    }
    for(k=1;k<=4;k++)
      for(i=1;i<=55;i++){
        ma[i] -= ma[1+(i+30) % 55];
        if(ma[i] < MZ) ma[i] += MBIG;
      }
    inext = 0;
    inextp = 31;
    *idum = 1;
  }
  if(++inext == 56) inext = 1;
  if(++inextp == 56) inextp = 1;
  mj = ma[inext] - ma[inextp];
  if(mj < MZ) mj += MBIG;
  ma[inext] = mj;
  return mj*FAC;
}
```

```
int SimulateResponsePatterns(char *outfile){
  int i,subject,sstate,idum;
  double r;
  FILE *fp,*fopen();

  if((fp = fopen(outfile,"w")) == NULL){
    fprintf(stderr, "i/o error: cannot open file %s!\n", outfile);
    return 1;
  }
  idum = -1 * (time(NULL) % 10000); /* seed by time */
  for(subject=0; subject<nsubjects; subject++){
    r = ran3(&idum);
    if(r <= cprob_K[0]){
      sstate = state[0];
    }else{
      for(i=1; i<nstates; i++){
        if(cprob_K[i-1] < r && r <= cprob_K[i]){
          sstate = state[i];
        }
      }
    }
    for(i=0; i<nitems; i++){
      r = ran3(&idum);
      if(sstate & (1 << (nitems-i-1))){
        (r <= beta[i]) ? fprintf(fp, "0") : fprintf(fp, "1");
      }else{
        (r <= eta[i]) ? fprintf(fp, "1") : fprintf(fp, "0");
      }
    }
    fprintf(fp, "\n");
  }
  fclose(fp);
  printf("\noutput is written in file %s\n\n", outfile);
  return 0;
}

/*************** output routines ***************/

void PrintErrorMessage(void){
  fprintf(stderr, "usage: simulate number_of_items structure_file"
                  " parameter_file output_file\n");
}

void PrintParameters(void){
  int i;

  for(i=0; i<nitems; i++){
    printf(" item %2d:  g = %.2lf, beta = %.2lf, eta = %.2lf\n",
           i+1,g[i],beta[i],eta[i]);
  }
}
```

```
void PrintMatrix(void){
  int i,j;

  for(i=0; i<nstates; i++){
    for(j=0; j<nstates; j++){
      printf(" %4.2lf", matrix[i][j]);
    }
    printf("\n");
  }
  printf("\n");
}

void CheckProbabilities(void){
  int i;

  for(i=0; i<nstates; i++){
    printf(" prob(%3i) = %1.6lf  cprob = %1.6lf\n",
           state[i],prob_K[i],cprob_K[i]);
  }
}

void PrintStartMessage(void){
  printf("*********************************\n");
  printf("**                             **\n");
  printf("**        Markov Models        **\n");
  printf("**     Computer Simulation     **\n");
  printf("**                             **\n");
  printf("*********************************\n\n");
  printf(" number of items:      %6i\n", nitems);
  printf(" number of states:     %6i\n", nstates);
  printf(" number of learnsteps: %6i\n", learnstep);
  printf(" number of subjects:   %6i\n", nsubjects);
  printf("\nmodel parameters:\n");
  PrintParameters();
  printf("\ntransition matrix:\n");
  PrintMatrix();
  printf("state probabilities:\n");
  CheckProbabilities();
}

/*************** main routine ***************/

int main(int argc, char *argv[]){
  if(argc != 5){
    PrintErrorMessage();
    exit(1);
  }
  if(ReadItemNumber(*++argv)){
    exit(1);
  }
  if(ReadKnowledgeStructure(*++argv)){
    exit(1);
```

```
  }
  if(ReadParameterValues(*++argv)){
    exit(1);
  }
  SetupMatrix();
  ComputeStateProbabilities();
  ComputeCumulativeProbabilities();
  PrintStartMessage();
  if(SimulateResponsePatterns(*++argv)){
    exit(1);
  }
  return 0;
}
```

A.2 Routine zur Berechnung des Diskrepanzindexes

```
/*
 *  distance.c: computes the distance between response
 *              patterns and the knowledge space, the
 *              distribution over the knowledege space,
 *              and beta and eta
 *  input: number_of_items
 *         responses.res (grouped responses)
 *         structure.dat (knowledge structure)
 *  author: Florian Wickelmaier (wickelmaier@web.de)
 *  last modified: 04/AUG/2001
 */

#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <math.h>

#define NITEMS  10
#define NPATTERNS (1<<NITEMS)
#define NSTATES 128

int nitems,nstates,state[NSTATES],freq[NPATTERNS],nsubjects,
  f[NITEMS];
double ev,sd,p[NSTATES],beta[NITEMS],eta[NITEMS];

/*************** input routines ***************/

int ReadItemNumber(char *number){
  nitems = atoi(number);
  if(nitems > NITEMS){
    fprintf(stderr, "maximal number of items is %i\n", NITEMS);
    return 1;
  }
  return 0;
}

int ReadData(char *datafile){
  int i,number,code;
  char pattern[NITEMS];
  FILE *fp,*fopen();

  nsubjects = 0;
  for(i=0;i<NPATTERNS;i++){
    freq[i] = 0;
  }
  if((fp = fopen(datafile,"r")) == NULL){
    fprintf(stderr, "i/o error: cannot open file %s!\n", datafile);
    return 1;
  }
  while(fscanf( fp, " %d %s", &number, pattern) != EOF){
```

```
    if(nitems != (int)strlen(pattern)){
      fprintf(stderr, "error in file %s!\n", datafile);
      return 1;
    }
    code = 0;
    for(i=0;i<nitems;i++){
      if(pattern[i] == '1'){
        code += 1 << (nitems-i-1);
      }
    }
    freq[code] = number;
    nsubjects += number;
  }
  fclose(fp);
  return 0;
}

int ReadKnowledgeStructure(char *structfile){
  int i,code;
  char pattern[NITEMS];
  FILE *fp,*fopen();

  if((fp = fopen(structfile,"r")) == NULL){
    fprintf(stderr, "i/o error: cannot open file %s!\n", structfile);
    return 1;
  }
  nstates = 0;
  while(fscanf( fp, " %s", pattern) != EOF){
    if(nitems != (int)strlen(pattern)){
      fprintf(stderr, "number of items doesn't match file %s!\n", structfile);
      return 1;
    }
    code = 0;
    for(i=0;i<nitems;i++){
      if (pattern[i] == '1'){
        code += 1 << (nitems-i-1);
      }
    }
    state[nstates++] = code;
  }
  fclose(fp);
  return 0;
}

/*************** general routines **************/

int ComputeDistance(int pattern, int state){
  int i,symdif,dist=0;

  symdif = pattern^state;
  for(i=0;i<nitems;i++){
    if(symdif&(1<<i)) dist++;
```

```
  }
  return dist;
}

int ComputeMinimum(int pattern){
  int j,minimum,dist;

  minimum = nitems;
  for(j=0;j<nstates;j++){
    dist = ComputeDistance(pattern,state[j]);
    (minimum > dist) ? (minimum = dist) : (minimum = minimum);
  }
  return minimum;
}

int ComputeNumberOfStates(int pattern){
  int j,dist,minimum,nstates_per_pattern=0;

  for(j=0;j<nstates;j++){
    dist = ComputeDistance(pattern,state[j]);
    minimum = ComputeMinimum(pattern);
    if(dist == minimum) nstates_per_pattern++;
  }
  return nstates_per_pattern;
}

void ComputeProbabilities(){
  int i,j,dist,minimum;

  for(j=0;j<nstates;j++){
    p[j] = 0.0;
  }
  for(j=0;j<nstates;j++){
    for(i=0;i<NPATTERNS;i++){
      if(freq[i]){
        dist = ComputeDistance(i,state[j]);
        minimum = ComputeMinimum(i);
        if(dist == minimum){
          p[j] += (double) freq[i]/ComputeNumberOfStates(i);
        }
      }
    }
  }
}

void ComputeDiscrepancyFunction(){
  int i,minimum;

  for(i=0;i<nitems;i++){
    f[i] = 0;
  }
  for(i=0;i<NPATTERNS;i++){
```

```
    if(freq[i]){
      minimum = ComputeMinimum(i);
      f[minimum] += freq[i];
/*      printf(" %5i \t\t %i\n", i,minimum); */
    }
  }
}

void ComputeMoments(){
  int i;
  double ev_sq,var;

  ev = 0.0;
  sd = 0.0;
  ev_sq = 0.0;
  for(i=0;i<nitems;i++){
    ev += (double) i * f[i]/nsubjects;
  }
  for(i=0;i<nitems;i++){
    ev_sq += (double) i*i * f[i]/nsubjects;
  }
  var = ev_sq - ev*ev;
  sd = sqrt(var);
}

void ComputeBeta(){
  int item,i,j,dist,minimum;
  double denom=0.0,num=0.0;

  for(item=0;item<nitems;item++){
    for(i=0;i<NPATTERNS;i++){
      if(freq[i]){
        for(j=0;j<nstates;j++){
          dist = ComputeDistance(i,state[j]);
          minimum = ComputeMinimum(i);
          if(dist == minimum){
            if(state[j]&(1<<(nitems-item-1))){
              denom += (double) freq[i]/ComputeNumberOfStates(i);
              if(!(i&(1<<(nitems-item-1)))){
                num += (double) freq[i]/ComputeNumberOfStates(i);
              }
            }
          }
        }
      }
    }
    beta[item] = num/denom;
    denom = 0.0;
    num = 0.0;
  }
}
```

```
void ComputeEta(){
  int item,i,j,dist,minimum;
  double denom=0.0,num=0.0;

  for(item=0;item<nitems;item++){
    for(i=0;i<NPATTERNS;i++){
      if(freq[i]){
        for(j=0;j<nstates;j++){
          dist = ComputeDistance(i,state[j]);
          minimum = ComputeMinimum(i);
          if(dist == minimum){
            if(!(state[j]&(1<<(nitems-item-1)))){
              denom += (double) freq[i]/ComputeNumberOfStates(i);
              if(i&(1<<(nitems-item-1))){
                num += (double) freq[i]/ComputeNumberOfStates(i);
              }
            }
          }
        }
      }
    }
    eta[item] = num/denom;
    denom = 0.0;
    num = 0.0;
  }
}

/**************** output routines ****************/

void PrintErrorMessage(){
  fprintf(stderr, "usage: distance number_of_items data_file"
                  " structure_file\n");
}

void PrintStartMessage(){
  printf("**********************************\n");
  printf("**                              **\n");
  printf("**      Knowledge Spaces        **\n");
  printf("**  Discrepancy Distribution    **\n");
  printf("**                              **\n");
  printf("**********************************\n\n");
  printf(" number of items:      %6i\n", nitems);
  printf(" number of states:     %6i\n", nstates);
  printf(" number of subjects:   %6i\n", nsubjects);
  printf("\n pattern:   minimum distance:\n");
}

void PrintDistribution(){
  int i;

  printf("\n distribution of the minimum\n");
  for(i=0;i<nitems;i++){
```

```
    if(f[i]) printf(" f(%i) = %i \t %lf\n", i,f[i],(double) f[i]/nsubjects);
  }
}

void PrintMoments(){
  printf("\n expected minimum distance (standard deviation)\n");
  printf(" %lf (%lf)\n", ev,sd);
}

void PrintProbabilities(){
  int j;

  printf("\n probabilities of the knowledge states\n");
  for(j=0;j<nstates;j++){
    printf(" p(%3i) = %lf\n", state[j],p[j]/nsubjects);
  }
}

void PrintParameters(){
  int i;

/*  printf("\n parameters: \t beta \t\t eta\n");
  for(i=0;i<nitems;i++){
    printf(" item(%i): \t %lf \t %lf\n", i,beta[i],eta[i]);
  } */
  printf("%8.6lf ", ev);
  for(i=0;i<nitems;i++) printf("%8.6lf ", beta[i]);
  for(i=0;i<nitems;i++) printf("%8.6lf ", eta[i]);
  printf("\n");
}

/**************** main routine ****************/

int main(int argc, char *argv[]){
  if(argc != 4){
    PrintErrorMessage();
    exit(1);
  }
  if(ReadItemNumber(*++argv)){
    exit(1);
  }
  if(ReadData(*++argv)){
    exit(1);
  }
  if(ReadKnowledgeStructure(*++argv)){
    exit(1);
  }
/*  PrintStartMessage(); */
  ComputeDiscrepancyFunction();
  ComputeProbabilities();
  ComputeMoments();
/*  PrintDistribution();
```

```
  PrintMoments();
  PrintProbabilities(); */
  ComputeBeta();
  ComputeEta();
  PrintParameters();
  return 0;
}
```

Ich versichere hiermit, daß ich die anliegende Arbeit mit dem Thema:

Empirische Untersuchung zur Erfassung von Wissen durch deterministische und probabilistische Wissensstrukturen

selbständig verfaßt und keine anderen Hilfsmittel als die angegebenen benutzt habe. Die Stellen, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen sind, habe ich in jedem einzelnen Fall durch Angabe der Quelle, auch der benutzten Sekundärliteratur, als Entlehnung kenntlich gemacht.