Diskussionsbeiträge der Projektgruppe Friedensforschung Konstanz, Nr. 65, 2010
www.regener-online.de
ISSN 1611-1818

Wilhelm Kempf

Quantifizierung qualitativer Daten[1]

Abstract: Obwohl der Mainstream der quantitativen Sozialforschung gern einen Überlegenheitsanspruch gegenüber der qualitativen Sozialforschung erhebt, bleiben die Voraussetzungen der Quantifizierung sozialwissenschaftlicher Daten jedoch meist ungeprüft. In der Praxis begnügt man sich nur allzu oft mit *ad hoc*-Quantifizierungen, die man zu Scores verrechnet, und dann so tut, *als ob* man damit bereits metrische Daten vorliegen hätte.

Seit den grundlegenden Arbeiten von Louis Guttman, Paul Lazarsfeld, Patrick Suppes, Georg Rasch und anderen Autoren um die Mitte des vorigen Jahrhunderts ist bekannt, dass auf diese Weise weder metrische Skalen konstruiert noch die in den Daten enthaltenen statistischen Informationen ausgeschöpft werden können, sofern die (zunächst qualitativen) Daten nicht bestimmten empirischen Gesetzmäßigkeiten genügen.

Der vorliegende Arbeitsbericht skizziert die Hauptergebnisse dieses Forschungszweiges, erläutert die mit der Quantifizierung sozialwissenschaftlicher Daten einhergehenden wissenschaftstheoretischen und mathematisch-statistischen Problemlagen anhand von Beispielen aus der psychologischen Leistungsmessung, der Fragebogenkonstruktion und der Inhaltsanalyse und empfiehlt die Anwendung eines Loss-of-Information-Indexes zur Quantifizierung des Informationsverlustes, welcher mit der Verwendung von Scores einhergeht, wenn die empirischen Voraussetzungen der Scorebildung verletzt sind.

1. Der immanent qualitative Charakter psychologischer Daten

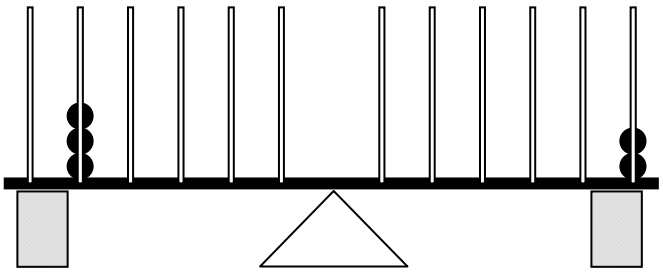
Abgesehen von physikalischen Messungen wie z.B. Reaktionszeitmessungen oder EEG-Aufzeichnungen hat es die Psychologie fast ausschließlich mit qualitativen Daten zu tun. In der klinischen Psychologie beobachtet man das Auftreten bestimmter Krankheitssymptome, in der Leistungsdiagnostik untersucht man, ob die Versuchspersonen (Vpn) bestimmte Testaufgaben lösen können, und mit der Fragebogenmethode wird erfasst, für welche aus einer Reihe von vorgegebenen Antwortmöglichkeiten sich die Vpn entscheiden. Bei offenen Interviews, die textinterpretativ oder inhaltsanalytisch ausgewertet werden, springt der qualitative Charakter der Daten noch deutlicher ins Auge. Auch hier gewinnt man die Daten aber letztlich dadurch, dass man bestimmte Fragen an den Text richtet und die Antworten, welche man aus dem Text herausliest, anschließend kategorisiert.

In all diesen Fällen liegen also zunächst qualitative Beobachtungen vor, deren statistische Weiterverarbeitung es erforderlich macht, dass sie in numerische Variablen übersetzt werden. Ein erster Schritt in diese Richtung besteht darin, die Antwortkategorien mit Zahlen zu benennen.

So ordnet man z.B. dem Auftreten eines Symptoms die Zahl Eins und seinem Ausbleiben die Zahl Null zu. Ob ein bestimmtes Symptom (i) bei einer bestimmten Vp (v) auftritt, wird dann durch den Zahlenwert

$$x_{vi} = \begin{cases} 1 \text{ wenn das Symptom auftritt} \\ 0 \text{ wenn nicht} \end{cases}$$

beschrieben. In ähnlicher Weise ordnet man der Antwort einer Vp auf eine Testaufgabe (i) den Zahlenwert $x_{vi} = 1$ zu, wenn die Antwort richtig war, und $x_{vi} = 0$, wenn dies nicht der Fall ist.



"Welche Seite neigt sich nach unten, wenn man die Holzklötzchen entfernt?"

Abbildung 1, Beispiel für eine Balkenwaagenaufgabe (nach Kempf, 2009, S. 124).

[1] Gefördert von der Deutschen Forschungsgemeinschaft (DFG), Kenziffer KE 300/8-1.

Mitunter klassifiziert man die Antworten nicht nur als richtig oder falsch, sondern protokolliert ganz genau, *welche* Antwort die Vp gegeben hat. Auch dabei kann man die verschiedenen Antwortmöglichkeiten wieder mit Zahlen benennen. So kann man z.B. bei Balkenwaagenaufgaben, wie sie bereits Inhelder & Piaget (1958) verwendet haben, um die Stufenentwicklung der Intelligenz zu untersuchen (vgl. Abb. 1), unterscheiden, ob die Vp meint, dass die Waage in Balance bleibt ($x_{vi} = 0$), dass sich die Seite mit dem größeren Gewicht nach unten neigt ($x_{vi} = 1$) oder die Seite mit der größeren Distanz ($x_{vi} = 2$).

In standardisierten Fragebögen legt man den Vpn eine Reihe von Aussagen vor und lässt sie beurteilen, ob oder wie weit sie der jeweiligen Aussage (i) zustimmen. Z.B. gibt man die folgenden Antwortkategorien vor, die man wieder mit Zahlen benennt:

$$x_{vi} = \begin{cases} 0 \text{ volle Zustimmung} \\ 1 \text{ teilweise Zustimmung} \\ 2 \text{ weder} - \text{noch} \\ 3 \text{ teilweise Ablehnung} \\ 4 \text{ volle Ablehnung} \end{cases}$$

In der Inhaltsanalyse kategorisiert man einen Text oder eine Textstelle (v) danach, ob darin ein bestimmtes Textmerkmal (i) vorkommt ($x_{vi} = 1$) oder nicht ($x_{vi} = 0$), oder man unterscheidet mehrere Ausprägungen des Textmerkmals, die man ebenfalls mit Zahlen benennt. Wenn man z.B. Texte analysiert, in denen über Konflikte berichtet wird, so kann man diese u.a. danach klassifizieren, welche Eigenschaften einer der Konfliktparteien zugeschrieben werden:

$$x_{vi} = \begin{cases} 0 \text{ keine Eiigenschaften} \\ 1 \text{ positive Eigenschaften} \\ 2 \text{ negative Eigenschaften} \\ 3 \text{ sowohl als auch} \end{cases}$$

In all diesen Fällen hat man es zunächst mit *qualitativ* verschiedenen Kategorien zu tun, die durch die zugeordneten Maßzahlen auf einer *Nominalskala* (lat. nomen = Name) gemessen werden, welche die Messobjekte (= die Antworten) lediglich mit Zahlen *benennt* und also nur dem Namen nach eine Skala ist.

Welche Zahlen man dafür verwendet, ist völlig beliebig. Einzige Bedingung ist die Eindeutigkeit der Zuordnung, so dass

- je zwei Messobjekten v und w, die bezüglich einer Variable (i) in die gleiche Kategorie fallen ($v =^o w$), die selbe Maßzahl ($x_{vi} = x_{wi}$) zugeordnet wird und
- je zwei Messobjekten, die bezüglich der Variable in verschiedene Kategorien fallen ($v \neq^o w$), verschiedene Maßzahlen ($x_{vi} \neq x_{wi}$) zugeordnet werden.

Entsprechend kann man aus der Gleichheit oder Ungleichheit der Maßzahlen jederzeit auf die Gleichheit oder Ungleichheit der empirischen Kategorien zurückschließen und jede *umkehrbar eindeutige* Transformation der Skalenwerte

(1) $X \rightarrow Y : x_{vi} \rightarrow y_{vi}$

liefert eine gleichwertige Skala. Eine Quantifizierung wird dadurch allerdings noch nicht geleistet. Auch wenn man sie auf einer Nominalskala abbildet, sind und bleiben die Antwortkategorien die Ausprägungen einer *qualitativen* Variablen.

2. Quantität und Qualität

Von einer *Quantifizierung* kann man erst dann sprechen, wenn die Maßzahlen (zumindest) eine empirische Ordnung der Messobjekte abbilden. Dies ist keineswegs trivial. Denn um die Messobjekte in eine empirische Ordnung bringen zu können, benötigt man – über die empirische Gleichheitsrelation ($v =^o w$) hinaus – eine empirische Ordnungsrelation ($v >^o w$). Damit eine zweistellige (empirische) Relation eine Ordnungsrelation darstellt, muss sie aber dieselben Eigenschaften aufweisen, wie numerische Ordnungsrelationen. Diese Eigenschaften sind:

- *Irreflexivität:* Kein Messobjekt ist größer als es selbst. Formal: $\neg(v >^o v)$.[2]
- *Asymmetrie:* Wenn v größer ist als w, dann ist w nicht größer als v. Formal: $v >^o w \rightarrow \neg(w >^o v)$.[3]

[2] Das Zeichen „$\neg$“ steht für die Negation („nicht“).

- *Transitivität:* Wenn v größer ist als w und w größer als z, dann ist v größer als z. Formal: $(v >^o w \wedge w >^o z) \rightarrow v >^o z$.[4]

Erst wenn diese Voraussetzungen empirisch erfüllt sind, kann man den Messobjekten Maßzahlen zuordnen, welche die empirische Rangordnung der Messobjekte widerspiegeln, so dass man von der numerischen Rangordnung der Messobjekte jederzeit auf ihre empirische Rangordnung zurückschließen kann, und jede *streng monoton wachsende* Transformation der Skalenwerte liefert eine gleichwertige Rang- bzw. *Ordinalskala*.

Die Ordinalskala spiegelt zwar eine Rangordnung der Messobjekte wider, der Abstand zwischen je zwei Messwerten ist jedoch nicht aussagekräftig. In der Praxis gibt man sich mit bloßen Ordinalskalen daher nicht gern zufrieden, zumal viele statistische Verfahren nur auf metrische Skalen angewendet werden dürfen. Deren Konstruktion erfordert über die o.g. zweistelligen Relationen ($v =^o w$ und $v >^o w$) hinaus jedoch auch mehrstellige empirische Relationen, so dass der empirische Unterschied zwischen je zwei Messobjekten auf die Differenz zwischen den ihnen zugeordneten Maßzahlen abgebildet werden kann und man von gleichen Maßzahldifferenzen auf gleiche empirische Unterschiede und von größeren Maßzahldifferenzen auf größere empirische Unterschiede zurückschließen kann.

Innerhalb der metrischen Skalen unterscheidet man – je nach Schärfe der Skala – zwischen Intervallskalen, Rationalskalen, Differenzenskalen und Absolutskalen. Welche (mehrstelligen) empirischen Relationen erforderlich sind, um einen bestimmten Skalentyp zu gewährleisten, wird von der sog. Allgemeinen Messtheorie (vgl. Suppes & Zinnes, 1963; Pfanzagl, 1968; Orth, 1974) untersucht.

Die *Intervallskala* ist die schwächste der metrischen Skalen. Sie bildet zwar die empirischen Unterschiede auf die Maßzahldifferenzen ab, verfügt jedoch weder über einen festen Nullpunkt noch über eine feste Maßeinheit und jede *positive lineare Maßzahltransformation*

$$(2) \qquad X \rightarrow Y : x_{vi} \rightarrow y_{vi} = a + bx_{vi} \qquad \text{mit } b > 0$$

liefert eine gleichwertige Skala. Ein Beispiel ist die Temperaturmessung auf der Celsius- oder Fahrenheitskala.

Die *Rationalskala* verfügt über einen festen Nullpunkt (nicht jedoch über eine feste Maßeinheit) und jede *proportionale Maßzahltransformation* (= positive lineare Maßzahltransformation mit a = 0)

$$(3) \qquad X \rightarrow Y : x_{vi} \rightarrow y_{vi} = bx_{vi} \qquad \text{mit } b > 0$$

liefert eine gleichwertige Skala. Als Beispiel kann die Längenmessung in Zentimeter oder Zoll dienen.

Die *Differenzenskala* verfügt im Unterschied dazu über eine feste Maßeinheit (nicht jedoch über einen festen Nullpunkt) und jede *Translation* der Maßzahlen (= positive lineare Maßzahltransformation mit b = 1)

$$(4) \qquad X \rightarrow Y : x_{vi} \rightarrow y_{vi} = a + x_{vi}$$

liefert eine gleichwertige Skala. Als Beispiel können logarithmische Längenskalen dienen, wie sie in der Feinmechanik Verwendung finden.

Die *Absolutskala* ist die schärfste aller Skalen. Sie verfügt sowohl über einen festen Nullpunkt als auch über eine feste Maßeinheit. Hier ist nur noch die *identische Transformation* (= positive lineare Maßzahltransformation mit a = 0 und b = 1)

$$(5) \qquad X \rightarrow Y : x_{vi} \rightarrow y_{vi} = x_{vi}$$

zulässig, welche die Maßzahlen in sich selbst überführt. Als Beispiel dafür kann man die Messung der Wahrscheinlichkeit eines Ereignisses anführen.

Zwischen den verschiedenen Skalentypen besteht die in Abb. 2 dargestellte Hierarchie. Je höher eine Skala in der Hierarchie steht, desto aussagekräftiger ist sie und desto strengeren Restriktionen unterliegen die zulässigen Maßzahltransformationen. Wendet man auf eine Skala eine Transformation an, die nur für einen tiefer stehenden Skalentyp zulässig ist, so wird die Skala dadurch nicht gänzlich zerstört. Das Skalenniveau sinkt jedoch auf jenes der tiefer stehenden Skala ab. Man erhält also eine schlechtere Skala, die weniger aussagekräftig ist. Auch wenn man auf eine Rational- oder Differenzenskala eine Transformation anwendet, die nur für den jeweils anderen Skalentyp zulässig ist, wird die Skala dadurch nicht gänzlich zerstört, aber das Skalenniveau sinkt auf jenes der darunter stehenden Intervallskala.

---

[3] Das Zeichen „$\rightarrow$“ steht für die Subjunktion („wenn – dann“).
[4] Das Zeichen „$\wedge$“ steht für die Konjunktion („und“).

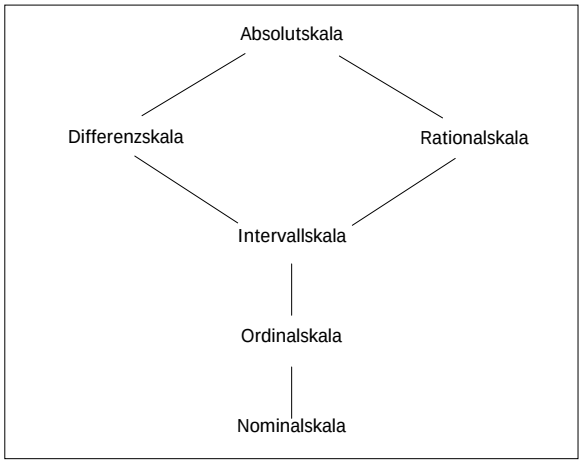


Abbildung 2, Hierarchie der Skalentypen (nach Kempf, 2009, S. 106)

3. Manifeste und latente Variablen

Die unmittelbar beobachtbaren (= manifesten) Variablen, mit denen man es in der Psychologie zu tun hat, beschreiben zunächst einmal nur die Antworten, welche eine Vp (oder ein Text) auf die an sie gerichteten Fragen gibt. Das eigentliche Interesse der Psychologie gilt aber nicht diesen manifesten Variablen, sondern zugrunde liegenden (= latenten) Variablen, für welche sie als Indikatoren dienen. So schließt man z.B. von Krankheitssymptomen auf eine bestimmte Erkrankung der Vp, von der Testleistung einer Vp auf ihre Leistungsfähigkeit oder von Fragebogenantworten auf eine Einstellung. Auch in der Inhaltsanalyse geht es zumeist nicht nur um die Beschreibung manifester Textinhalte, sondern um deren latenten Gehalt, d.h. um die Regeln, welchen die durch den Text repräsentierten (subjektiven) Wirklichkeitskonstruktionen folgen (vgl. Kempf, 2008, S.39ff).

Die Skaleneigenschaft der manifesten Variablen, die man beobachtet, und jene der latenten Variablen, auf welche man daraus schließt, ist jedoch nicht dasselbe. Selbst dort, wo man die manifesten Variablen mittels physikalischer Messung auf einer metrischen Skala beschreiben kann, bleibt das Skalenniveau der latenten Variable in der Regel dahinter zurück. So ist z.B. „Hunger“ eine subjektseitig definierte (latente) Variable, für die sich verschiedene objektseitig definierte (manifeste) Indikatoren angeben lassen, die sich – wie z.B. die Dauer des Nahrungsentzuges – z.T. sogar auf einer Rationalskala messen lassen. Zwischen diesem Indikator und dem Hunger besteht ein streng monoton wachsender Zusammenhang: Je länger der Nahrungsentzug ist, desto größer ist der Hunger. So lange über die genaue Form dieses Zusammenhanges jedoch nichts bekannt ist, kann man die Dauer des Nahrungsentzuges aber noch so genau messen, über den Hunger wird man daraus nur ordinale Informationen beziehen.

Um von manifesten Variablen auf latente Variablen schließen zu können, muss eine gesetzmäßige Beziehung angegeben werden, wie die zu erschließende, latente Variable mit den beobachteten, manifesten Indikatoren zusammenhängt. Z.B. postuliert man in der Leistungsdiagnostik:

- Je leistungsfähiger eine Vp ist, desto eher wird sie Aufgaben (Items) einer bestimmten Art lösen können.
- Und umgekehrt schließt man dann: Je eher eine Vp Aufgaben dieser Art lösen kann, desto leistungsfähiger ist sie.

Die Leistungsfähigkeit wird dabei als eine quantitative Variable (Θ) gedacht, deren Ausprägung bei einer gegebenen Vp mit θ_v bezeichnet wird. Dass man die Vpn aufgrund ihrer qualitativen, manifesten Antworten (richtig vs. falsch) auf einer latenten Dimension anordnen kann, ist keineswegs trivial, sondern setzt voraus, dass alle Items dieselbe latente Dimension messen. Mit anderen Worten: Sie setzt voraus, dass (1) jedes Item eine empirische Ordnungsrelation zwischen je zwei Vpn definiert und (2) dass alle Items dieselbe Ordnungsrelation definieren.

Eine Antwort auf die Frage, wie dies möglich ist, wurde erstmals von Guttman (1950) gegeben, der die empirische Ordnungsrelation $v >^{\circ} w$ zwischen je zwei Vpn v und w durch die Regel

(6) $$\left(\left[x_{vi} = 1\right] \wedge \left[x_{wi} = 0\right]\right) \Rightarrow v >^{\circ} w$$

definiert, worin

(7) $$x_{vi} = \begin{cases} 1 \text{ wenn die Antwort richtig ist} \\ 0 \text{ wenn nicht} \end{cases}.$$

Immer dann, wenn eine Vp v ein Item i richtig beantwortet, auf welches eine andere Vp w eine falsche Antwort gibt, schließen wir also darauf, dass Vp v leistungsfähiger ist als w. Wird ein Item von beiden oder von keiner der Vpn richtig beantwortet, so sagt dies dagegen nichts darüber aus, wie sie sich in der gemessenen Leistungsdimension voneinander unterscheiden.

Somit definiert jedes Item eine Ordnungsrelation zwischen den Vpn, und die Vpn können aufgrund ihrer Antworten genau dann auf einer Leistungssdimension angeordnet werden, wenn diese Relationen einander nicht widersprechen. D.h. es darf nicht vorkommen, dass ein Item die Relation $v >^{o} w$, ein anderes jedoch die Relation $w >^{o} v$ definiert. Diese Forderung, die sich unmittelbar aus der Asymmetrie von Ordnungsrelationen ergibt, hat zur Folge, dass man die Items so anordnen können muss, dass

(8) $$x_{vi} = \begin{cases} 1 \text{ wenn } \theta_v > \delta_i \\ 0 \text{ wenn } \theta_v \leq \delta_i \end{cases},$$

worin θ_v den Rangplatz der Vp und δ_i den (Schwierigkeits-) Rangplatz des Items bezeichnet. Die Eigenschaft der Vpn und die Schwierigkeit der Items werden damit auf derselben Ordinalskala gemessen (vgl. Abb. 3).

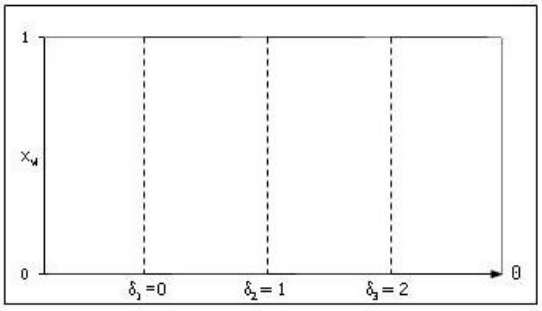


Abbildung 3, Die Guttman-Skala (nach Kempf, 2008, S. 178).

Werden die Rangplätze mit nichtnegativen ganzen Zahlen durchnummeriert, dann ist der Rangplatz einer Vp auf der Guttman-Skala gleich der Anzahl der von ihr gegebenen richtigen Antworten

(9) $$\theta_v = \sum_{i=1}^{k} x_{vi}$$

und die Rangplätze der Items resultieren aus der Häufigkeit, mit der sie positiv beantwortet werden, wobei jenes Item, das am häufigsten gelöst wurde, den Schwierigkeitsrangplatz Null erhält, u.s.w.

Ob die Items tatsächlich alle dieselbe Leistungsdimension messen, ist an einer einfachen (aber sehr restriktiven) empirischen Gesetzmäßigkeit abzulesen. Wenn man die Items nach ihrer Lösungshäufigkeit angeordnet hat, dürfen nämlich (z.B. bei k=3 Items) nur die in Tab. 1 dargestellten Antwortmuster vorkommen:

0	0	0
1	0	0
1	1	0
1	1	1

Tabelle 1, Zulässige Antwortmuster einer Guttman-Skala mit k=3 Items.

Erst die empirische Gesetzmäßigkeit, wonach von den 2^k möglichen Antwortmustern tatsächlich nur die in Tab. 1 dargestellten k+1 Muster vorkommen und nicht etwa ein Muster der Form (1, 0, 1), erlaubt es somit, die Vpn auf einer quantitativen Dimension Θ anzuordnen und so ihre Leistungsfähigkeit (ordinal) zu messen. Eine metrische Information ist in der Guttman-Skala aber nicht enthalten: Die Abstände zwischen den Skalenwerten zweier Items sagen nichts darüber aus, wie stark sich die Items in ihrer Schwierigkeit unterscheiden, und die Abstände zwischen den Skalenwerten zweier Vpn sagen nichts darüber aus, wie sehr sie sich in ihrer Leistungsfähigkeit voneinander unterscheiden.

Dass es auch zweckmäßig ist, von den (qualitativen) manifesten Variablen auf eine (quantitative) latente Dimension schließen zu wollen, auf der man die Vpn gemäß der Anzahl ihrer richtigen Antworten oder – allgemeiner formuliert – aufgrund ihres Summenscores

$$(10) \qquad x_{vo} = \sum_{i=1}^{k} x_{vi}$$

anordnen kann, liegt jedoch keineswegs auf der Hand. So kommt es z.B. in der klinischen Diagnostik weniger auf die Anzahl der Symptome an, welche eine Vp zeigt, als auf die Syndrome, zu welchen diese kombiniert sind, und für die Inhaltsanalyse hat bereits Kracauer (1952) darauf aufmerksam gemacht, dass es nicht die Häufigkeit bestimmter Textmerkmale ist, welche die Bedeutung eines Textes ausmachen, sondern die Muster, welche sie bilden.

Der Schweregrad einer Erkrankung mag zwar als quantitative Dimension gedacht werden, die verschiedenen Krankheiten, für welche die entsprechenden Syndrome einen Indikator darstellen, sind jedoch qualitative (latente) Variablen, und mit dem latenten Gehalt eines Textes verhält es sich nicht anders.

Selbst in der Leistungsdiagnostik kann es ggf. zweckmäßiger sein, nicht die Leistungsmenge (= den Summenscore), sondern die Antwortmuster der Vpn zu betrachten und auch die latente Variable als qualitative Variable zu konzipieren. Als Beispiel betrachten wir ein Experiment von Siegler (1976) zur Entwicklung des formal operationalen Denkens bei Kindern. Darin verwendet Siegler die bereits oben erwähnten Balkenwaagenaufgaben (vgl. Abb. 1) und unterscheidet sechs verschiedene Aufgabentypen, die in Tab. 2 dargestellt sind.[5]

Aufgabentyp	Beispiel			
	Links		Rechts	
	G	D	G	D
Balance (B)	2	3	2	3
Gewicht (G)	1	1	5	1
Distanz (D)	3	2	3	3
Konflikt-Gewicht (K-G)	1	3	4	1
Konflikt-Distanz (K-D)	3	3	6	1
Konflikt-Balance (K-B)	2	2	4	1

Tabelle 2, Die sechs Aufgabentypen nach Siegler (1976). G= Gewicht, D = Distanz.

Um die Testleistung der Vpn zu eruieren, betrachtet Siegler jedoch nicht einfach die Leistungsmenge (z.B. die Anzahl der gelösten Aufgaben), sondern klassifiziert die Vpn aufgrund ihrer manifesten Antwortmuster danach, auf welcher von vier qualitativ verschiedenen Entwicklungsstufen sie operieren.[6] Ähnlich wie bei der Guttman-Skala sind auch in Sieglers Modell nur bestimmte Antwortmuster zulässig (d.h. mit dem Entwicklungsmodell vereinbar). Auch hier setzt die Rückführung der Itemantworten auf eine – in diesem Fall qualitative – latente Variable, somit eine empirische Gesetzmäßigkeit voraus.

Erst die empirische Gesetzmäßigkeit, wonach von den $3^6 = 729$ möglichen Antwortmustern tatsächlich nur die in Tab. 3 dargestellten 28 verschiedenen Muster vorkommen,[7] erlaubt den Rückschluss von den (manifesten) Antwortmustern der Vpn auf die (latente) Entwicklungsstufe, auf welcher sie sich befinden. Obwohl die verschiedenen Stufen einer gewissen Entwicklungslogik folgen (und einen sukzessiven Wissenszuwachs der Vpn abbilden; vgl. Fußnote 5), werden sie damit aber nur auf einer Nominalskala gemessen.

[5] Bei den Balance-Aufgaben sind die Gewichte und die Distanzen auf beiden Seiten der Balkenwaage gleich. Bei den Gewicht-Aufgaben sind die Distanzen dieselben, die Gewichte aber verschieden. Bei den Distanz-Aufgaben verhält es sich umgekehrt: Die Gewichte sind gleich, aber die Distanzen verschieden. Bei den Konflikt-Aufgaben sind Gewichte und Distanzen gegenläufig. Konflikt-Gewicht: Höheres Gewicht überwiegt größere Distanz. Konflikt-Distanz: Größere Distanz überwiegt höheres Gewicht. Konflikt-Balance: Höheres Gewicht und größere Distanz kompensieren sich gegenseitig.

[6] Auf der ersten Entwicklungsstufe wissen die Kinder lediglich, dass das Verhalten der Balkenwaage etwas mit Gewichten zu tun hat. Das größere Gewicht entscheidet. Auf der zweiten Stufe entscheidet ebenfalls das größere Gewicht. Bei gleichen Gewichten wird aber auch bereits die Distanz beachtet. Auf der dritten Stufe spielen Distanz und Gewicht eine gleichwertige Rolle. Widersprechen sie einander, kann das Kind nur raten. Auf der vierten Stufe spielen Distanz und Gewicht eine gleichwertige Rolle. Widersprechen sie einander, so entscheidet das größere Drehmoment (= Produkt aus Gewicht x Distanz).

[7] Die für die Stufen 2 und 4 charakteristischen Muster (1, 1 ,2, 1, 1, 1) bzw. (0, 1, 2, 1, 2, 0) können wegen des Rateprozesses bei den Konfliktaufgaben auch bei Vpn der Stufe 3 auftreten. Um zu entscheiden, welcher Entwicklungsstufe eine Vp zuzuordnen ist, wenn sie eines dieser Antwortmuster zeigt, ist es daher erforderlich, den Vpn von jedem Konfliktaufgabentyp mehrere Aufgaben vorzulegen.

Stufe	Aufgabentyp						Stufe	Aufgabentyp					
	B	G	D	K-G	K-D	K-B		B	G	D	K-G	K-D	K-B
1	0	1	0	1	1	1	Fort-	0	1	2	1	1	1
2	1	1	2	1	1	1	set-	0	1	2	1	1	2
3	0	1	2	0	0	0	zung	0	1	2	1	2	0
	0	1	2	0	0	1	3	0	1	2	1	2	1
	0	1	2	0	0	2		0	1	2	1	2	2
	0	1	2	0	1	0		0	1	2	2	0	0
	0	1	2	0	1	1		0	1	2	2	0	1
	0	1	2	0	1	2		0	1	2	2	0	2
	0	1	2	0	2	0		0	1	2	2	1	0
	0	1	2	0	2	1		0	1	2	2	1	1
	0	1	2	0	2	2		0	1	2	2	1	2
	0	1	2	1	0	0		0	1	2	2	2	0
	0	1	2	1	0	1		0	1	2	2	2	1
	0	1	2	1	0	2		0	1	2	2	2	2
	0	1	2	1	1	0	4	0	1	2	1	2	0

Tabelle 3, Prognostizierte Antwortmuster für 6 Aufgabentypen und 4 Entwicklungsstufen. 0 = „Balance“, 1 = „Gewicht“, 2 = „Distanz“.

4. Statistische Modelle psychologischer Daten

Indem sie nur ganz bestimmte Antwortmuster zulassen, stellen sowohl die Guttman-Skala als auch das Modell von Siegler äußerst restriktive Anforderungen an die Daten, die in der Praxis schon deshalb kaum je erfüllt sind, weil psychologische Daten in der Regel einer Zufallsstreuung unterliegen. Neben den in Tab. 1 bzw. 3 dargestellten *idealtypischen* Antwortmustern werden sich daher auch andere finden, die davon mehr oder minder stark abweichen.

Um diese Zufallsvariation modellieren zu können, geht man in der psychologischen Testtheorie davon aus, dass jeder Vp und jedem Testitem nicht eine bestimmte, feststehende Antwort, sondern eine zufällige Verteilung möglicher Antworten entspricht. Die tatsächlich beobachtete Antwort (x_{vi}) wird somit als die Realisation einer Zufallsvariable X_{vi} betrachtet (vgl. Lazarsfeld, 1950; Novick, 1966) und die statistischen Modelle der Testtheorie sind zwischen zwei Polen angesiedelt:

Den einen Pol bildet das sog. *saturierte Modell* (SM), das davon ausgeht, dass die Antworten einer gegebenen Vp keinerlei Zufallsvariation unterliegen,[8] so dass die Likelihood der Antwortmatrix[9] durch $m^k - 1$ unabhängige Modellparameter (p_g) dargestellt werden kann, welche die Wahrscheinlichkeit beschreiben, mit der eine zufällig herausgegriffene Vp das Antwortmuster $x_g = (x_{g1},...,x_{gk})$ zeigt.

Den anderen Pol bildet das *Pure Random Modell* (PR), das davon ausgeht, dass sich weder die Vpn noch die Items systematisch voneinander unterscheiden, so dass die Zufallsvariable X_{vi} für alle Vpn und alle Items dieselbe (und die gesamte beobachtete Variation der Daten daher reine Zufallsvariation) ist. Die Likelihood der Antwortmatrix hängt dann lediglich von m - 1 unabhängigen Modellparametern ab, welche die (item- und personenunabhängigen) Kategorienwahrscheinlichkeiten $prob\{X_{vi} = x\} = p_{\bullet\bullet x}$ darstellen.

Zwischen diesen beiden Polen, von denen das SM die beste (= maximale Likelihood) und das PR die schlechteste (= minimale Likelihood) aller möglichen Beschreibungen der Daten leistet, liegen die Modelle der *probabilistischen Testtheorie*, welche die Wahrscheinlichkeitsdichte $p_{vix} = prob\{X_{vi} = x\}$ der Antwortvariablen gemäß der Funktionsgleichung

(11) $$p_{vix} = f_{ix}(\theta_v)$$

durch die Ausprägung einer quantitativen oder qualitativen latenten Variable Θ zu erklären versucht. Im Unterschied dazu betrachtet die *klassische Testtheorie* lediglich die Erwartungswerte der Antwortvariablen und spaltet die manifesten Antworten der Vpn gemäß

(12) $$x_{vi} = \tau_{vi} + f_{vi}$$

in zwei Faktoren auf, von denen einer der als True-Score $\tau_{vi} = E(X_{vi})$ bezeichnete Erwartungswert der Antwortvariable und der andere die als Messfehler $f_{vi} = x_{vi} - \tau_{vi}$ bezeichnete Abweichung der beobachteten Antwort von diesem Erwartungswert ist.

[8] Vgl. als Spezialfälle die Guttman-Skala und/oder das Modell von Siegler.
[9] D..h. die Wahrscheinlichkeit, in einer Stichprobe von n Vpn genau jene $n \times k$ Antworten zu beobachten, die tatsächlich vorliegen.

Dass die Items eines Tests oder Fragebogens dasselbe messen, wird in der Klassischen Testtheorie durch das Konzept der essentiellen τ-Äquivalenz operationalisiert, wonach sich die erwarteten Scores je zweier Items (i und j) lediglich durch eine von den Vpn unabhängige Konstante $c_{ij} = \tau_{vi} - \tau_{vj}$ unterscheiden, in welcher der Schwierigkeitsunterschied der beiden Items zum Ausdruck kommt (vgl. Lord & Novick, 1968; Steyer & Eid, 1993).

Graphisch dargestellt bedeutet dies, dass die in Abb. 4 dargestellten Profillinien alle parallel verlaufen.[10]

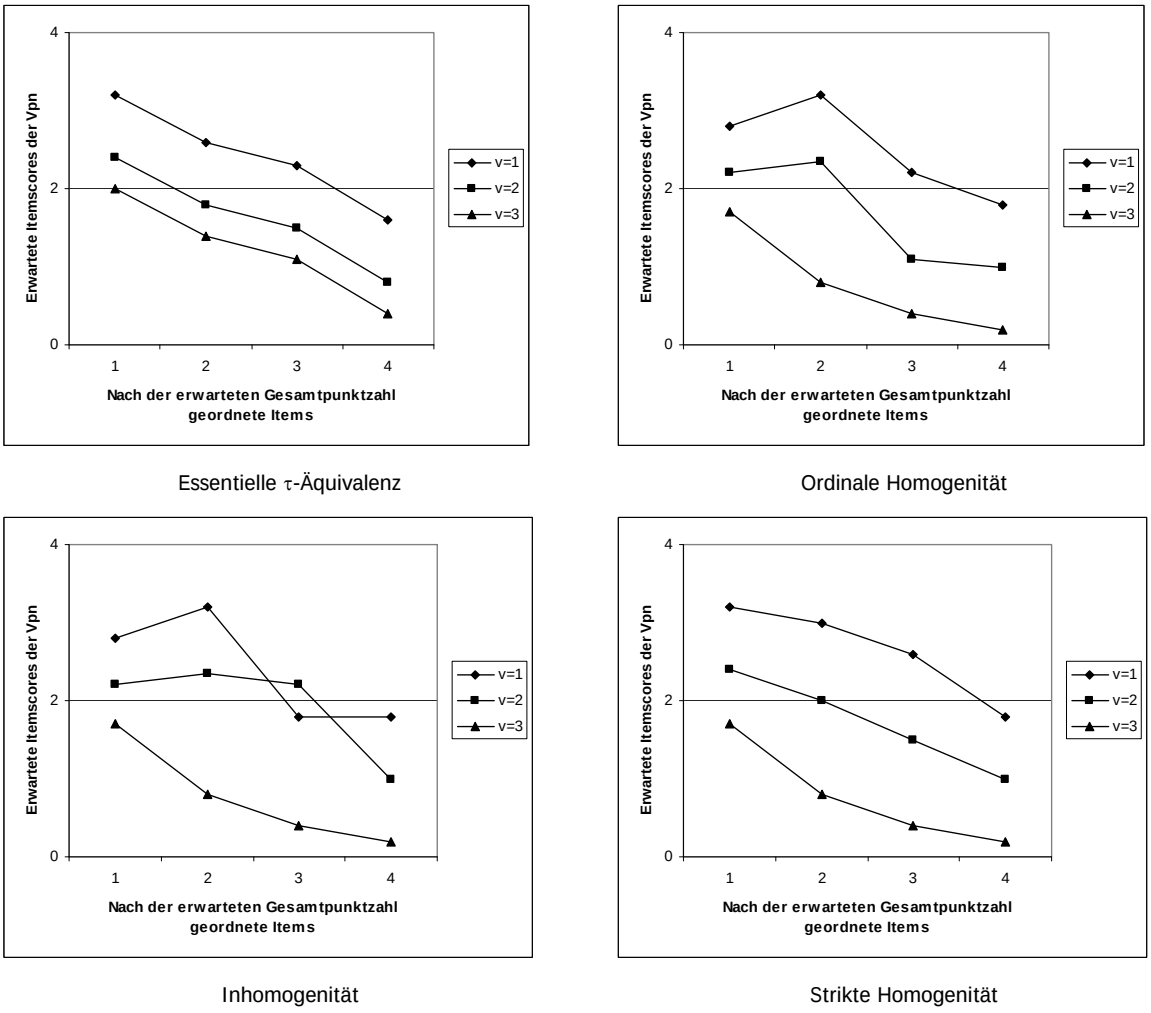


Abbildung 4, Profillinien der Vpn.

In Hinblick auf die Quantifizierung qualitativer Daten kann das Konzept der essentiellen τ-Äquivalenz aber insofern nicht voll befriedigen, als es bereits voraussetzt, dass die Antwortkategorien auf einer metrischen Skala gemessen werden. Zudem ist die Annahme der essentiellen τ-Äquivalenz auch strenger als zur Anordnung der Vpn auf einer latenten Dimension erforderlich. Um die Vpn auf Grundlage ihrer erwarteten Itemscores auf einer latenten Dimension Θ anordnen zu können, genügt es, dass die Regel

(13) $\tau_{vi} > \tau_{wi} \Leftrightarrow v >^{o} w$

unabhängig von der Auswahl der Items stets dieselbe empirische Ordnungsrelation zwischen den Vpn definiert. Ist dies der Fall, so sind die Items *ordinal homogen* und die Profillinien überschneidungsfrei. Überschneiden sich die Profillinien dagegen, so sind die Items *inhomogen* und die Rückführung der Itemantworten auf eine latente Dimension Θ ist nicht möglich.

Selbst wenn die Items alle dieselbe Ordnungsrelation $v >^{o} w$ definieren, bedeutet dies aber noch nicht, dass die Vpn auch auf Grundlage ihrer erwarteten *Summenscores*

[10] Um die Profillinien zu konstruieren, trägt man die erwarteten Itemscores der Vpn (τ_{vi}) auf der Ordinate eines Koordinatensystems ab, auf dessen Abszisse die Items nach der erwarteten Gesamtzahl der bei ihnen erzielten Punkte (τ_{oi}) angeordnet sind, so dass (von links nach rechts) das Item mit der höchsten erwarteten Gesamtpunktzahl zuerst und jenes mit der niedrigsten zuletzt kommt.

(14) $\tau_{vo} = \sum_{i=1}^{k} \tau_{vi}$

in eine Rangreihe gebracht werden können. Damit der Summenscore die gesamte statistische Information darüber enthält, welche Position eine Vp auf dem latenten Kontinuum Θ einnimmt, müssen auch die Items nach ihrer Schwierigkeit geordnet sein, so dass die Regel

(15) $\tau_{vi} > \tau_{vj} \Leftrightarrow i <^{o} j$

unabhängig von der Auswahl der Vpn stets dieselbe Schwierigkeitsordnung der Items definiert und die Antwortwahrscheinlichkeiten durch das Rasch-Modell (RM)

(16) $p_{vix} = \frac{\exp(x\theta_v - \delta_{ix})}{\sum_{y=0}^{m-1} \exp(y\theta_v - \delta_{iy})}$ mit $\delta_{i0} = 0$

(Rasch, 1960, 1961) dargestellt werden können. Ist dies der Fall, so sind die Items *strikt homogen* (oder Rasch-homogen) und die Profillinien sind nicht nur überschneidungsfrei, sondern sie haben darüber hinaus einen monoton fallenden Verlauf. Sind die Antwortkategorien x = 0, 1, ...,m-1 geordnet, so ergeben sich die Kategorienschwierigkeiten (δ_{ix}) in Gleichung (16) als Summe aller Schwellen (α_{ij}), die man überschreiten muss, um in Kategorie x zu antworten, so dass

(17) $\delta_{ix} = \sum_{j=1}^{x} \alpha_{ij}$ für x = 1,...,m-1

(vgl. Andrich , 1978a,b), wobei die Fähigkeit der Vpn (θ_v) und die Schwierigkeit der Schwellen (α_{ij}) auf derselben Differenzenskala gemessen werden (vgl. Abb. 5). Ob die Antwortkategorien tatsächlich auf einer Ordinalskala gemessen werden, kann man schließlich daran ablesen, ob die Schwellenparameter gemäß $\alpha_{i1} < \alpha_{i2} < ... < \alpha_{im-1}$ geordnet sind.

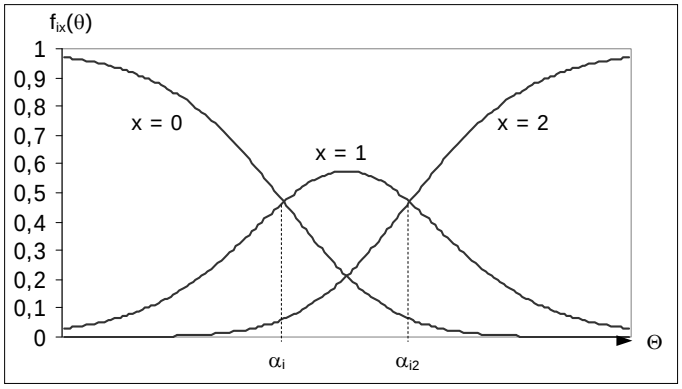


Abbildung 5, Itemfunktionen des Rasch-Modells (nach Kempf, 2008, S. 216)

Da der Summenscore im RM eine erschöpfende Statistik für den Personenparameter θ_v darstellt, teilt er die Vpn in k(m-1) + 1 ordinalskalierte manifeste Klassen ein, so dass die Likelihood der Antwortmatrix unabhängig von den Personenparametern dargestellt werden kann und ausschließlich von den Itemparametern δ_{ix} und den Klassengrößenparametern p_g abhängt, welche die Wahrscheinlichkeit beschreiben, mit der eine zufällig herausgegriffene Vp den Summenscore x_{vo} = g zeigt.

Handelt es sich bei der latenten Variable dagegen um eine qualitative Variable, so ergibt sich das Modell der Latent-Class-Analyse (LCA). Im Unterschied zum RM teilt die LCA die Vpn in h nominalskalierte latente Klassen ein, die durch *Typen von Antwortmustern* (Syndrome) charakterisiert sind und allen Vpn, welche derselben Klasse angehören, dieselben Antwortwahrscheinlichkeiten

(18) $p_{vix} = p_{gix}$ für θ_v = g

zuweisen (vgl. z.B. Abb. 6).

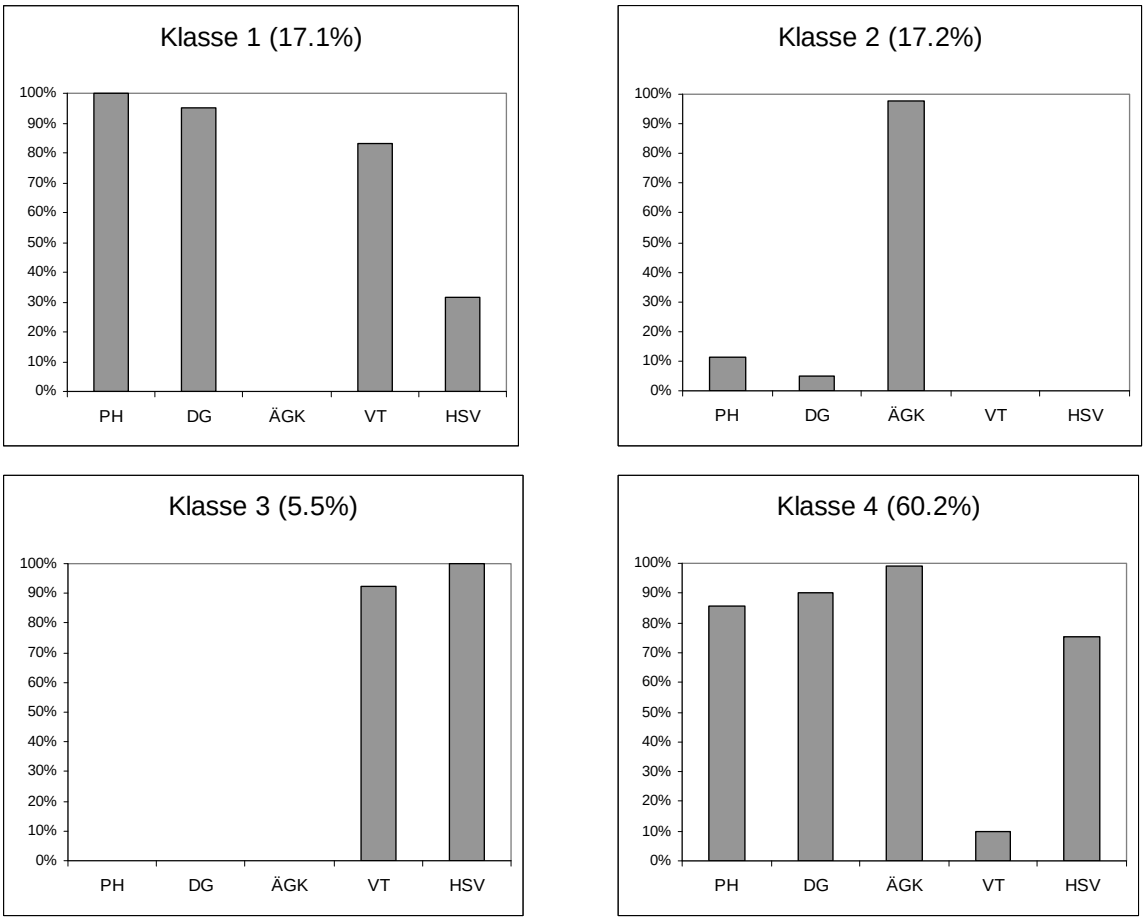


Abbildung 6, Depressive Sympomatik verschiedener Patientengruppen. PH = herabgesetzte Psychomotorik; DG = gehemmter Denkablauf; ÄGK = ängstlich-gequält-klagsame Stimmung; VT = vital-traurige Stimmng; HSV = Hypochondrien, Schuld- und Verarmungsgedanken (nach Kempf, 2008, S. 188).

Die Likelihood der Antwortmatrix hängt dann lediglich von den klassenspezifischen Antwortwahrscheinlichkeiten p_{gix} und den Klassengrößeparametern p_g ab, welche die Wahrscheinlichkeit beschreiben, mit der eine zufällig herausgegriffene Vp der Klasse g angehört. Um die Anzahl der latenten Klassen zu bestimmen, greift man auf verschiedene Informationsmaße wie z.B. auf Akaikes (1987) Informationskriterium

(19) $$AIC = -2\ln(L) + 2n(P)$$

zurück, worin ln(L) den Logarithmus der Likelihood der Datenmatrix und n(P) die Anzahl der zu ihrer Darstellung erforderlichen Modellparameter bezeichnet. Je mehr latente Klassen postuliert werden, desto besser können die Daten beschrieben werden, desto mehr Modellparameter müssen aber auch geschätzt werden. Je kleiner der AIC-Index ausfällt, desto besser ist daher der Kompromiss zwischen Genauigkeit (= große Likelihood) und Sparsamkeit (= wenig Parameter) der Beschreibung der Daten.

Stellt man o.B.d.A. die klassenspezifischen Antwortwahrscheinlichkeiten p_{gix} durch die logistische Funktion

(20) $$p_{gix} = \frac{\exp\left(x\theta_{gi} - \delta_{gix}\right)}{\sum_{y=0}^{m-1}\exp\left(y\theta_{gi} - \delta_{giy}\right)} \qquad \text{mit } \delta_{gi0} = 0,$$

dar (vgl. Rost, 1988), so kann man auch im Rahmen der LCA wieder überprüfen, ob die Antwortkategorien auf einer Ordinalskala gemessen werden. In Analogie zum RM kann man die (klassenspezifischen) Kategorienschwierigkeiten (δ_{gix}) dann als Summe der (klassenspezifischen) Schwellen (α_{gij}) darstellen, die man überschreiten muss, um in Kategorie x zu antworten:

(21) $$\delta_{gix} = \sum_{j=1}^{x}\alpha_{gij} \qquad \text{für } x = 1,\ldots,m\text{-}1.$$

Dass die Antwortkategorien ordinalskaliert sind, zeigt sich nun wieder daran, dass die Schwellenparameter (innerhalb jeder der latenten Klassen) gemäß $\alpha_{gi1} < \alpha_{gi2} < \ldots < \alpha_{gim-1}$ geordnet sind.

Da die in Abb. 4 dargestellten Beziehungen für jede einzelne Vp gelten, gelten sie auch für jede Klasse von Vpn. Dies kann man sich zunutze machen, um die Skalenqualität der latenten Variable zu untersuchen, indem man die Profillinien der erwarteten Itemscores innerhalb der latenten Klassen (τ_{gi}) berechnet.

- Überschneiden sich die Profillinien der Klassen, so ist der Test *inhomogen*, die latente Variable ist *qualitativ* und wird auf einer *Nominalskala* gemessen.
- Sind die Profillinien der Klassen überschneidungsfrei, so ist der Test *ordinal homogen*, die latente Variable ist *quantitativ* und wird auf einer *Ordinalskala* gemessen.
- Sind die Profillinien der Klassen überschneidungsfrei und monoton fallend, so ist der Test *möglicherweise strikt homogen*, die latente Variable ist *quantitativ* und kann evtl. mittels des RM auf einer Differenzenskala gemessen werden.

Ob letzteres tatsächlich der Fall ist, kann jedoch nur entschieden werden, indem man die Modellgeltung des RM überprüft.[11]

Ein Mindesterfordernis für die Modellgeltung - sowohl der LCA als auch des RM - besteht darin, dass das jeweilige Modell die Daten nicht signifikant schlechter beschreibt als das SM. Ob dies zutrifft, kann mittels eines Likelihood-Quotienten-Tests (LR-Test) überprüft werden, dessen Prüfgröße

$$(22) \qquad -2\ln(\lambda) = 2\left|\ln(L_0) - \ln(L_1)\right|$$

asymptotisch nach χ^2 mit df = $n(P_1) - n(P_0)$ Freiheitsgraden verteilt ist. Darin bezeichnen L_0 die Likelihood und $n(P_0)$ die Anzahl der unabhängigen Modellparameter des Modells, dessen Geltung überprüft werden soll (= Nullhypothese). L_1 und $n(P_1)$ bezeichnen jene des SM (= Alternativhypothese).

5. Homogenitätsanalyse

Obwohl RM und LCA deutlich weniger restriktiv sind als die in Abschnitt 3 beschriebenen deterministischen Modelle, stellen sie immer noch recht strenge Anforderungen an die Daten. Namentlich das RM erweist sich häufig als zu streng, so dass man davon ausgehen muss, dass die in der quantitativen Psychologie oft routinemäßig und ohne Überprüfung ihrer Voraussetzungen verwendeten Summenscores in der Regel mit einem Verlust an statistischer Information einhergehen und die latente Variable nicht einmal auf einer Ordinalskala messen. Insbesondere Einstellungsfragebögen sind nur selten strikt homogen und erfüllen selbst dann, wenn die Items nahezu synonym lauten, bestenfalls das Kriterium der ordinalen Homogenität.

Auch und gerade dann, wenn die Annahme der strikten Homogenität verworfen werden muss, kann die psychometrische Analyse eines Fragebogens mittels RM und LCA jedoch wichtige Aufschlüsse über die Struktur der Einstellung liefern, welche er erfasst.

Das folgende Datenbeispiel geht auf eine Stichprobe von n = 411 Vpn zurück, welche drei Items zur Messung von sekundärem Antisemitismus auf einer Skala von 0 = „völlige Ablehnung“ bis 4 = „völlige Zustimmung“ bewerteten. Die Items lauteten:

1. „Jahrzehnte nach Kriegsende sollten wir nicht mehr so viel über die Judenverfolgung reden, sondern endlich einen Schlussstrich unter die Vergangenheit ziehen“.
2. „Man sollte endlich mit dem Gerede über unsere Schuld gegenüber den Juden Schluss machen“.
3. „Das Deutsche Volk hat (k)eine besondere Verantwortung gegenüber den Juden“.

Tab. 4 gibt die Godness-of-Fit-Statistiken des PR, des RM, der LCA-Lösungen für h = 1 bis 6 latente Klassen (LCh) und des SM wieder.

Ersichtlich kann das PR die Daten nur sehr schlecht erklären (kleinste Likelihood, größtes AIC) und muss aufgrund des hoch signifikanten LR-Tests verworfen werden. Das RM erklärt die Daten zwar etwas besser (größere Likelihood, kleineres AIC), muss aufgrund des hoch signifikanten LR-Tests aber ebenfalls verworfen werden. Unter den LCA-Lösungen beschreibt die 5-Klassen Lösung die Daten am besten (minimales AIC) und ebenso gut wie das SM (nicht-signifikanter LR-Test), das zwar eine etwas größere Likelihood, aber ein schlechteres AIC zeigt.

[11] Zu den verschiedenen Modelltests, mittels derer man die Geltung des RM untersuchen kann, vgl. z.B. die Lehrbücher von Rost (1996) und Kempf (2008).

Modell	ln(L)	n(P)	-2ln(λ)	df	p	AIC
PR	-1933,93	4	643,52	120	< 0.001	3875,86
RM	-1728,14	23	231,95	101	< 0.001	3502,28
LC1	-1899,37	12	574,41	112	< 0.001	3822,74
LC2	-1742,37	25	260,41	99	< 0.001	3534,74
LC3	-1693,22	38	162,12	86	< 0.001	3462,45
LC4	-1668,18	51	112,03	73	0.002	3438,36
LC5	-1638,66	64	52,99	60	0.7274	3405,32
LC6	-1636,28	77	48,22	47	0.4232	3426,55
SM	-1612,17	124				3472,34

Tabelle 4, Goodness-of-Fit-Statistiken

Die Skala ist daher nicht strikt - sondern bestenfalls ordinal - homogen und wenn man den Summenscore als Maß für die sekundär antisemitische Einstellung der Vpn verwendet, so gehen

$$LI = 100 \times \frac{\ln(L_{LC5}) - \ln(L_{RM})}{\ln(L_{LC5}) - \ln(L_{PR})} = 30{,}3\%$$

der durch die 5-Klassen Lösung der LCA verwerteten Information verloren.

Wie die in Abb. 7 dargestellten Profillinien der Klassen zeigen, ist die Skala ordinal homogen, so dass jedes der drei Items dieselbe Rangordnung zwischen je zwei Klassen definiert (überschneidungsfreie Profillinien). Die Profillinien verlaufen jedoch nicht monoton, was darauf hinweist, dass sich die Schwierigkeitsrelation zwischen den Items mit zunehmender Ausprägung antisemitischer Einstellungen verschiebt (Verletzung der strikten Homogenität). Während Vpn, die den sekundär-antisemitischen Aussagen (eher) ablehnend gegenüberstehen (Klasse 1 und 2), die Forderung, man solle „mit dem Gerede" über unsere Schuld gegenüber den Juden endlich Schluss machen (Item 2), deutlich stärker zurückweisen als die schlichte Schlussstrichforderung, die in Item 1 zum Ausdruck kommt, ist es bei ausgewachsenen Antisemiten (Klasse 5) gerade umgekehrt. Sie stimmen gerade der Aussage in Item 2 stärker zu, welche nicht nur einen Schlussstrich unter die Vergangenheit fordert, sondern die Schuldfrage als bloßes „Gerede" abwehrt und sich damit in die Nähe von Holocaust-Leugnung begibt. Diese für die Diagnose antisemitischer Einstellungen höchst relevante Information geht jedoch verloren, wenn man lediglich die Summenscores der Vpn betrachtet.

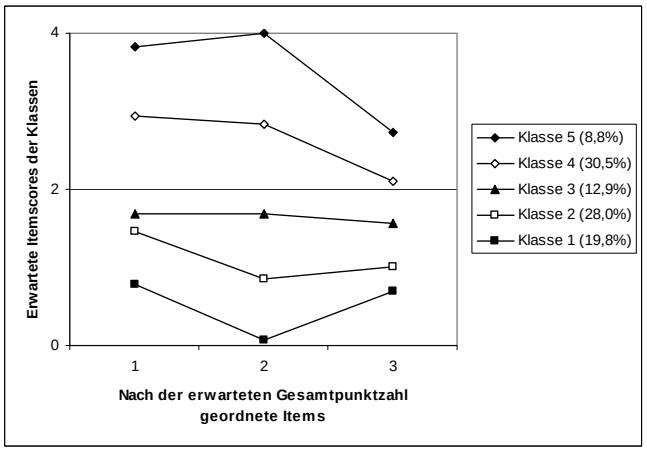


Abbildung 7, Profillinien der latenten Klassen

6. Der Loss-of-Information-Index

Für die statistische Weiterverrechnung von Fragebogendaten – z.B. mittels Strukturgleichungsmodellen – mag es oft zweckmäßig erscheinen, an der Verwendung des Summenscores festzuhalten, obwohl die Voraussetzungen der Scorebildung nicht erfüllt sind. Wenn man dies tut, sollte man aber wenigstens Rechenschaft darüber ablegen, welcher Informationsverlust damit einhergeht.

Der bereits oben angewendete **L**oss-of-**I**nformation-Index

$$(23) \qquad LI = 100 \times \frac{\ln(L_{LCA}) - \ln(L_{RM})}{\ln(L_{LCA}) - \ln(L_{PR})}$$

leistet dies, indem er den mit der Scorebildung einhergehenden Informationsverlust $\ln(L_{LCA}) - \ln(L_{RM})$ zu der durch die Klassenbildung verwerteten statistischen Information $\ln(L_{LCA}) - \ln(L_{PR})$ in Beziehung setzt.

Cronbachs Koeffizient Alpha, den man üblicherweise verwendet, um die interne Konsistenz von Tests oder Fragebögen abzuschätzen, ist für die Beurteilung der Angemessenheit der Scorebildung kein geeignetes Maß. Er liefert lediglich eine untere Schranke für die Reliabilität der Scores (vgl. Lord & Novick, 1968). Reliabilität und Homogenität sind jedoch zwei völlig unabhängige Konzepte.

Da der Koeffizient Alpha die Reliabilität genau dann exakt wiedergibt, wenn die Items (zumindest) essentiell τ-äquivalent sind, könnte man ihn zwar zur Abschätzung der essentiellen τ-Äquivalenz benutzen, indem man ihn zur Reliabilität in Beziehung setzt. In der Praxis scheitert dies jedoch daran, dass alle anderen Reliabilitätsmaße noch strengere Voraussetzungen treffen. Und selbst wenn die Items essentiell τ-äquivalent sind, geht die Scorebildung immer noch mit einem Verlust an statistischer Information einher. Ohne statistischen Informationsverlust können die Scores nur dann verwendet werden, wenn die Items im Sinne des RM strikt homogen sind.

Literatur

Akaike, Hirotugu (1987). Factor Analysis and AIC. Psychometrika, 52, 317-332.

Andrich, David (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 2, 581-594.

Andrich, David (1978b). A rating formulation for ordered response categories. Psychometrika, 43, 4, 561-573.

Cronbach, Lee.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Guttman, Louis (1950). The basis of scalogram analysis. In: Stouffer, Samuel.A., Guttman, Louis, Suchman, Edward A., Lazarsfeld, Paul F., Star, Shirley.A. & John A. Clausen (eds.). Studies in social psychology in world war II, Vol. IV. Measurement and Prediction. Princeton, N.J.: Princeton Universtity Press, 60-90.

Inhelder, Bärbel & Jean Piaget. (1958). The growth of logical thinking from childhood to adolescence. New York: Basic Books.

Kempf, Wilhelm (2008). Forschungsmethoden der Psychologie. Zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik. Band 2: Quantität und Qualität. Berlin: regener.

Kempf, Wilhelm (2009). Forschungsmethoden der Psychologie. Zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik. Band 1: Theorie und Empirie. 3. Auflage. Berlin: regener.

Kracauer, Siegfried (1952). The challenge of qualitative content analysis. Public Opinion Quaterly, 16, 631-642.

Lazarsfeld, Paul F. (1950). Logical and mathematical foundations of latent structure analysis. In: Stouffer, Samuel.A., Guttman, Louis., Suchman, Edward A., Lazarsfeld, Paul F., Star, Shirley A. & J.A. Clausen (eds.). Studies in social psychology in world war II, Vol. IV. Measurement and Prediction. Princeton/N.J.: Princeton University Press, 362-412.

Lord, Frederick .M. & Melvin R. Novick (1968). Statistical theories of mental test scores. Reading (Mass.): Addison-Wesley.

Novick, Melvin R. (1966). The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3, 1-18.

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

Pfanzagl, Johann (1968). Theory of measurement. Würzburg: Physica.

Rasch, Georg (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche.

Rasch, Georg (1961). On general laws and the meaning of measurement in psychology. Berkeley: University of California Press.

Rost, Jürgen (1988). Quantitative und qualitative probabilistische Testtheorie. Bern: Huber.

Rost, Jürgen (1996). Lehrbuch Testtheorie Testkonstruktion. Bern: Huber.

Siegler, Robert S. (1976). Three aspects of cognitive development. Cognitive Psychology, 8, 481-520.

Steyer, Rolf & Michael Eid (1993). Messen und Testen. Ein Lehrbuch. Berlin. Springer.

Suppes, Patrick., Zinnes, Joseph.L. (1963). Basic measurement theory. In: Luce, R. Duncan., Bush, Robert R. & Eugene Galanter (eds.). Handbook of mathematical psychology. New York: Wiley, 1-76.

Warm, Thomas A. (1989). Weighted likelihood estimation of ability in item response models. Psychometrika, 54, 427-450.