

Neue Ergebnisse zur Jagodzinski-Langeheine-Debatte in der ZA-Information 1987

Langeheine, Rolf; Pol, Frank van de; Pannekoek, Jeroen

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:

Langeheine, R., Pol, F. v. d., & Pannekoek, J. (1995). Neue Ergebnisse zur Jagodzinski-Langeheine-Debatte in der ZA-Information 1987. *ZA-Information / Zentralarchiv für Empirische Sozialforschung*, 37, 38-50. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-201119>

Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Neue Ergebnisse zur Jagodzinski-Langeheine-Debatte in der ZA-Information 1987

von Rolf Langeheine¹, Frank van de Pol und Jeroen Pannekoek²

Zusammenfassung

In den beiden Heften der ZA-Information 1987 haben Jagodzinski und Langeheine eine Debatte darüber geführt, ob Chi-Quadrat basierte Teststatistiken zur Beurteilung der Anpassungsgüte einer Reihe von latent class Modellen für die durch sparseness gekennzeichnete 3×3×3 Tabelle des Postmaterialismus Panels sinnvoll sind. In dieser Arbeit berichten wir über eine Reevaluation der im Zentrum der Debatte stehenden Modelle mit Hilfe des non-naïven, parametrischen Bootstrap.

Abstract

This paper is tied to a debate of Jagodzinski and Langeheine documented in the two 1987 issues of the ZA-Information about whether chi-square based statistics make sense in assessing model fit of a variety of latent class models fit to the 3×3×3 sparse data postmaterialism panel table. The models which have been focused on in the debate are reevaluated using non-naïve, parametric bootstrap procedures.

1. Zur Erinnerung

In der ZA-Information 19 ist *Jagodzinski* (1986) der Frage nachgegangen, wie man die Zuverlässigkeit materialistischer und postmaterialistischer Antworten mit Hilfe von Modellen für nominalskalierte Daten beurteilen kann. Die Datenbasis dazu bestand aus dem sog. Postmaterialismus Panel. Kategorisierungen in einer 3kategorialen Variable mit den Kategorien

¹ Dr. **Rolf Langeheine** ist wissenschaftlicher Mitarbeiter des Instituts für die Pädagogik der Naturwissenschaften an der Universität Kiel, Abt. Päd.-Psychol. Methodenlehre, Olshausenstr. 62, D-24098 Kiel

² Dr. **Frank van de Pol** und Dr. **Jeroen Pannekoek** sind wissenschaftliche Mitarbeiter von Statistics Netherlands, Department of Statistical Methods, P.O. Box 959, NL-2270 AZ Voorburg. Die in dieser Arbeit vertretenen Sichtweisen sind die der Autoren. Sie decken sich nicht notwendigerweise mit denen von Statistics Netherlands.

MAT (Personen, die als Materialisten zugeordnet wurden), PMA (Postmaterialisten) und MIX (einem Mischtyp) lagen für wiederholte Messungen an drei Zeitpunkten für eine Stichprobe von 152 Personen vor.

Die resultierende 3×3×3 Kontingenztabelle hat *Jagodzinski* mit drei Modellen konfrontiert, die sich in der Annahme hinsichtlich der Formalisierung unzuverlässigen Antwortverhaltens unterscheiden.

Modell M1 unterstellt unzuverlässiges, d.h. rein zufälliges Antwortverhalten bei allen Personen.

Modell M2 entspricht dem ursprünglichen Black & White Modell von *Converse* (1964). Dieses Modell nimmt an, daß sich die Stichprobe in zwei Gruppen aufteilen läßt, die in ihrem Antwortverhalten über die Zeit differieren. Es gibt eine Gruppe von Personen mit völlig zuverlässigen und stabilen Wertorientierungen (von Materialisten, Postmaterialisten und dem Mischtyp) und eine zweite Gruppe von Personen, die nur Zufallsantworten geben. *Jagodzinski* hat diese beiden Modelle verworfen und das Modell M3 akzeptiert.

Modell M3 erweitert Modell M2, indem für die drei zeitstabilen Typen bestimmte Antwortunsicherheiten zum Zeitpunkt $t = 1$ zugelassen werden: Materialisten dürfen auch MIX-Antworten geben, aber keine PMA-Antworten. Ebenso dürfen Postmaterialisten PMA- und MIX- aber keine MAT-Antworten geben. Für den MIX-Typ sind alle drei Kategorien zugelassen.

Zur Schätzung dieser Modelle hat *Jagodzinski* eine relativ anspruchslose Methode (die OLS-Schätzung) verwendet und auf statistische Tests zur Prüfung der Modellgültigkeit verzichtet. Er hat dies damit begründet, daß die Fallzahl ($N = 152$) zu gering ist, was zu zahlreichen Nullzellen in der Kontingenztabelle führt.

Langeheine (1987a) hat dagegen für die Verwendung einer gut etablierten, anspruchsvollen Schätzmethode (die ML-Schätzung) und die Durchführung statistischer Tests auf Modellverträglichkeit (mittels Likelihood-Quotienten Test) argumentiert. Er hat zugleich gezeigt, wie sich *Jagodzinski*'s Modelle sowie eine Anzahl weiterer Modelle mittels LCA (Latent Class Analyse; vgl. *Goodman* 1974; zu neueren Übersichten siehe *Langeheine* 1988; *Langeheine* und *Rost* 1993; *Clogg* 1995) formalisieren lassen.

In Anbetracht einer signifikanten Likelihood-Quotienten Chi-Quadrat Statistik ($L^2 = 31.81$, Freiheitsgrade = $df = 19$, $p = .033$) hat *Langeheine* das Modell M3 verworfen.

Er hat Modell M4 vorgeschlagen, das sich von Modell M3 darin unterscheidet, daß es Antwortunsicherheit wie in Modell M3 erlaubt, aber an allen drei Zeitpunkten und nicht nur zu $t = 1$. Gegenüber dem 4-Klassen-Modell M3 hat das Modell M4 nur drei Klassen, da alle

Antwortunsicherheit durch diese drei Klassen abgebildet wird, eine vierte Klasse sich somit erübrigt. Modell M4 erwies sich als gut passend für die Daten ($L^2 = 13.81$, $df = 12$, $p = .312$).

Modell M4 kann auch als Meßfehlermodell interpretiert werden, in dem eine latente kategoriale Variable durch drei meßfehlerbehaftete Indikatoren (hier Messungen zu drei Zeitpunkten) gemessen wird. Die Analyse der geschätzten Fehlerraten zeigte, daß diese zwar über die Zeit abnehmen, sich aber nur unerheblich unterscheiden. Mit Modell M6 hat **Langeheine** daher ein Modell angepaßt, das gegenüber Modell M4 erheblich sparsamer ist, da ein zeitkonstantes Meßmodell (bedingte Wahrscheinlichkeiten) und somit identische Fehlerraten für alle Zeitpunkte angenommen werden. Modell M6 erwies sich sowohl als mit den Daten verträglich ($L^2 = 21.24$, $df = 20$, $p = .383$) als auch als nicht signifikant schlechter als Modell M4, in das es geschachtelt ist (L^2 Differenz = 7.43, $df = 8$).

Im folgenden gehen wir nur auf die Modelle M2, M3, M4 und M6 ein, da sie im Zentrum der Debatte stehen. Hinsichtlich der detaillierten Beschreibung dieser Modelle sei auf **Jagodzinski** (1986) und **Langeheine** (1987a) verwiesen.

Im Kern der Debatte (**Jagodzinski** 1987a, b; **Langeheine** 1987b) stand die Frage, ob die Überprüfung der Modellgültigkeit mittels Chi-Quadrat Statistiken bei dünn besetzten Tabellen (sparse data) angemessen ist. Im folgenden Abschnitt gehen wir daher zunächst auf dieses Problem und eine Reihe von ad hoc oder mehr oder wenig theoretisch wohl definierte Strategien ein, die für derartige Fälle vorgeschlagen wurden. Im Anschluß daran stellen wir das Bootstrap Verfahren zur Bestimmung von p-Werten für Chi-Quadrat Maße der Anpassungsgüte bei der Analyse kategorialer Daten vor und berichten schließlich über eine Reevaluation der Modelle M2, M3, M4 und M6 mittels Bootstrap.

2. Sparse data und Strategien zur Prüfung der Modellgültigkeit

Chi-Quadrat basierte Statistiken sind die am häufigsten verwendeten Statistiken in der Analyse kategorialer Daten, sowohl zur Prüfung der Anpassungsgüte spezifischer Modelle bei der Modellselektion als auch beim Vergleich von Modellen. Die bekanntesten zwei solcher Statistiken sind das Pearson Chi-Quadrat (X^2) und das Likelihood-Quotienten Chi-Quadrat (L^2). Wie **Read** und **Cressie** (1988) zeigen, gehören diese beiden sowie eine Reihe weiterer Statistiken zur Familie der sog. power divergence statistics mit

$$2[\lambda(\lambda + 1)]^{-1} \sum_{y=1}^Y f_y \left[\left(f_y / \hat{F}_y \right)^\lambda - 1 \right]$$

wobei Y die Anzahl der Zellen der Kontingenztabelle angibt, f_y für die beobachtete Häufigkeit von Zelle y steht und \hat{F}_y für die entsprechende nach einem Modell erwartete Häufigkeit

Die verschiedenen Statistiken unterscheiden sich lediglich durch den Parameter λ . L^2 ergibt sich für $\lambda \rightarrow 0$. Für $\lambda = 1$ resultiert X^2 . *Read* und *Cressie* empfehlen eine Kompromiß-Statistik (im folgenden mit RC bezeichnet) mit $\lambda = 2/3$. Wie sie zeigen, hat RC eine Verteilung, die auch für erwartete Häufigkeiten von mindestens 1 durch die Chi-Quadrat Verteilung generell gut approximiert wird. Ihre Vermutung geht dahin, daß RC auch bei vielen erwarteten Häufigkeiten kleiner 1 eine gute Wahl ist.

Das Problem bei diesen Statistiken ist, daß sie asymptotisch der Chi-Quadrat Verteilung folgen, aber die asymptotische Annäherung gilt nur, wenn die Stichprobengröße gegen Unendlich geht, d.h., wenn die Zahl der Beobachtungen in jeder Zelle der Tabelle "groß" ist. Darüber, wie groß "groß" sein sollte, um groß genug zu sein, hat es immer wieder Debatten gegeben (vgl. *Read* und *Cressie* 1988). Es gibt eine ganze Reihe von Monte Carlo Studien, in denen die Chi-Quadrat Approximation für L^2 und X^2 bei log-linearen Modellen untersucht wurde, in jüngerer Zeit auch für LCA-Fragestellungen (*Everitt* 1988; *Holt* und *Macready* 1989; *Collins, Fidler, Wugalter* und *Long* 1993). *Collins* et al. haben z.B. das Verhalten von L^2 , RC und X^2 bei unterschiedlich stark besetzten Zellen untersucht, indem sie ein bekanntes, wahres latentes Klassen-Modell auf einen Satz von Zufallsdaten angepaßt haben. Die Schlußfolgerung aus ihrer Studie ist, daß "none of the three goodness-of-fit indices is a clear choice for latent class model evaluation" (*Collins* et al. 1993 : 385). In ihrer Studie unterschieden sich sowohl die Mittelwerte der simulierten Verteilungen (besonders für L^2 und RC) als auch die Standardabweichungen (besonders für X^2) von den Erwartungswerten nach der Chi-Quadrat Verteilung.

Das Problem all solcher Simulationsstudien ist, daß ein Design festgelegt werden muß, in dem bestimmte Faktoren variiert werden. Die Anzahl dieser Faktoren ist bei latenten Klassen-Modellen ungleich größer als bei log-linearen Modellen. Die Folge ist, daß die Übertragung der Ergebnisse solcher Studien auf den konkreten Fall zumindest zweifelhaft ist, da es für das gerade zur Diskussion stehende Modell in der Regel keine entsprechenden Monte-Carlo Ergebnisse gibt.

Sparse data sind also ein relativ altes Problem, das sich heute um so mehr stellt, da mit moderner Software und leistungsfähigen PCs auch sehr große Tabellen mit vielen Variablen in angemessener Zeit analysiert werden können. Solche Tabellen können mehrere tausend oder sogar millionen und mehr Zellen haben, von denen oft nur ein Bruchteil besetzt ist. Das eigentliche Problem dabei sind nicht beobachtete Nullzellen oder gering besetzte Zellen, sondern niedrige Erwartungswerte. Während dies für die Parameterschätzung (auch bei der LCA) im allgemeinen keine Probleme aufwirft, ist die Beurteilung der Güte der Modellanpassung mittels Chi-Quadrat Statistiken zweifelhaft. Verschiedene Autoren haben deshalb unterschiedliche Strategien vorgeschlagen, um diesem Problem zu begegnen.

Eine Möglichkeit besteht darin, Zellen zusammenzufassen, die einen bestimmten Erwartungswert unterschreiten (vgl. z.B. *Andersen* 1982, 1985, 1988; *Rost* 1988). Es gibt allerdings viele verschiedene Möglichkeiten der Zusammenfassung. Man kann z.B. alle Zellen zusammenfassen, die einen bestimmten kritischen Wert unterschreiten. Oder man kann solange zusammenfassen, bis der kritische Wert überschritten ist und darauf neu beginnen. Dabei kann man sich noch überlegen, ob nach Zufall zusammengefaßt wird, ob dies für benachbarte Zellen erfolgen soll, oder ob man dabei Modelleigenschaften berücksichtigt. *Andersen* faßt z.B. im Fall des Rasch Modells nur Zellen mit demselben Rohscore zusammen. Der Effekt solcher Zusammenfassungen ist allerdings nicht systematisch untersucht worden.

Viele Forscher verzichten völlig auf statistische Tests und verwenden dagegen deskriptive Fit-Indizes. Dies gilt nicht nur für die Analyse kategorialer Daten mittels log-linearer, latent class oder latent trait Modelle, sondern auch für die Analyse kontinuierlicher Daten, z.B. mit Kovarianz-Struktur Modellen. Informationstheoretische Indizes wie AIC (*Akaike* 1974; *Bozdogan* 1987), BIC (*Schwarz* 1978) oder der index of dissimilarity D (vgl. *Shockey* 1988) werden von neueren Programmen zur Analyse kategorialer Daten routinemäßig ausgegeben. Das Problem mit derartigen Indizes ist allerdings, daß sie lediglich helfen können, das "beste" Modell aus den Modellen zu identifizieren, die untersucht wurden. Bislang gibt es keine Möglichkeit der präzisen Bestimmung eines "guten" Wertes. *Langeheine, Stern* und *van de Pol* (1994) berücksichtigen daher die kombinierte Information aus statistischen Tests und deskriptiven Indizes, vorausgesetzt, verschiedene Chi-Quadrat basierte Statistiken (L^2 , RC und X^2) führen alle zu derselben Entscheidung über die Anpassung eines Modells.

Eine dritte Möglichkeit besteht darin, die Referenzverteilung von Chi-Quadrat Statistiken bei einem gegebenen Modell zu simulieren. Genau dies haben bereits *Aitkin, Anderson* und *Hinde* (1981) getan, um zu einer Entscheidung über die angemessene Zahl latenter Klassen für eine Tabelle von 38 dichotomen Variablen bei einer Stichprobe von lediglich 468 Personen zu gelangen. Obwohl *Aitkin* et al. eine Arbeit von *Hope* (1968) zitieren, und ihre Simulation somit einer parametrischen Bootstrap Prozedur zu entsprechen scheint, geben sie nur wenig Details über das genaue Vorgehen. Ihre Ergebnisse beruhen zudem auf sehr wenigen (jeweils 19) Simulationen, da für eine größere Zahl von Zufallsdatensätzen erhebliche Computerzeit benötigt worden wäre. Offensichtlich waren extrem aufwendige Rechenzeiten der Grund dafür, daß diese Prozedur für viele Jahre als nicht realistisch angesehen wurde. Inzwischen hat sich die Situation drastisch verändert, so daß derartige Simulationen mit schnellen PCs in angemessener Zeit durchgeführt werden können. Im folgenden Abschnitt stellen wir ein entsprechendes Bootstrap Verfahren vor.

3. Die Bestimmung von p-Werten für Chi-Quadrat Statistiken mittels Bootstrap

Wird die Angemessenheit einer theoretischen Verteilung zur Prüfung einer empirisch ermittelten Teststatistik in Frage gestellt, so können Resampling Methoden eingesetzt werden mit dem Ziel, die Verteilung dieser Statistik zu identifizieren. Zur wiederholten Ziehung von Stichproben gibt es verschiedene Möglichkeiten (vgl. z.B. *Efron* 1982; *Efron* und *Tibshirani* 1993). Beim Jackknife werden einzelne Fälle ausgelassen. Beim Bootstrap werden vollständige neue Stichproben gezogen. Bei der Methode balancierter halber Stichproben wird die Stichprobe auf vielerlei Art halbiert. Das Bootstrap ist für die Modelltestung besonders gut geeignet.

PANMARK (*van de Pol, Langeheine* und *de Jong* 1991) ist ein Programm zur Analyse von latent class und Markov Modellen. Die jüngste Version dieses Programms enthält eine Option zur Bestimmung von p-Werten für die Chi-Quadrat Teststatistiken L^2 , RC und X^2 mittels Bootstrap. Die in PANMARK implementierte Bootstrap Methode entspricht dem sog. non-naive bootstrapping, das *Bollen* und *Stine* (1992) für Kovarianz Struktur Modelle vorgestellt haben. Der wesentliche Unterschied zwischen naivem und nicht-naivem Bootstrap besteht darin, daß bei ersterem Zufallsstichproben aus den Daten gezogen werden, bei letzterem jedoch aus den (erwarteten) Modellhäufigkeiten eines zur Diskussion stehenden Modells. *Bollen* und *Stine* zeigen, warum es falsch (naiv) ist, Stichproben aus den Daten zu ziehen. Die Daten sind eine Stichprobe aus der Population. Zieht man wiederum eine Stichprobe aus dieser Stichprobe, so erfolgt zweifache Ziehung. Für den asymptotischen Fall konnten *Bollen* und *Stine* zeigen, daß diese Prozedur eine Verteilung generiert, deren erwarteter Mittelwert (z.B. $E(L^2)$) doppelt so groß ist wie er sein sollte: beim naiven Bootstrap ist $E(L^2) = 2df$. Werden die Stichproben dagegen aus den Modellhäufigkeiten gezogen, so ist der Erwartungswert der Verteilung im asymptotischen Fall korrekt: $E(L^2) = df$. Auf den Unterschied dieser beiden Arten von Stichprobenziehung gehen verschiedene andere Autoren ein. Was *Bollen* und *Stine* non-naive vs. naive bootstrap bezeichnen, nennen *Efron* und *Tibshirani* (1993) parametric vs. non-parametric bootstrap und *Collins* et al. (1993) Monte Carlo sampling vs. bootstrap resampling.

Eine vereinfachte Monte Carlo Prozedur wurde bereits von *Hope* (1968) vorgestellt, die sie *Barnard* (1963) zuschreibt. Diese Prozedur wird als vereinfacht bezeichnet, weil nur eine geringe Anzahl von Stichproben gezogen wird (zur Entscheidung auf einem α -Niveau von 5% sind nur 19 Stichproben nötig). In den 60er Jahren war dies nötig zur Minimierung der Rechenzeit (vgl. die Anmerkungen zu *Aitkin* et al. im vorigen Abschnitt). *Hope* zeigt allerdings, daß die Power dieser Prozedur mit zunehmender Zahl von Stichproben steigt, was intuitiv verständlich ist.

Die in PANMARK implementierte Prozedur arbeitet wie folgt:

- 1) Es wird angenommen, daß ein Modell wahr ist. Die Parameter werden mittels ML-Methode geschätzt und aus ihnen werden die Modellhäufigkeiten berechnet. Durch Vergleich von Daten (f_y) und Modellhäufigkeiten (\hat{F}_y) resultieren die Teststatistiken L^2 , RC und X^2 .
- 2) Die unter dem Modell geschätzten Proportionen $\hat{\mathbf{P}} = \hat{F}_y / N$ werden als Populations-Proportionen betrachtet.
- 3) Aus dieser multinomialen Verteilung $\hat{\mathbf{P}}$ mit bekannten Parametern wird eine Zufallsstichprobe gezogen.
- 4) Für diese Stichprobe werden dasselbe Modell angepaßt und die Teststatistiken berechnet.
- 5) Diese Prozedur wird K-mal wiederholt, so daß für jede Statistik eine Verteilung resultiert.
- 6) Durch Vergleich der unter 1) berechneten Statistik mit der unter 5) simulierten Verteilung resultiert ein Bootstrap p-Wert, der angibt, wie wahrscheinlich es ist, daß das Modell dieser Verteilung entstammt. Für kleines p (große Teststatistik) wird das Modell verworfen.

Die technischen Details der Stichprobenziehung werden in *Langeheine, Pannekoek* und *van de Pol* (1995) beschrieben.

Die Validität dieser Prozedur sehen wir dadurch bestätigt, daß für Tabellen mit gut besetzten Zellen (oder auch sehr geringem Anteil von sparseness) in etwa die gleichen p-Werte resultieren wie nach der theoretischen Chi-Quadrat Verteilung. Wir weisen allerdings darauf hin, daß die Bootstrap p-Werte im Fall von dünn besetzten Tabellen nicht ohne Bias sind, wenngleich sie viel besser sind als die Werte nach der theoretischen Chi-Quadrat Verteilung (*Efron* 1982). Der Bias rührt daher, daß anstelle der Populations-Proportionen die geschätzten Proportionen verwendet werden. Diese Proportionen werden dann ungenau geschätzt, wenn die Modellparameter ungenau geschätzt werden, z.B. bei Überparametrisierung. Ein Indikator für Überparametrisierung sind geschätzte Parameter an den Grenzen des erlaubten 0-1 Bereichs. Ein Modell sollte also nicht so viele Parameter enthalten, daß ihre Werte ungenau geschätzt werden. Bei sparsamen Modellen scheint dieser Bias kein großes Problem zu sein.

4. Ergebnisse

PANMARK fordert die Spezifikation von Abbruchkriterien für die Anzahl der zu ziehenden Bootstrap Stichproben. Einerseits kann eine Maximalzahl von Stichproben spezifiziert werden. Dieses Kriterium kann durch ein zweites außer Kraft gesetzt werden, nach dem die Prozedur abgebrochen wird, wenn die Standardabweichung der L^2 Bootstrap p-Werte kleiner wird als ein durch den Benutzer vorgegebener Wert. Für die in Tabelle 1 wiedergegebenen Ergebnisse haben wir ein Maximum von 300 Stichproben spezifiziert und das zweite Kriterium auf $sd = .02$ gesetzt. Für die Modelle M2 und M3 wurde dies mit 193 bzw. 214 Stichproben erreicht.

Tabelle 1: Deskriptive sowie inferenzstatistische Maße der Anpassungsgüte

Modell	BIC	D	df	L^2	p	RC	p	X^2	p	Bootstrap			
										Stichproben	p(L^2)	p(RC)	p(X^2)
M2	782.7	.132	23	42.44	.008	36.39	.038	36.09	.040	193	.057	.166	.218
M3	792.2	.094	19	31.81	.033	27.29	.098	27.19	.100	214	.051	.126	.182
M4	809.3	.050	12	13.81	.312	12.84	.381	14.18	.289	300	.263	.297	.313
M6	776.6	.074	20	21.24	.383	22.38	.320	26.42	.152	300	.557	.507	.433

BIC = $-2 (\text{Log-Likelihood}) + q \log (N)$, wobei q = Anzahl geschätzter Parameter

$$D = \text{index of dissimilarity} = \sum_{y=1}^Y \left| f_y - \hat{F}_y \right| / 2 N$$

Tabelle 1 enthält deskriptive Indizes (BIC und D), die Teststatistiken L^2 , RC und X^2 mit bei gegebenen df assoziierten p-Werten nach der theoretischen Chi-Quadrat Verteilung sowie die Bootstrap p-Werte.

Nach den p-Werten aufgrund der theoretischen Chi-Quadrat Verteilung wird Modell M2 bei einem kritischen Wert von $\alpha = .05$ übereinstimmend von allen drei Statistiken abgelehnt. Modell M3 würde nach L^2 abgelehnt, nach RC und X^2 jedoch akzeptiert. Für die Modelle M4 und M6 führen alle Statistiken übereinstimmend zur Akzeptanz.

Was sagen die Bootstrap p-Werte? Insgesamt scheint der Modellfit leicht bis erheblich besser zu sein, mit Ausnahme von p(L^2) und p(RC) für Modell M4. Es zeigt sich sogar, daß allein aufgrund der Tests alle Modelle akzeptiert werden könnten. Allerdings liegt p(L^2) für

die Modelle M2 und M3 an der Grenze der Akzeptanz (die Standardabweichungen der p-Werte für diese beiden Modelle betragen .017 und .015). Es zeigt sich auch, daß die p-Werte für die Modelle M2 und M3 erheblich stärker differieren als die für die Modelle M4 und M6. Solange die Frage nicht geklärt ist, welche dieser drei Statistiken der theoretischen Chi-Quadrat Verteilung am ehesten folgt, plädieren wir dafür, einem Modell den Vorzug zu geben, bei dem alle Bootstrap p-Werte möglichst dicht beieinander liegen. Das sind in diesem Fall die Modelle M4 und M6, wobei die Präferenz für Modell M6 spricht. Dies ist nicht nur ein sparsames Modell, sondern es paßt nach den Bootstrap p-Werten auch am besten für die Daten.

Wir möchten auch einen Befund erwähnen, der nicht nur typisch für diese Daten ist. Für Modell M2 ist $L^2 > X^2$, für Modell M6 dagegen $L^2 < X^2$. Für die vorliegenden Daten bleibt die Relation für die non-Bootstrap p-Werte bei den Bootstrap p-Werten erhalten ($p(L^2) < p(X^2)$ für Modell M2 und $p(L^2) > p(X^2)$ für Modell M6). Es kann aber auch Fälle geben, bei denen sich diese Relation drastisch umkehrt (vgl. *Langeheine* et al. 1995).

Schließlich zu den deskriptiven Indizes. Forscher, die BIC favorisieren, würden sich eindeutig für Modell M6 entscheiden (je kleiner BIC, desto besser). Man beachte allerdings, daß die Werte für alle Modelle nicht weit auseinander liegen. BIC hat wie der index of dissimilarity D den Vorteil, daß sich auch nicht geschachtelte Modelle vergleichen lassen. Da bislang allerdings nichts über die Verteilungen dieser beiden Indizes bekannt ist, läßt sich nicht sagen, ob ein BIC (D) wirklich besser ist als ein anderes. D kann interpretiert werden als der Anteil der Fälle der erwarteten Häufigkeiten, die reklassifiziert werden müßten, um die beobachteten Häufigkeiten perfekt zu reproduzieren. D-Werte nahe Null zeigen also eine gute Anpassung, aber die Obergrenze von D ist kleiner als 1 und variiert mit dem in Frage stehenden Modell. Im vorliegenden Fall spricht D am ehesten für die Modelle M4 und M6.

Führt man Simulationen durch, so stellt sich die Frage, wie groß die Zahl der Stichproben sein sollte, damit man den Ergebnissen vertrauen kann. Sicherlich sind 19 Stichproben (*Hope* 1968; *Aitkin* et al. 1981) nicht ausreichend. Aber selbst eine so kleine Zahl sagt mehr als gar nichts. *Van der Heijden, 't Hart* und *Dessens* (1994) haben zur Entscheidung über die Anzahl der Klassen in einer LCA über antisoziales Verhalten (19 dichotome Variablen, N = 2918 Jugendliche) 50 Stichproben mittels parametrischem Bootstrap verwendet. Sie sehen diese Zahl als bei weitem zu klein an, um eine gute Schätzung der Monte Carlo Verteilung für L^2 zu erhalten. *Bollen* und *Stine* (1992) haben für ihre beiden Beispiele 500 (artifizielle Daten) und 250 (reale Daten) Replikationen benutzt. *Collins* et al. (1993) geben 200 Stichproben als wahrscheinlich mehr als ausreichend an. Wie bereits erwähnt, hat *Hope* (1968) gezeigt, daß die Power des Tests mit zunehmender Anzahl der Stichproben steigt.

Natürlich hängt der notwendige Rechenaufwand von der Größe der Daten (Anzahl der Zellen der Tabelle) und der Größe des Modells (Anzahl der zu schätzenden Parameter) ab.

Man beachte, daß latent class Modelle iterativ geschätzt werden. Das erfordert u.U. eine große Zahl von Iterationen nicht nur für die Schätzung eines Modells für die Daten, sondern auch für jede Bootstrap Stichprobe. Über das Verhältnis von Aufwand und Ertrag muß daher von Fall zu Fall entschieden werden.

Für die vorliegenden Daten und Modelle können auch mehrere tausend Stichproben mit einem schnellen PC (486-66 oder Pentium) in angemessener Zeit verarbeitet werden. In Tabelle 2 geben wir daher die Bootstrap p-Werte für L^2 , RC und X^2 (zur besseren Übersichtlichkeit auf zwei Stellen hinter dem Komma gerundet) für die vier Modelle bei zunehmender Zahl von Stichproben. Der Trend ist klar: für die vorliegenden Daten stabilisieren sich die Werte zwischen 1000 und 2000 Stichproben, aber die Schlußfolgerungen unterscheiden sich nicht wesentlich von denen bei 300 Stichproben.

Tabelle 2: Bootstrap p-Werte bei zunehmender Zahl von Stichproben
(300, 1000, 2000, 3000, 4000)

Modell	p(L ²)					p(RC)					p(X ²)				
	300	1000	2000	3000	4000	300	1000	2000	3000	4000	300	1000	2000	3000	4000
M2*	.06	.05	.05	.04	.04	.17	.13	.14	.13	.14	.22	.19	.19	.19	.20
M3*	.05	.09	.07	.07	.07	.13	.20	.18	.19	.19	.18	.27	.24	.25	.24
M4	.26	.27	.29	.29	.29	.30	.37	.37	.37	.37	.31	.37	.36	.36	.37
M6	.56	.59	.58	.57	.58	.51	.51	.52	.52	.52	.43	.45	.44	.44	.44

* Für die Modelle M2 und M3 wurden anstelle von 300 193 bzw. 214 Stichproben gezogen

5. Diskussion

Da wir den Großteil der diskussionswürdigen Punkte in die vorangegangenen Abschnitte integriert haben, verbleiben nur wenige Punkte, die angesprochen werden sollen.

Die Debatte zwischen *Jagodzinski* (1986, 1987a,b) und *Langeheine* (1987a,b) entzündete sich an der Frage, ob Chi-Quadrat basierte Teststatistiken zur Beurteilung der Modellanpassung für die durch sparseness gekennzeichnete 3×3×3 Tabelle des Postmaterialismus Panels sinnvoll sind. Mittels parametrischem Bootstrap haben wir eine Antwort auf diese und nur diese Frage zu geben versucht (d.h., die restlichen Punkte der Kontroverse bleiben dabei

außer Acht). Unsere Ergebnisse bestätigen die Schlußfolgerung von *Langeheine* (1987a), daß den Modellen M4 bzw. M6 der Vorzug zu geben ist.

Statistischer Modell-Fit ist allerdings nicht das einzige entscheidende Kriterium. Modelle sollen auch plausibel sein. Man kann sich z.B. die Frage stellen, ob es plausibel ist, Antwortunsicherheit nur am ersten Zeitpunkt anzunehmen und später gar nicht mehr (wie in Modell M3). Zudem ist zu bedenken: Kein Modell ist wahr; jedes Modell ist eine Simplifikation der Wirklichkeit. Es erscheint uns aber sinnvoll, das Modell zu wählen, das die Wirklichkeit aufgrund statistischer Tests am besten zu beschreiben scheint. Von den vier Modellen ist Modell M2 nicht plausibel, Modell M6 scheint eine gute Darstellung der Wirklichkeit zu geben. Aber auch die Modelle M3 und M4 erlauben einen interessanten Blick auf die Daten.

Wir möchten auch nicht unerwähnt lassen, daß sich das Aggregationsniveau der Daten nicht ohne Einfluß auf die Ergebnisse erwies (vgl. *Langeheine* 1987b; *Jagodzinski* 1987b). Die desaggregierten Daten bestanden aus einer 12^3 -Tabelle, die zu einer 3^3 -Tabelle bzw. einer 6^3 -Tabelle aggregiert wurden. Da sparseness in den größeren Tabellen (6^3 , 12^3) in noch viel extremerem Ausmaß gegeben ist, wäre es interessant zu sehen, wie die Bootstrap Analyse verschiedener Modelle für diese Tabellen ausfallen würde. Über diese Daten verfügen wir leider nicht mehr.

Der Befund, daß die Bootstrap p-Werte im vorliegenden Fall für die Modelle M2 und M3 zwischen den drei Statistiken L^2 , RC und X^2 differieren, mag Pessimisten zu der Folgerung verleiten, daß das Bootstrap uns auch nicht aus dem Sumpf befreien kann (das Bootstrap geht zurück auf die Abenteuer des Baron *von Münchhausen*, der sich vor dem Ertrinken rettete, indem er sich an seinen eigenen Haaren aus einem Sumpf zog - Bootstrapper benutzen dazu ihre Schnürsenkel). "Formal" könnte man dieses Dilemma lösen, indem man lediglich eine dieser Statistiken simuliert, und zwar diejenige, die für das Problem am adäquatesten erscheint (*Hope* 1968). Das allerdings ist eine schwierige Frage, und deshalb empfiehlt *Hope*, mehr als eine Statistik zu benutzen. Intuitiv würde man erwarten, daß alle drei Bootstrap p-Werte zu derselben Schlußfolgerung führen, selbst wenn die Verteilungen der Statistiken gegeneinander verschoben sind. Es scheint allerdings so zu sein, daß es derartige Unterschiede geben kann, wenn die Asymptotik nicht gilt. L^2 , RC und X^2 sind unterschiedliche Statistiken, die dann nicht dieselbe Verteilung haben. Daher plädieren wir dafür, ein Modell zu verwerfen, wenn eines der drei Bootstrap p's dies nahelegt. Unsere Erfahrungen aus einer Reihe von Datensätzen zeigen, daß es sich bei solchen Fällen in der Regel um restriktive Modelle handelt. Verschiedene dieser im Ausmaß von sparseness variierende Datensätze und verschiedene Modelle werden in *Langeheine* et al. (1995) mittels Bootstrap reevaluiert.

Schließlich möchten wir anmerken, daß das hier präsensierte Bootstrap zur Evaluation der Anpassungsgüte eines bestimmten Modells benutzt wurde. *Bollen* und *Stine* (1992) zeigen

für Kovarianz Struktur Modelle, wie man geschachtelte Modelle über die Differenzverteilung der entsprechenden Bootstrap Werte gegeneinander testen kann. Im Prinzip ließe sich dies auch für kategoriale Daten durchführen. Es wären dann beide Modelle für jede Stichprobe anzupassen, wobei die Stichproben aus den geschätzten Proportionen des restriktiveren Modells zu ziehen wären. Auf diese Weise ließe sich die oft gehörte Versicherung, daß asymptotische Tests für Vergleiche von geschachtelten Modellen auch im Fall von sparseness gültig sind (z.B. *Agresti* und *Wang* 1987; *Haberman* 1978), mit den Ergebnissen aus Bootstrap Tests vergleichen.

Literatur

- Agresti, A.* and *Wang, M.C.* 1987:
An empirical investigation of some effects of sparseness in contingency tables, in: *Computational Statistics & Data Analysis* 5: 9-21.
- Aitkin, M., Anderson, D* and *Hinde, J.* 1981:
Statistical modelling of data on teaching styles, in: *Journal of the Royal Statistical Society, Series A* 144: 419-461.
- Akaike, H.* 1974:
A new look at the statistical model identification, in: *IEEE Transactions on Automatic Control* 19: 716-723.
- Andersen, E.B.* 1982:
Latent trait models and ability parameter estimation, in: *Applied Psychological Measurement* 6: 445-461.
- Andersen, E.B.* 1985:
Estimating latent correlations between repeated testings, in: *Psychometrika* 50: 3-16.
- Andersen, E.B.* 1988:
Comparison of latent structure models, in: R. Langeheine and J. Rost (eds.), *Latent trait and latent class models*, New York: Plenum: 207-229.
- Barnard, G.A.* 1963:
Contribution to discussion, in: *Journal of the Royal Statistical Society, Series B* 25: 294.
- Bollen, K.A.* and *Stine, R.A.* 1992:
Bootstrapping goodness-of-fit measures in structural equation models, in: *Sociological Methods & Research* 21: 205-229.
- Bozdogan, H.* 1987:
Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions, in: *Psychometrika* 52: 345-370.
- Clogg, C.C.* 1995:
Latent class models: Recent developments and prospects for the future, in: *G. Arminger, C.C. Clogg and M.E. Sobel* (eds.), *Handbook of statistical modeling in the social sciences*, New York: Plenum: 311-359.
- Collins, L.M., Fidler, P.F., Wugalter, S.E.* and *Long, J.D.* 1993:
Goodness-of-fit testing for latent class models, in: *Multivariate Behavioral Research* 28: 375-389.
- Converse, P.E.* 1964:
The nature of belief systems in mass publics, in: *D.E. Apter* (ed.), *Ideology and discontent*, New York: The Free Press: 206-261.
- Efron, B.* 1982:
The jackknife, the bootstrap and other resampling plans. CBMS-NSF regional conference series in applied mathematics (Society for Industrial and Applied Mathematics, Philadelphia, PA).
- Efron, B.* and *Tibshirani, R.* 1993:
An introduction to the bootstrap, New York: Chapman & Hall.
- Everitt, B.* 1988:
A Monte Carlo investigation of the likelihood ratio test for number of classes in latent class analysis, in: *Multivariate Behavioral Research* 23: 531-538.

- Goodman, L.A.** 1974:
The analysis of systems of qualitative variables when some of the variables are unobservable. Part I - A modified latent structure approach, in: *American Journal of Sociology* 79: 1179-1259.
- Haberman, S.J.** 1978:
Analysis of qualitative data. Vol. 1: Introductory topics, New York: Academic Press.
- Holt, J.A. and Macready, G.B.** 1989:
A simulation study of the difference chi-square statistic for comparing latent class models under violation of regularity conditions, in: *Applied Psychological Measurement* 13: 221-231.
- Hope, A.C.A.** 1968:
A simplified Monte Carlo significance test procedure, in: *Journal of the Royal Statistical Society, Series B* 30: 582-598.
- Jagodzinski, W.** 1986:
Black & White statt LISREL? Wie groß ist der Anteil von "Zufallsantworten" beim Postmaterialismusindex? in: *ZA-Information* 19: 30-51.
- Jagodzinski, W.** 1987a:
Über einige Anwendungs- und Interpretationsprobleme "anspruchsvoller" Schätzverfahren (Entgegnung auf den Beitrag von Langeheine), in: *ZA-Information* 20: 56-63.
- Jagodzinski, W.** 1987b:
Entgegnung₂ zu Beitrag₂ von Langeheine, in: *ZA-Information* 21: 77-81.
- Langeheine, R.** 1987a:
Black & White, anfängliche Antwortunsicherheit, Mover-Stayer, Third Force oder was? Ein paar weitere Überlegungen zu Jagodzinski's Analyse des Postmaterialismus Panels, in: *ZA-Information* 20: 44-55.
- Langeheine, R.** 1987b:
Die zweite Diskussionswelle über die Auswertungsprobleme eines 3-Wellen Panels kategorialer Daten: Einige Anmerkungen zur Entgegnung von Jagodzinski, in: *ZA-Information* 21: 70-76.
- Langeheine, R.** 1988:
New developments in latent class theory, in: **R. Langeheine and J. Rost** (eds.), *Latent class and latent trait models*, New York: Plenum: 77-108.
- Langeheine, R. und Rost, J.** 1993:
Latent Class Analyse, in: *Psychologische Beiträge* 35: 177-198.
- Langeheine, R., Stern, E. and van de Pol, F.** 1994:
State mastery learning: Dynamic models for longitudinal data. *Applied Psychological Measurement* 18: 277-291.
- Langeheine, R., Pannekoek, J. and van de Pol, F.** 1995:
Bootstrapping goodness-of-fit measures in categorical data analysis. Manuscript.
- Read, T.R.C. and Cressie, N.A.C.** 1988:
Goodness-of-fit statistics for discrete multivariate data, New York: Springer.
- Rost, J.** 1988:
Test theory with qualitative and quantitative latent variables, in: **R. Langeheine and J. Rost** (eds.), *Latent trait and latent class models*, New York: Plenum: 147-171.
- Schwarz, G.** 1978:
Estimating the dimension of a model, in: *Annals of Statistics* 6: 461-464.
- Shockey, J.W.** 1988:
Latent class analysis: An introduction to discrete data models with unobserved variables, in: **J.S. Long** (ed.), *Common problems/proper solutions: Avoiding error in quantitative research*, Beverly Hills: Sage: 288-315.
- Van de Pol, F., Langeheine, R. and de Jong, W.** 1991:
PANMARK user manual: PANel analysis using MARKov chains. Voorburg: Netherlands Central Bureau of Statistics.
- Van der Heijden, P., 't Hart, H. and Dessens, J.** 1994:
A parametric bootstrap procedure to perform tests in a LCA of anti-social behaviour, erscheint in: **J. Rost and R. Langeheine** (eds.), *Applications of latent trait and latent class models in the social sciences*.