Wie viele Fälle werden gebraucht? Ein Monte-Carlo-Verfahren zur Bestimmung ausreichender Stichprobengrößen und Teststärken(power) bei Strukturgleichungsanalysen mit kategorialen Indikatorvariablen

Urban, Dieter; Mayerl, Jochen

Wie viele Fälle werden gebraucht? Ein Monte-Carlo-Verfahren zur Bestimmung ausreichender Stichprobengrößen und Teststärken (power) bei Strukturgleichungsanalysen mit kategorialen Indikatorvariablen

von Dieter Urban und Jochen Mayerl [1]

Zusammenfassung

Zur Festlegung ausreichender Fallzahlen für eine robuste Schätzung von Strukturgleichungsmodellen (SEM-Analyse) gibt es verschiedene Daumenregeln. Diese basieren jedoch auf Ergebnissen von Simulationsstudien, die in der Regel mit anwendungsfremden Modellspezifikationen durchgeführt wurden, und sie berücksichtigen stets auch nur wenige Daten- und Modell-Spezifika. Insbesondere sind sie nicht für SEM-Analysen mit kategorialen Indikatorvariablen und WLS-Methode geeignet. Als Alternative wird hier ein Verfahren aufgezeigt, das von ***Muthén/Muthén*** *in verschiedenen Veröffentlichungen vorgeschlagen wurde, und in dem in vier Schritten mittels Monte-Carlo-Simulation mehrere Kriterien zur Festlegung einer ausreichenden Fallzahl für ein spezielles Modell zu überprüfen sind (der Grad der Verzerrung bei der Schätzung von Effektparametern und Standardfehlern, der Grad der Abdeckung (coverage), die Teststärke (power) für Signifikanztests einzelner Effektparameter). Das ursprünglich für die SEM-Analyse mit kontinuierlichen Variablen gedachte Verfahren wird hier auf die speziellen Bedingungen einer kategorialen SEM-Analyse bezogen. Die dazu notwendigen EDV-Steuerfiles werden in der Syntax der SEM-Software „Mplus“ vorgestellt.*

---

1 Dr. ***Dieter Urban*** ist Professor, ***Jochen Mayerl*** (MA) ist Wissenschaftlicher Mitarbeiter am Institut für Sozialwissenschaften, Abteilung für Soziologie I, Universität Stuttgart, 70174 Stuttgart.

Wir danken ***Linda K. Muthén*** und ***Bengt O. Muthén*** für Hinweise und Informationen, die sie uns u.a. über das Internet-Diskussionsforum „Mplus Discussion“ (www.statmodel.com/discussion) zur Verfügung stellten.

Abstract

How Many Cases Are Needed? Using Monte-Carlo Simulation to Determine Sufficient Sample Size and Power When Analyzing Structural Equation Models with Categorical Outcome Variables.

There are several rules of thumb trying to determine a sufficient number of cases for a robust estimation of a structural equation model (SEM analysis). All of these rules generalize from results of simulation studies relying on model and data specifications that could be far away from those of a particular research project. Specifically, these rules do not consider the conditions of SEM analyses with categorical output variables and WLS estimation. In this paper an alternative procedure that was proposed by ***Muthén/Muthén*** *in several publications is shown. Using the results of Monte-Carlo simulations, this procedure assesses sample size by several criteria to find out a sufficient number of cases for SEM estimation in a particular research project. Criteria are: parameter and standard error biases, coverage and power for single parameter tests. It is shown how to apply the procedure originally developed for SEM analysis with continuous outcome variables to the analysis of structural models with categorical indicators. In addition, all input files needed to conduct the four step procedure are documented in the syntax of the SEM software package Mplus.*

1 Einleitung

Wie viele Fälle werden benötigt, um für ein bestimmtes Strukturgleichungsmodell ein stabiles oder zumindest halbwegs robustes Schätzergebnis zu erzielen? Welche Fallzahlen werden mindestens gebraucht, um den Resultaten einer Strukturgleichungsmodellierung vertrauen zu können? Ein jeder Anwender dieser Analysetechnik (im Folgenden auch „SEM-Analyse“ genannt) kennt das Problem. Reichen die vorhandenen 100, 200 oder 1000 Beobachtungsfälle aus, um für die freien Parameter eines Strukturgleichungsmodells, welches theoretisch interessiert, das aber auch zumeist in sehr anspruchsvoller Weise und mit einer hohen (viel zu hohen?) Komplexität spezifiziert wurde, mehr als nur numerische Zufallstreffer im Rahmen eines bestimmten statistischen Schätzverfahrens (ML, GLS, WLS/ADF, o.a.) geliefert zu bekommen?

Die einschlägige Methodenliteratur hilft dabei in aller Regel nicht viel weiter. In ihr ist zunächst einmal zu lesen, dass alle gängigen Schätzverfahren zur Bestimmung der freien Parameter in Strukturgleichungsmodellen auf einer asymptotischen Schätztheorie beruhen, nach der bei einer möglichst großen Fallzahl (und bei Einhaltung bestimmter Annahmen) vertrauenswürdige Schätzwerte zu erhalten sind.[2] Wann eine solche, möglichst große Fallzahl gegeben ist, sagt die statistische Schätztheorie leider jedoch nicht. Das verunsichert den Anwender natürlich noch mehr, und er greift umso dankbarer nach bestimmten Daumenregeln, die ebenfalls in der Literatur angeboten werden (wie z.B. nach der N:q-Hypothese, mehr dazu im nächsten Abschnitt), um sein Entscheidungsproblem überhaupt „in den Griff" zu bekommen. Leider sind diese Daumenregeln jedoch stets umstritten (auch in der Literatur) und basieren zumeist auf Simulationsstudien mit spezifischen Modellspezifikationen, die ganz anders als diejenigen des leidgeplagten Anwenders aussehen können, so dass sie häufig auch nicht zu übertragen sind. So wurden und werden z.B. als Richtlinien bei Analysen mit kontinuierlichen Variablen häufig die Simulationsstudien von ***Boomsma*** (1982, 1983) und die neuere Studie von ***Hoogland*** (1999) herangezogen.

Noch um ein Vielfaches größer ist das Leid desjenigen Anwenders, der so vermessen ist, die Werte seiner sozialwissenschaftlichen Einstellungs- und Verhaltensindikatoren nicht als Messungen kontinuierlicher Variablen anzunehmen bzw. zu definieren (was sie in aller Regel auch nicht sind, was aber fast alle SEM-Praktiker, die Autoren der vorliegenden Studie eingeschlossen, als – unter bestimmten Umständen – verzeihliche Sünde betrachten und deshalb tun). Alle Anwender, die ihre Indikatorvariablen als kategoriale Variablen (geordnet oder ungeordnet) betrachten und auch als solche in die SEM-Analyse einbeziehen wollen, erfahren in der Literatur und in einschlägigen Internet-Diskussionsforen (SEMNET, Mplus-Discussion, o.Ä.) höchst Verwirrendes über die dafür benötigten Mindest-Fallzahlen. Wenn sie sich z.B. der LISREL-Strategie verschrieben haben, und deren Analyseweg über die Berechnung polychorischer bzw. polyserieller Korrelationskoeffizienten und WLS-Schätzung gehen wollen, wird ihnen nahe gelegt, mindestens 2000 oder sogar 5000 Fälle in die Schätzung einzubeziehen, da ansonsten insbesondere die Schätzwerte für die Standardfehler extrem verzerrte Werte annehmen können.[3] Wenn sie aber die Mplus-Strategie bevorzugen und die dortige Implementation der WLS-Schätzung für Modelle mit kategorialen Indikatorvariablen akzeptieren,[4] werden

2 So ist z.B. zu lesen: „Because the GLS estimates are only asymptotically correct, large samples are required for the estimates to be trustworthy." (***Xie*** 1989: 340).

3 ***Olsson*** et al. (2000) empfehlen 2000 bis 3000 Fälle. ***Yuan/Bentler*** (1994) empfehlen sogar 2000 bis 5000 Fälle pro Gruppe von Variablen-Werten, da die Anzahl der Elemente in der weight matrix mit einem Exponenten von 4 für die Anzahl der Variablen anwachse.

4 Als „Mplus-Strategie" wird hier die Schätzmethode bezeichnet, die im EDV-Programmpaket Mplus implementiert ist (vgl. www.statmodel.com). Vgl. dazu ***Muthén*** 1983, 1984, 1993; ***Muthén/Satorra*** 1995; ***Xie*** 1989; ***Muthén/Muthén*** 2001: 339-343 (Appendix 1: Regression

bereits brauchbare Schätzungen mit nur 150 Fällen (bei geringer Anzahl freier Parameter und symmetrischer Werteverteilung) als machbar erklärt.[5]

Einen Ausweg aus dieser Misere scheinen Monte-Carlo-Simulationen zu ermöglichen, in denen für ein spezielles, zu testendes Strukturgleichungsmodell (und nicht für ein beliebiges bzw. nach methodischen Kriterien spezifiziertes Modell!) bestimmte Parameterwerte als richtig angenommen werden und sodann große Mengen von Zufallsstichproben für jeweils unterschiedliche Fallzahlen gezogen werden, um diese Werte zu replizieren. Über das Ausmaß der dabei erhaltenen Abweichungen zwischen „wahren“ und geschätzten Werten lässt sich dann ermitteln, welche Fallzahlen als akzeptabel und welche als nicht ausreichend gelten können.

Linda K. Muthén und ***Bengt O. Muthén*** haben unlängst in einem viel beachteten Vorschlag diese Methode zur Festlegung einer sinnvollen Stichprobengröße für die Schätzung von Strukturgleichungsmodellen mit latenten Variablen vorgestellt (***Muthén***/***Muthén*** 2002a, 2002b; ***Muthén*** 2002). Allerdings eignet sich der von ihnen präsentierte Verfahrensweg nur für die Bestimmung von Mindest-Fallzahlen, wenn alle Indikatorvariablen als kontinuierliche Variablen in eine ML-Schätzung einbezogen werden.[6] Im Folgenden wollen wir deshalb die von ***Muthén***/***Muthén*** vorgeschlagene Methodik benutzen, um einen Verfahrensweg aufzuzeigen, mit dem sinnvolle Mindest-Fallzahlen auch für Modelle mit kategorialen Indikatorvariablen, die mit einer WLS-Methode geschätzt werden müssen, zu ermitteln sind.[7]

Um diesen Weg für alle SEM-Anwender so leicht wie möglich gangbar zu machen, werden wir die dazu notwendigen vier Analyseschritte anhand eines Beispiels veranschaulichen (Abschnitt 4) und auch die dabei benutzten vier EDV-Inputfiles in der Syntax des SEM-Programmpakets „Mplus“[8] im Anhang dokumentieren. Zuvor werden wir noch typische Praxistipps und Daumenregeln zur Festlegung von

with a categorical dependent variable), 345-352 (Appendix 2: The general modeling framework).

5 Vgl. die Mitteilung von ***Linda K. Muthén*** im Internet-Diskussionsforum „Mplus-Discussion“ (www.statmodel.com/discussion) vom 23.5.2001.

6 Dies liegt u.a. daran, dass sie zur Bestimmung ausreichender Fallzahlen eine Teststärke-Analyse (power analysis) nach dem ***Satorra-Saris*** Ansatz vornehmen (in: ***Muthén/Muthén*** 2002b), der nur unter Bedingungen multivariater Normalverteilung gültig ist (vgl. dazu ***Saris/Satorra*** 1993; ***Satorra/Saris*** 1985; ***Saris/Stronkhorst*** 1984; ***Muthén/Curran*** 1997).

7 Für Analysen mit geordnet-kategorialen Variablen, deren Werte mindestens fünf Ausprägungen und keine allzu starken Abweichungen vom Idealbild einer Normalverteilung aufweisen, kann auch alternativ zur WLS-Methode die von ***Satorra***/***Bentler*** (1986) vorgeschlagene robuste Variante der ML-Schätzung eingesetzt werden.

8 Vgl. ***Muthén/Muthén*** 2001 (www.statmodel.com).

Mindest-Fallzahlen für SEM-Analysen vorstellen (Abschnitt 2) sowie die von uns benutzten Kriterien zur Bestimmung des Zusammenhangs zwischen Stichprobengröße und SEM-Analysequalität erläutern (Abschnitt 3).

2 SEM-Schätzung und Stichprobenumfang

Im Allgemeinen ist es allen SEM-Anwendern bekannt: die Ergebnisse von SEM-Analysen können durch zu kleine Stichprobengrößen in unerwünschter Weise beeinflusst werden. So tendiert die χ^2-Teststatistik zum Beispiel bei einem niedrigen Stichprobenumfang zu überhöhten Werten und die H_0 wird zu oft verworfen (vgl. ***Bollen*** 1989: 268). Ebenso kann eine Verzerrung der Parameterschätzungen und insbesondere der geschätzten Standardfehler auftreten, wovon die Schätzung der Konfidenzintervalle beeinflusst wird: unterschätzte Standardfehler können zur Überschätzung und überschätzte Standardfehler zur Unterschätzung der Signifikanz von Effekten führen (vgl. ***Muthén***/***Muthén*** 2002: 600). Des Weiteren nimmt die Breite des Konfidenzintervalls mit abnehmender Fallzahl zu. Eine zu kleine Stichprobe kann folglich zu einer deutlichen Verschlechterung der Qualität der Modellschätzung führen, wodurch eine zuverlässige substantielle Interpretation beeinträchtigt wird (vgl. ***Algina***/***Olejnik*** 2000).

In systematischer Argumentation wird die Qualität einer SEM-Analyse in ganz entscheidendem Maße von der Robustheit einer SEM-Schätzung bestimmt. Die Robustheit einer SEM-Schätzung betrifft wiederum:

a) die Robustheit aller geschätzten Effektparameter
b) die Robustheit der Standardfehler aller geschätzten Effektparameter
c) die Robustheit der χ^2-Teststatistik und aller darauf basierenden Fit-Indizes

Und alle drei Robustheiten werden durch die jeweils zur Verfügung stehende Fallzahl positiv beeinflusst: je höher die Fallzahl, umso höher ist die Robustheit der SEM-Schätzung. Allerdings wirkt die Fallzahl nicht in direkter Weise auf die Robustheit, sondern in Abhängigkeit von vielen anderen Merkmalen des jeweils analysierten Modells und der ausgewerteten Daten sowie in Abhängigkeit vom jeweils benutzten Schätzverfahren (ML-, GLS-, WLS/ADF-Schätzung). Zu den fallzahlrelevanten Modell- und Datenmerkmalen gehören insbesondere:[9]

- die Gesamtzahl der Indikatoren im Modell (k)
- die durchschnittliche Anzahl von Indikatoren pro Faktor (k/f)

---

9 Diese Auflistung und die folgenden Praxishinweise folgen den Analysen von ***Hoogland*** 1999 und ***Marsh/Hau*** 1999.

- die durchschnittliche Faktorladung (L) in allen Messmodellen
- die durchschnittliche Schiefe aller Werte-Verteilungen (S)
- die Standardabweichung der Schiefe (sd(s))
- die durchschnittliche Kurtosis der Verteilungen aller gemessenen Variablen (K)

Hinsichtlich des Zusammenhangs zwischen der Robustheit von SEM-Schätzungen (im Sinne der drei oben genannten Bezüge) und der zur Verfügung stehenden Fallzahlen wurden verschiedene Regelmäßigkeiten beobachtet und Praxisempfehlungen bzw. Daumenregeln abgeleitet, die hier vor allem hinsichtlich ihrer praktischen Konsequenzen für die SEM-Forschung vorgestellt werden sollen:

(ad a) Im Vergleich zu GLS- und WLS/ADF-Schätzung[10] ist die ML-Schätzung am wenigsten von der Kombination „geringe Fallzahl und hohe Kurtosis" betroffen. Generell sind ML-Schätzungen stabiler und von größerer Präzision, was ihren Fit angeht (***Olsson*** et al. 2000). Wenn ein Modell mindestens drei Indikatoren pro Faktor aufweist, und die durchschnittliche Faktorladung über 0.5 liegt, dann gelten bei $N \geq 200$ die ML-Parameterschätzwerte als unverzerrt. Dabei ist der Quotient (k/f) die wichtigste Determinante. So wurde z.B. für die ML-Schätzung eines Strukturmodells mit $k=12$, $k/f=4$ und $K=3$ eine Mindest-Fallzahl von $N=200$ empfohlen (***Hoogland*** 1999).

In verschiedenen Simulationen (mit ML-Schätzung) konnte gezeigt werden, dass die Nachteile kleiner Stichproben durch hohe k/f-Werte kompensiert werden können (vor allem, aber nicht nur, wenn die Faktorladungen höher als 0.6 sind).[11] Für viele ML-Schätzungen sind demnach bei k/f von 6 bis 12 auch Fallzahlen von 50 durchaus ausreichend.

Es konnte auch gezeigt werden, dass bei hoher Reliabilität (d.h. bei geringer Fehlervarianz) bereits Fallzahlen von 50 bis 100 Einheiten ausreichen können, um ML-Schätzwerte mit nur geringen Verzerrungen zu erhalten. Voraussetzung dafür sind jedoch normalverteilte Indikatoren (vgl. ***Hoyle***/***Kenny*** 1999).

Die Ergebnisse der GLS-Schätzung sind auch schon bei kleinen Fallzahlen akzeptabel, setzen aber Modelle mit nur sehr kleinem Spezifikationsfehler voraus (***Olsson***

10 WLS: „Weighted-least-squares"-Schätzverfahren; ADF: „Asymptotic-distribution-free"-Schätzverfahren. Das WLS-Verfahren ist theoretisch identisch mit dem ADF-Verfahren. Beide sind nur unterschiedliche Umsetzungen des gleichen Schätzverfahrens in verschiedenen Software-Paketen.

11 Vgl. ***Marsh/Hau*** 1999.

et al. 2000). Eine genügend große Robustheit sei bei $N \geq 50 * k$ gegeben, weil für die Robustheit der GLS-Schätzung die Gesamtheit der zur Verfügung stehenden Indikatoren (k) die wichtigste Determinante zu sein scheint. Im oben genannten Beispiel (mit $k=12$, $k/f=4$, $K=3$), für das sich bei einer ML-Schätzung eine Mindest-Fallzahl von 200 empfiehlt, ergibt sich somit für eine GLS-Schätzung eine empfehlenswerte Fallzahl von $N=600$.

Die WLS/ADF-Schätzung braucht große bis sehr große Fallzahlen und eine Modellspezifikation, die nur sehr gering fehlerhaft ist (***Olsson*** et al. 2000). Eine genügend große Fallzahl könnte sich aus der Gesamtzahl aller Indikatoren (k) in Abhängigkeit von der durchschnittlichen Kurtosis aller Werteverteilungen (K) ergeben, denn es besteht ein starker Effekt von $K/\sqrt{N}$ auf die Verzerrung der WLS/ADF-Schätzung. Dies bedeutet für Empfehlungen (nach ***Hoogland*** 1999) hinsichtlich der Mindest-Fallzahl bei WLS/ADF-Schätzungen:

$N \geq 50 * k$, wenn K zwischen -1.0 und 0.0
$N \geq 100 * k$, wenn K zwischen 0.0 und 3.3
$N \geq 250 * k$, wenn K zwischen 3.3 und 6.0

Daraus würde als Empfehlung für eine sinnvolle WLS/ADF-Fallzahl bei einem Modell mit $k=12$, $k/f=4$, $K=3$ (s.o.) eine Stichprobengröße von $N=1200$ folgen.

(ad b) Für den Zusammenhang von Stichprobengröße und Robustheit der Standardfehler können aus den Ergebnissen der Simulationen von ***Hoogland*** (1999) folgende Praxisempfehlungen abgeleitet werden:

Hinsichtlich der Robustheit von ML- und GLS-Schätzung sind L und K am wichtigsten. Sie sollten möglichst bei 0.00 liegen. Das hieße:

wenn $L \geq 0.7$ und $K=0$ wird $N=200$ empfohlen,
wenn $L \geq 0.5$ und $K=0$ wird $N=400$ empfohlen,
wenn $K \geq 2.0$ sollte ML-robust benutzt werden.

Demgegenüber verzerrt die WLS/ADF-Schätzung die Standardfehler so sehr, dass sie nur bei sehr großen Fallzahlen eingesetzt werden sollte. Als Daumenregel zur Maximierung der Robustheit von Standardfehlern bei WLS/ADF-Schätzungen könnte nach ***Hoogland*** (1999) gelten:

$N \geq 10k(k+1)$ bei $K=0$ (z.B. bei $k=12$: $N=1560$)
$N \geq 15k(k+1)$ bei $K \leq 5.70$ (z.B. bei $k=12$: $N=2340$)

(ad c) Auch der Zusammenhang zwischen der Robustheit der χ^2-Teststatistik (sowie aller darauf basierender Fit-Indizes) und einer genügend großen Fallzahl ist von vielen Determinanten abhängig. Generell könnte gelten:[12] Insbesondere bei angemessen spezifizierten Modellen mit normalverteilten Variablen zeigen χ^2-Tests bei ML-Schätzungen mit unterschiedlichsten Fallzahlen (N = 100, 200, 500 oder 1000) keine verzerrten Ergebnisse. Hingegen sind die χ^2-Testergebnisse bei WLS/ADF-Schätzungen nur dann nicht verzerrt, wenn die Fallzahlen sehr groß sind (ab N = 1000). Nach den Ergebnissen von ***Hoogland*** (1999) könnten folgende Daumenregeln sinnvoll sein:

für ML-Schätzungen bei alpha=0.05:

$N \geq \max(3Ldf\,(1 + 10S - 10sd(s) + K), 100)$

Beispiel: N = 826 für df = 51, L = 0.6, S = 1.0, sd(s) = 0.5, K = 3.0

für GLS-Schätzungen bei alpha = 0.05:

$N \geq \max(2(1-L)\,df\,(1 + 10S - 10sd(s) + K), 100)$

Beispiel: N = 367 für df = 51, L = 0.6, S = 1.0, sd(s) = 0.5, K = 3.0

für WLS/ADF-Schätzungen bei alpha = 0.05:

$N \geq 45 * df$

Beispiel: N = 2295 für df = 51

Die oben genannten, minimalen Fallzahlen (200-400) zur Erreichung robuster ML-Schätzwerte werden auch von Monte-Carlo-Simulationen bestätigt, in denen experimentell möglichst viele der zuvor genannten Einflussfaktoren konstant gehalten werden und nur die Fallzahl variiert wird. ***Jackson*** (2001, 2003) konnte z.B. in verschiedenen Simulationen zeigen, dass vieles für eine sinnvolle Stichprobengröße in ML-Schätzungen bei Fallzahlen ab einer Größenordnung von 200 bis 400 Einheiten spricht. So hatte in einem „wahren“ Modell der GFI (ein direkter Fit-Index) bei einer Fallzahl von 200 einen Wert von 0.92, der dann bis zu einer Fallzahl von 400 auf 0.96 anstieg, aber sich bei einer Erhöhung auf eine Fallzahl von 800 nur noch um 2,6 % vergrößerte, während er sich bei der Erhöhung von 50 auf 400 Fälle noch um 29 % vergrößerte (***Jackson*** 2001).

Generell betrachtet ist somit die notwendige Fallzahl für SEM-Analysen von sehr vielen Determinanten abhängig. Da hilft auch die oftmals benutzte Universal-Daumenregel[13] von 5, 10 oder 20 Beobachtungen für jeden zu schätzenden Parame-

---

12 Vgl. dazu ***Curran*** et al. 1996.

13 Diese Daumenregel wird hier als „universell“ bezeichnet, weil sie beansprucht, unabhängig vom jeweils eingesetzten Schätzverfahren sowie unabhängig von den jeweils empirisch gegebenen Datenverteilungen und den forschungspraktisch begründeten Modellspezifikationen gültig zu sein.

ter[14] nicht viel weiter, die zudem wohl auch falsch ist. Zumindest konnte ***Jackson*** (2001, 2003) in verschiedenen Monte-Carlo-Simulationen nicht eindeutig nachweisen, dass der NPPP-Wert (number of participants per parameter) bzw. das N:q-Verhältnis (Anzahl von Beobachtungen, N, pro zu schätzendem Parameter, q) einen direkten Einfluss auf die Varianz der geschätzten Koeffizienten und die Höhe der Fit-Indizes hat. Deutliche statistische Hinweise gibt es demnach allein dafür, dass das N:q-Verhältnis einen negativen Effekt auf zwei von sieben Fit-Indizes hat (nämlich auf den „chi-square bias“ und den RMSEA-Index): je mehr Beobachtungen pro Parameter vorliegen, umso kleiner wird der Chi-Square Bias[15] und der RMSEA-Index. Allerdings ist das Ausmaß des nachgewiesenen Effektes nicht dramatisch und oftmals auch ohne praktische Relevanz. Zum Beispiel reduzierte sich der RMSEA-Index von 0.005 auf 0.002, wenn bei N = 800 das N:q-Verhältnis durch zusätzliche Modell-Constraints von 20:1 (800 Fälle und 40 Parameter) auf 400:1 (800 Fälle und 2 Parameter) vergrößert wurde. So hat wohl die absolute Fallzahl eine größere Bedeutung für die Erhöhung der Modellanpassung (Fit) und die Vermeidung von Parameterverzerrungen (s.o.) als das Verhältnis von Fallzahl zu Parameterzahl.[16] Zudem haben ***Marsh*/*Hau*** (1999) nachgewiesen, dass auch, wenn die Fallzahl kleiner als die Anzahl der zu schätzenden Parameter ist, stabile Schätzungen erreicht werden können.

3 Kriterien zur Bestimmung des Stichprobenumfangs

Die oben genannten Beispiele für Praxistipps/Daumenregeln zur Festlegung eines genügend großen Stichprobenumfangs bei einer SEM-Schätzung haben gezeigt, dass das Problem einer genügend großen Fallzahl nicht durch Anwendung einer einzigen, einheitlichen Generalformel zu lösen ist: „In reality, there is no rule of thumb that applies to all situations“ (***Muthén/Muthén*** 2002a: 599). Wie viele Fälle insgesamt benötigt werden, scheint in erster Linie vom spezifizierten Modell und den zur Verfügung stehenden Daten abzuhängen. Denn es ist unstrittig, dass für gute Schätzungen von komplexen Modellen mehr Fälle benötigt werden als für einfache Modelle.[17] Und auch die Verwendung von kategorialen Indikatoren vergrößert die benötigten Fallzahlen erheblich.

---

14 Vgl. zu dieser Universal-Daumenregel: ***Bollen*** 1989, ***Kline*** 1998, ***Tanaka*** 1987.

15 Wird berechnet als: (χ^2-Wert - df) / (df) (***Jackson*** 2003: 133).

16 Zu diesem Resümee kommt auch ***Jackson*** 2001, 2003.

17 Nach ***Stone/Sobel*** (1990) verdoppelt sich bei einer ML-Schätzung sofort die Mindest-Stichprobengröße von ca. 200 auf 400 Fälle, wenn im Modell eine latente statt einer manifesten Mediatorvariablen enthalten ist.

Wie viele Fälle insgesamt benötigt werden, hängt aber auch davon ab, in welcher Hinsicht die Qualität einer SEM-Analyse durch Festlegung einer genügend großen Fallzahl optimiert werden soll bzw. in welcher Hinsicht durch eine genügend große Fallzahl ein bestimmter SEM-Qualitätsstandard zu gewährleisten ist. Denn eine bestimmte Fallzahl „may be large enough for unbiased parameter estimates, unbiased standard errors, and good coverage, but it may not be large enough to detect an important effect in the model" (***Muthén/Muthén*** 2002a: 600).

Im Folgenden wollen wir deshalb das von ***Muthén/Muthén*** (2002a) vorgeschlagene Vorgehen zur Ermittlung einer ausreichenden Anzahl von Beobachtungsfällen benutzen, um die besonders heikle Frage nach der Zahl notwendiger Fälle für eine SEM-Analyse mit kategorialen Indikatorvariablen zu beantworten. Dieses Verfahren nutzt zum einen eine Monte-Carlo-Simulation am empirisch zu analysierenden und nicht an einem abstrakt-methodisch bestimmten Strukturgleichungsmodell, und es liefert zum anderen Informationen zu insgesamt fünf verschiedenen Kriterien, mit denen die Relevanz verschiedener Stichprobenumfänge für den Erfolg bzw. die Qualität einer SEM-Analyse zu beurteilen ist. Diese fünf Kriterien sind:[18]

1. Der Grad der Verzerrung (bias), mit dem jeder Effektparameter im Modell geschätzt wird. Dieser sollte nicht größer als 10% sein. Er wird hier ermittelt, indem die Modellschätzung in Form einer Monte-Carlo-Simulation mit den Populationswerten als Startwerten sehr häufig wiederholt wird (in unserem Beispiel insgesamt 5000-mal). Die Abweichung des Durchschnittswertes aller Replikationen vom „wahren" Populationswert sollte dann nicht größer als 10% der Größe des Populationswertes selbst sein.[19]
2. Der Grad der Verzerrung (bias), mit dem jeder Standardfehler im Modell geschätzt wird. Dafür gilt der gleiche Grenzwert und die gleiche Methodik wie für den Grad der Verzerrung aller Effektparameter (s.o.).
3. Der Grad der Verzerrung (bias) für den Standardfehler eines oder mehrerer besonders wichtiger Effektparameter, für den/die hier auch die Teststärke (s.u.) bestimmt werden soll. Er sollte nicht oberhalb von 5 % liegen.
4. Der Grad der Abdeckung (coverage) für alle Effektparameter sollte oberhalb von 0.90 liegen. Mit dem Abdeckungsgrad wird hier der Anteil der Replikationen der Monte-Carlo-Simulation bezeichnet, deren 95%-Konfidenzintervall den „wahren" Parameter der Population enthalten.

---

18 Vgl. dazu ***Muthén/Muthén*** 2002a: 605f.

19 Um dies festzustellen, muss nur der Durchschnittswert vom Populationswert subtrahiert werden und dieser Differenzwert durch den Populationswert dividiert werden. Der sich daraus ergebende Kennwert sollte nicht größer als $|0.010|$ sein.

5. Die Teststärke (power) (s.u.), die notwendig ist, um einen wichtigen Effektparameter zu identifizieren (vgl. 3.). Diese ergibt sich hier aus dem Anteil von Replikationen der Monte-Carlo-Simulation, für welche die falsche Null-Hypothese, dass ein Parameter gleich null ist, mit einer Irrtumswahrscheinlichkeit von 5% zurückgewiesen wird. Sie sollte entsprechend eines Vorschlags von ***Cohen*** (1988) und den Konventionen der gängigen Forschungspraxis nahe einem Anteil von 80% bzw. einem Wert von 0.80 liegen.

Da es in sozialwissenschaftlichen Forschungen noch immer eher unüblich ist, sich mit der Teststärke von Hypothesentests zu beschäftigen, sollen dazu an dieser Stelle noch einige Erläuterungen gegeben werden:

Die übliche Praxis bei Anwendung von Signifikanztests zielt darauf ab, eine Null-Hypothese (H_0), an deren Richtigkeit der Forscher nicht glaubt, mit einer möglichst geringen Irrtumswahrscheinlichkeit (α) zu verwerfen. Wenn z.B. die Irrtumswahrscheinlichkeit mit 0.021 unterhalb des allgemein akzeptierten Grenzwertes von 0.05 liegt, wird dies in der Forschungspraxis als Hinweis darauf gewertet, dass die H_0 falsch ist, folglich verworfen werden muss, und deshalb die Alternativ-Hypothese (H_A), an deren Berechtigung der Forscher glaubt, richtig sein könnte. Freilich ist eine solche Entscheidung stets mit einem Fehler, dem so genannten „Typ I-Fehler“, versehen: die H_0 könnte auch irrtümlich, also fälschlicherweise (obwohl sie richtig ist) verworfen worden sein. Zur Vorbeugung gegenüber einem solchen falschen Entscheid sollte dann auch der Typ I-Fehler möglichst gering sein und im Regelfall unterhalb von 5% liegen.

Eine solche Testpraxis entspricht einer konservativen Entscheidungsstrategie. Um eine neue Hypothese als möglicherweise richtige Hypothese überhaupt akzeptieren zu können, muss möglichst viel gegen die Null-Hypothese sprechen, die behauptet, dass an der neuen Alternativ-Hypothese überhaupt nichts dran ist. Jedoch könnten die Wälle, die zum Schutz vor übereilten Schlüssen errichtet wurden, auch zu hoch sein, um einer neuen Hypothese überhaupt eine Chance zu geben. Dadurch könnte eine falsche H_0 irrtümlicherweise viel zu lange Bestand haben und daran gehindert werden, durch eine neue, innovative Hypothese ersetzt zu werden. Die Wahrscheinlichkeit eines solchen Irrtums wird also umso größer, je höher die Schutzmauern um H_0 gezogen werden, oder je kleiner der Typ I-Fehler gesetzt wird. Diese Wahrscheinlichkeit, dass die H_0, obwohl sie falsch ist, irrtümlicherweise beibehalten wird, wird als Typ II-Fehler bezeichnet. Die Wahrscheinlichkeiten der beiden Fehler verlaufen also gegensinnig: Wird die Wahrscheinlichkeit (α) des Typ I-Fehlers abgesenkt, steigt im Gegenzug die Wahrscheinlichkeit (β) des Typ II-Fehlers und umgekehrt. Beide sollten natürlich möglichst klein sein, aber beide können nicht gleich klein sein. Deshalb ist ein Kompromiss zu suchen, und der liegt üblicherweise bei einem $\alpha \leq 0.05$ und einem $\beta \leq 0.20$.

Als Teststärke (power) wird die Differenz von 1-β bezeichnet. Sie sollte mithin ≥ 0.80 sein. Die Teststärke kann als die Wahrscheinlichkeit definiert werden, eine falsche Null-Hypothese richtigerweise zu verwerfen, was im Kontext von SEM-Analysen die Möglichkeit zur Entdeckung von schwerwiegenden Spezifikationsfehlern bedeutet. Und damit wird auch der Zusammenhang zwischen Teststärke und Stichprobenumfang deutlich: Da sich Typ I- und Typ II-Fehler gegenläufig verändern, sinkt die Teststärke bei gleicher Stichprobengröße, wenn die Irrtumswahrscheinlichkeit α herabgesetzt wird. Somit muss die Fallzahl angehoben werden, um eine höhere Teststärke bei einem festgesetzten α zu erhalten, da eine Vergrößerung des Stichprobenumfangs zu einer Verkleinerung des Standardfehlers[20] und damit zu einer Erhöhung der Teststärke führt.[21] Eine zu geringe Teststärke aufgrund einer zu kleinen Fallzahl kann demnach auch dazu führen, dass eine Alternativ-Hypothese keine Chance bekommt, die H_0 zu ersetzen. Denn wenn mit der „konservativen Teststrategie" (s.o.) eine nicht ausreichende Test-Signifikanz (α) substantiell als Zurückweisung der alternativen H_A-Hypothese interpretiert wird, kann das Scheitern der H_A auch daran liegen, dass im Test- bzw. Untersuchungsdesign keine ausreichende Teststärke gegeben ist.

Im Kontext von SEM-Analysen ist die Teststärke-Bestimmung schwieriger als bei einfachen statistischen Modellen (z.B. bei t-Tests oder Korrelationsanalysen), in denen Alternativ-Hypothesen immer nur zu wenigen Parametern zu formulieren sind. Im Prinzip kann in der SEM-Analyse jeder fixierte Parameter (auch jeder via Modellannahme unausgesprochen fixierte Parameter) falsch sein und eine sehr große Anzahl von alternativen Werten annehmen. Zudem hängt die Teststärke auch nicht nur von der Stichprobengröße ab. Weitere Bestimmungsgrößen der Teststärke in der SEM-Analyse sind: die Größe des α-Fehlers (s.o.), die Anzahl fehlender Werte (***Muthén/Muthén*** 2002a), Spezifikationsfehler im Analysemodell (***Kaplan*** 1997, 2000), Positionierungen von Variablen innerhalb des Modells (***Kaplan*** 1997), Werteverteilungen von Variablen (***Muthén/Muthén*** 2002a), die Reliabilität der Indikatorvariablen (***Hoyle/Kenny*** 1999), die Anzahl von Indikatoren pro Faktor (***Bollen*** 1989: 348) und die Effektstärken der Variablenbeziehungen (***Bollen*** 1989: 343).

---

20 Der Standardfehler des Regressionskoeffizienten lässt sich berechnen als: $SE_b = \sqrt{var(b)} / \sqrt{N}$.

21 Mit einer Verkleinerung des Standardfehlers (SE_b) verkleinert sich im Hypothesentest (Annahme: H_0 ist richtig) das Konfidenzintervall ($KIB = b \pm (k * SE_b)$), damit steigt der Typ I-Fehler, was wiederum ein Abfallen des Typ II-Fehlers (β) zur Folge hat, und ein kleinerer Typ II-Fehler führt zu einer größeren Teststärke (power = 1- β).

Für die Teststärke-Analyse/Poweranalyse im Rahmen von Strukturgleichungsmodellierungen liegen verschiedene Verfahrensvorschläge vor, die dazu in aller Regel eine Monte-Carlo-Simulation mit ML-Schätzverfahren für Modelle mit kontinuierlichen Indikatorvariablen nutzen (z.B. ***MacCallum*** et al. 1996, ***Muthén/Muthén*** 2002b, ***Satorra/Saris*** 1985). Nachfolgend soll jedoch erstmals in systematischer Weise gezeigt werden, in welcher Weise das Simulationsverfahren auch eingesetzt werden kann, um den notwendigen Stichprobenumfang für Strukturgleichungsmodellierungen mit kategorialen Indikatoren zu ermitteln.

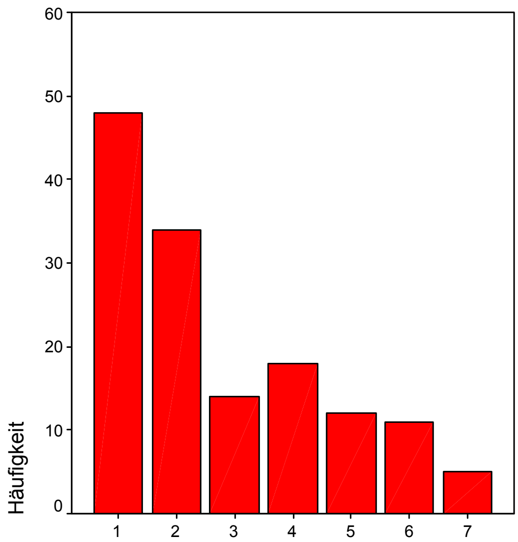
4 Die vier Schritte der Monte-Carlo-Simulation, am Beispiel verdeutlicht

Das zu analysierende Modell, mit dem beispielhaft die Schritte zur Bestimmung von Stichprobengröße und Teststärke ausgeführt werden sollen, stammt aus der Xenophobie-Forschung und soll hier nur in aller Kürze vorgestellt werden.[22] In diesem Modell (vgl. dazu die Abbildung 2) übt das latente Konstrukt „Autoritarismus" (F2/AU) einen kausalen Effekt auf die ausländerablehnende Einstellung (F1/AA) aus. Die beiden exogenen Variablen „Bildung" (x1) und „subjektive Schichteinstufung" (x2) üben sowohl einen direkten Effekt als auch jeweils einen indirekten Effekt (über den Mediator „Autoritarismus") auf die Ausländerablehnung aus. Die sieben Indikatoren y1 - y7 enthalten Messwerte auf siebenstufigen Ratingskalen, die in mündlicher Befragung von 142 Personen erhoben wurden. Die univariaten Verteilungen dieser ordinalen (kategorial geordneten) Indikatoren sind teilweise extrem schief (vgl. Abbildung 1), so dass sich eine Analyse des Modells mit einem Schätzverfahren für kontinuierlich-normalverteilte Variablen strikt verbietet. Stattdessen soll das Modell mit einem WLS-Schätzverfahren nach der Mplus-Strategie[23] berechnet werden, welche auch Monte-Carlo-Simulationen unter kategorialen Modellbedingungen zulässt.

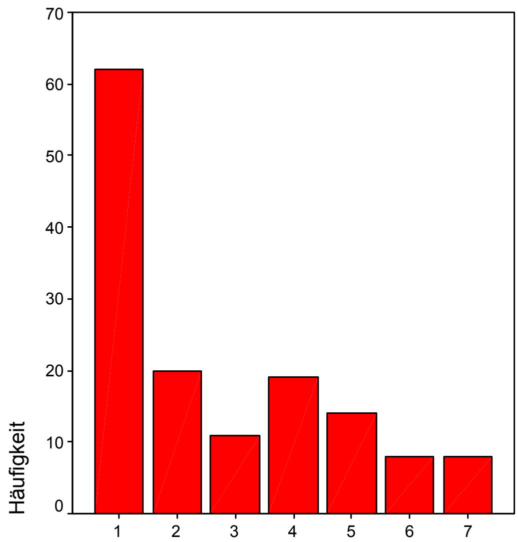
22 Für weitere theoretische und substantielle Informationen zu diesem Modell vgl. ***Urban/Mayerl*** (2003).

23 Vgl. dazu die Literaturhinweise in Fußnote 4.

Abbildung 1 Univariate Werteverteilungen von zwei y-Indikatoren

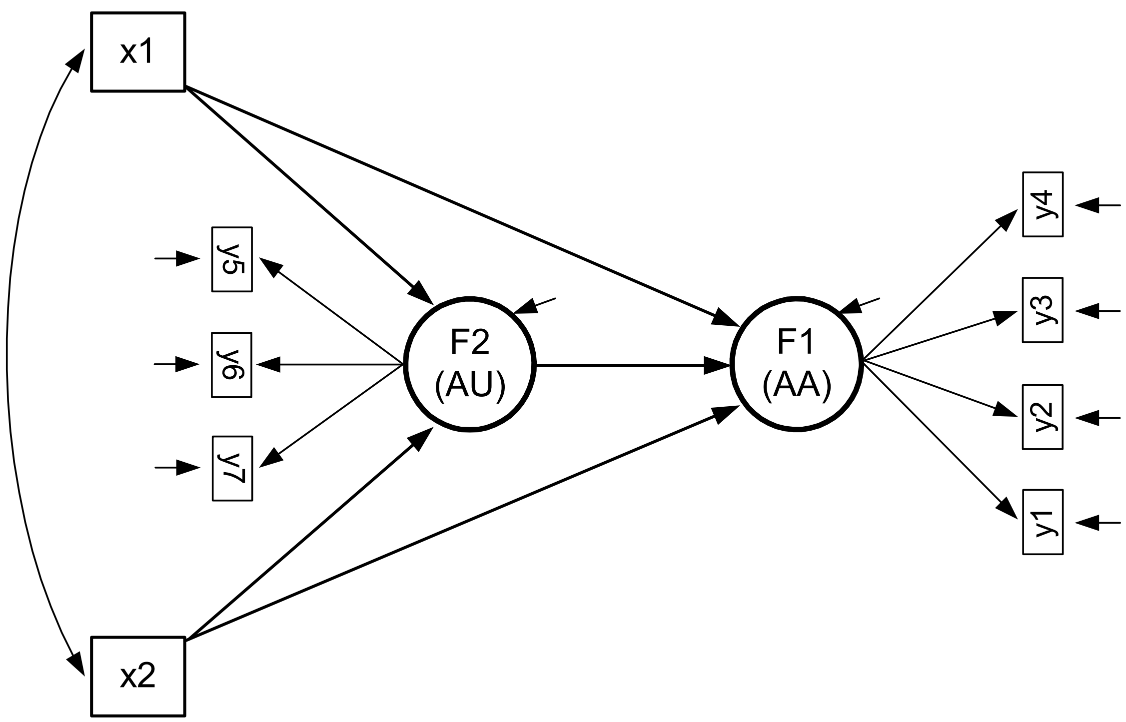


y2: „bei knapper Arbeit Jobs nur für Deutsche“
1: lehne völlig ab; 7: stimme voll und ganz zu



y5: „dankbar für führende Köpfe“
1: lehne völlig ab; 7: stimme voll und ganz zu

Abbildung 2 Beispielmodell zum Zusammenhang zwischen Autoritarismus und Ausländerablehnung



Der nachfolgende Vorschlag zur Bestimmung der Stichprobengröße für Strukturgleichungsmodellierungen mit kategorialen Indikatoren besteht aus insgesamt vier Schritten. Die Monte-Carlo-Methode wird dabei in den Schritten 2, 3 und 4 eingesetzt. Im Einzelnen betreffen die vier Schritte folgende Aufgaben:

Schritt 1: Modellschätzung zur Bestimmung der Populationswerte

Schritt 2: Erstellung einer Populationskovarianzmatrix

Schritt 3: Evaluation der Modellschätzung mittels Analyse von Parameter-, Standardfehlerverzerrungen und Abdeckungen (coverage)

Schritt 4: Bestimmung der Teststärken

Die modell-notwendige Stichprobengröße ergibt sich aus den Schritten 3 und 4 und ist erreicht, sobald durch stufenweise Anhebung der Stichprobengröße in der Monte-Carlo-Simulation für alle der in Abschnitt 3 genannten Kriterien zufrieden stellende Ergebnisse vorliegen.

Im Folgenden werden die vier notwendigen Analyseschritte anhand des oben vorgestellten Beispielmodells erläutert. Die programmtechnische Realisation eines jeden Analyseschritts wird in der Mplus-Syntax im Anhang gezeigt.

4.1 Erster Schritt: Modellschätzung

Ausgangspunkt einer jeden Monte-Carlo-Simulation ist eine hypothetische Bestimmung der Populationswerte aller Parameter des zu analysierenden Strukturmodells: „These values can be obtained from theory or previous research" (***Muthén***/***Muthén*** 2002a: 601). Im vorliegenden Beispiel (Abb. 2) sollen die WLS-Schätzwerte des Strukturmodells, die mit den Daten der oben genannten Studie ermittelt wurden, als Populationswerte verwendet werden. Somit dient der erste Schritt des Verfahrens der Bestimmung der Populationswerte aller zu schätzenden Parameter und der sich daraus ergebenden Residualvarianzen.

Die im ersten Schritt erzielten WLS-Schätzwerte für alle Effektparameter zeigt Abbildung 3.[24]

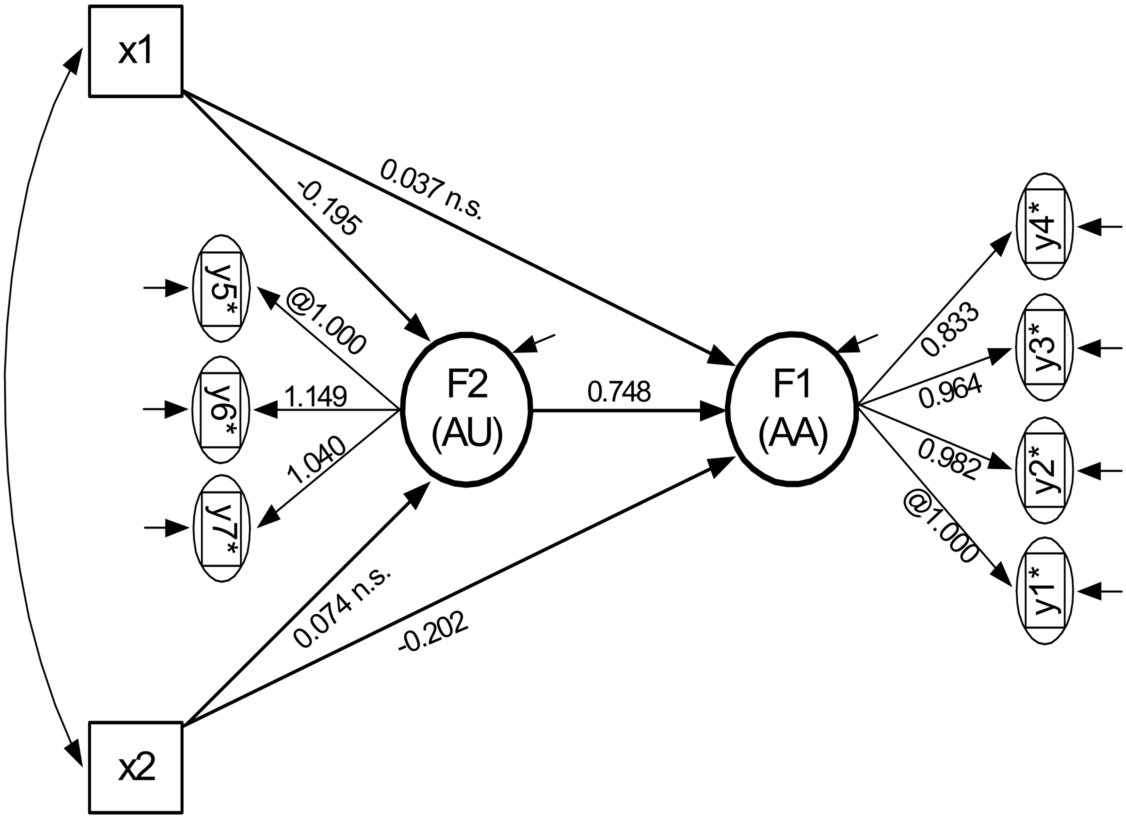
Die Anpassungswerte sind für die Zwecke dieser Arbeit insgesamt zufrieden stellend. Der χ^2-Wert beträgt 45.83 (mit p=0.00 und df=23 gegenüber einem χ^2-Wert

24 Die geschätzten Residualvarianzen der latenten Konstrukte betragen: 0.335 (F1) und 0.572 (F2), diejenigen der Indikatoren betragen: 0.345 (y1*), 0.368 (y2*), 0.391 (y3*), 0.545 (y4*), 0.428 (y5*), 0.244 (y6*) und 0.381 (y7*). Die Thresholds aller sieben kontinuierlich-normalverteilten y*-Indikatoren bewegen sich (mit minimalen Abweichungen) zwischen -1.500 und 1.000.

des Baseline-Modells von 722.03 mit p = 0.00 und df = 35), der CFI beträgt 0.97 und der RMSEA 0.08.[25]

Die in Schritt 1 gewonnenen Populationswerte werden nun in Schritt 2 zur Generierung einer Populationskovarianzmatrix benutzt. Erst dann können in den Schritten 3 und 4 die erwünschten Ergebnisse zum notwendigen Stichprobenumfang des Modells berechnet werden.

Abbildung 3 Beispielmodell mit unstandardisierten Koeffizienten



Erläuterungen:

Koeffizientenwerte, die mit @ markiert sind, wurden in der Schätzung auf 1.00 fixiert. Die y*-Indikatoren werden in der WLS-Schätzstrategie als kontinuierlich-normalverteilte Hintergrund-Indikatoren ermittelt (vgl. dazu die in Fußnote 4 angegebene Literatur). Da es sich bei den y*-Indikatoren um latente Indikatoren handelt, deren empirische Werte nur als kategorial skalierte Ausprägungen von y-Variablen empirisch ermittelt werden konnten, sind sie hier als umkreiste Rechtecke grafisch dargestellt.

25 Bessere Anpassungswerte für dasselbe Modell können mit dem robusten WLS-Schätzer WLSMV erzielt werden, der bislang nur in Mplus implementiert ist.

4.2 Zweiter Schritt: Erzeugung der Populationskovarianzmatrix

Die im zweiten Verfahrensschritt zu generierende Populationskovarianzmatrix wird für die Monte-Carlo-Simulationen der Schritte 3 und 4 benötigt. Sie bezieht sich auf die kontinuierlich verteilten y*-Indikatoren, die zwar in der empirischen Erhebung nur als kategoriale y-Indikatoren beobachtet werden konnten, deren kontinuierliche Verteilungen aber aufgrund der Schätzung von Threshold-Werten im ersten Schritt unseres Verfahrensweges[26] zu ermitteln sind.

Die Kovarianzelemente werden mit Hilfe der in Schritt 1 bestimmten, hypothetischen Parameterwerte für Faktorladungen, Regressionskoeffizienten, Faktor(ko)varianzen und Residualvarianzen gewonnen. Die entsprechende Kovarianzmatrix der Population kann mittels einer ML-Schätzung in einer Monte-Carlo-Simulation mit einer einzigen Replikation und einer möglichst hohen (fiktiv festzusetzenden) Fallzahl generiert werden (in der vorliegenden Studie werden 100000 Fälle verwendet). Für die Schätzung ist allein eine Identitätskovarianzmatrix zu definieren, die die modellbedingte Matrixstruktur festlegt, und eine Fixierung aller Modellparameter auf die in Schritt 1 gewonnenen Parameterwerte vorzunehmen (vgl. dazu die Syntax für Schritt 2 im Anhang).

Die folgende Abbildung 4 zeigt die Identitätskovarianzmatrix und (in der ersten Zeile) den Vektor der arithmetischen Mittelwerte der abhängigen und unabhängigen Modellvariablen (insgesamt neun Variablen) im Format einer üblichen Datensatz-Textdatei.

Abbildung 4 Vektor der arithmetischen Mittel und Identitätskovarianzmatrix (hier dargestellt als Datenfile „identity.dat"; vgl. Syntax im Anhang)

```
0 0 0 0 0 0 0 0 0
1
0 1
0 0 1
0 0 0 1
0 0 0 0 1
0 0 0 0 0 1
0 0 0 0 0 0 1
0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 1
```

26 Vgl. die Werte in Fußnote 24. Zur Schätzung von kontinuierlichen y*-Verteilungen aufgrund von kategorialen y-Informationen vgl. ***Bollen*** 1989: 439-442.

Die gesuchte Kovarianzmatrix der Population kann der Residualstatistik des Monte-Carlo-Outputs entnommen werden. Sie wird in Abbildung 5 vorgestellt (die erste Zeile der Abbildung enthält den Vektor der arithmetischen Mittel aller y*-Indikatoren). Diese Populationswerte werden zur Datengenerierung in den Monte-Carlo-Simulationen der folgenden Schritte 3 und 4 für die Bestimmung der notwendigen Stichprobengrößen und Teststärken benötigt.

Abbildung 5 Vektor der arithmetischen Mittel und Populationskovarianzmatrix aller y*-Indikatoren (hier dargestellt als Datenfile „pop.dat“; vgl. Syntax im Anhang)

```
0 0 0 0 0 0 0
1.033
0.676 1.032
0.664 0.652 1.031
0.573 0.563 0.553 1.023
0.438 0.430 0.422 0.365 1.043
0.504 0.494 0.485 0.419 0.707 1.057
0.456 0.448 0.439 0.380 0.640 0.735 1.047
-0.109 -0.107 -0.105 -0.091 -0.195 -0.224 -0.203 1.000
-0.147 -0.144 -0.141 -0.122 0.074 0.085 0.077 0.001 1.000
```

4.3 Dritter Schritt: Evaluation der Modellschätzung

Der dritte Verfahrensschritt dient der Evaluation der Modellschätzung hinsichtlich der in Abschnitt 3 vorgestellten Qualitätskriterien: Grad der Parameterverzerrung, Grad der Standardfehlerverzerrung, Grad der Abdeckung (coverage). Mittels Monte-Carlo-Simulation soll untersucht werden, ob der für das Beispielmodell gegebene Stichprobenumfang von N = 142 zur Erfüllung dieser Kriterien ausreicht, oder, falls nicht, mit welcher Stichprobengröße die genannten Kriterien erfüllt werden könnten.

Im Input-File zum dritten Verfahrensschritt (vgl. Anhang) müssen als Erstes die Schwellenwerte (Cutpoints) der kategorialen Variablen definiert werden. In Anlehnung an den Wertebereich der empirischen Thresholds (vgl. Schritt 1) werden im vorliegenden Beispiel für alle sieben der siebenstufigen Indikatoren die sechs Cutpoints -1,5, -1, -0,5, 0, 0,5 und 1 verwendet. Als Datenbasis dient der Simulation die in Schritt 2 gewonnene Kovarianzmatrix plus dem Vektor der arithmetischen Mittel der y*-Indikatoren. Alle Modellparameter werden mit den Populationswerten aus Schritt 1 als Startwerte unter Verwendung des WLS-Verfahrens geschätzt. Dabei wird zunächst eine Monte-Carlo-Simulation mit einer Stichprobengröße von 142 Fällen durchgeführt, die der oben skizzierten empirischen Datengrundlage entspricht. In Tabelle 1a werden die dementsprechenden Ergebnisse vorgestellt.

In den Tabellen 1a und 1b stammen die Werte der Spalte „hyp. Parameterwert“ aus Schritt 1 und benennen die hypothetischen Parameterwerte der Population. In der Spalte „durchschn. Parameter“ werden die durchschnittlichen Parameterschätzwerte über alle Replikationen der Monte-Carlo-Simulation aufgeführt. Wie der Spalte „Parameterverzerrung“ in Tabelle 1a bei 142 Fällen zu entnehmen ist, weist lediglich ein Parameter eine Verzerrung von mehr als 10% auf (fett gedruckt). Ganz anders hingegen verhält es sich bei den Standardfehlerverzerrungen (vorletzte Spalte). Deren Verzerrungsgrad ist bei 142 Fällen durchweg viel zu hoch. Auch die Coverage-Werte sind für die meisten Parameter nicht zufrieden stellend, d.h. der Anteil an Replikationen, die den wahren Parameterwert im 95%-Konfidenzintervall beinhalten, ist bei fast allen Parametern zu gering. Insgesamt betrachtet, reicht also eine Fallzahl von nur 142 Beobachtungen für eine stabile Schätzung des Analysemodells nicht aus.

Weitere Monte-Carlo-Simulationen mit größeren Fallzahlvorgaben zeigen, dass sich erst ab einer Fallzahl von 400 weitestgehend akzeptable Schätzwerte ergeben. Denn dann erfüllen die Parameter- und Standardfehlerverzerrungen sowie die Coverage-Werte für fast alle Parameter die hier angesetzten Kriterien. Einzige Ausnahme ist die Schätzung der Residualvarianz des latenten Konstrukts F2. Diese weist eine leicht erhöhte Standardfehlerverzerrung von 11% auf, was evtl. vernachlässigt werden könnte, wenn nicht auch der zugehörige Coverage-Wert bei lediglich 0.884 liegen würde. Eine erneute Anhebung der Stichprobengröße auf 500 Fälle erbringt jedoch zufrieden stellende Werte für alle Schätzwerte (vgl. Tabelle 1b).

Diesen Ergebnissen zufolge ist für das untersuchte Beispielmodell bei Verwendung von kategorialen Indikatoren eine Stichprobengröße von mindestens 500 Fällen anzustreben. Mit kleineren Abstrichen ist auch eine Fallzahl von 400 noch akzeptabel, allerdings sollte die Fallzahl für unser Beispiel nicht darunter liegen.

Tabelle 1a) Parameterbias, Standardfehlerbias und Coverage bei 142 Fällen

		hyp. Para-meter-wert	durchschn. Parameter	Parame-terverzer-rung (1)	Standard-abwei-chung	durchschn. Standard-fehler	Standard-fehlerver-zerrung (2)	95%-Cover-age
F1 (AA) BY	y1 (Ausl1)	1,000						
	y2 (Ausl2)	0,982	0,9860	0,00	0,0752	0,0574	**-0,24**	**0,874**
	y3 (Ausl3)	0,964	0,9712	0,01	0,0742	0,0576	**-0,22**	**0,885**
	y4 (Ausl5)	0,833	0,8526	0,02	0,0835	0,0619	**-0,26**	**0,845**
F2 (AU) BY	y5 (Auto2)	1,000						
	y6 (Auto3)	1,149	1,1347	-0,01	0,0949	0,0717	**-0,24**	**0,849**
	y7 (Auto4)	1,040	1,0391	0,00	0,0868	0,0674	**-0,22**	**0,888**
F1 ON	F2	0,748	0,7884	0,05	0,0959	0,0709	**-0,26**	**0,827**
	X1 (Bildung)	0,037	0,0455	**0,23**	0,0723	0,0596	**-0,18**	**0,887**
	X2 (Schicht)	-0,202	-0,1989	-0,02	0,0697	0,0587	**-0,16**	**0,900**
F2 ON	X1 (Bildung)	-0,195	-0,1960	0,01	0,0793	0,0694	**-0,12**	0,913
	X2 (Schicht)	0,074	0,0731	-0,01	0,0770	0,0685	**-0,11**	0,920
Residual-varianzen	F1	0,335	0,3156	-0,06	0,0722	0,0535	**-0,26**	**0,815**
	F2	0,572	0,6173	0,08	0,0822	0,0615	**-0,25**	**0,777**

Tabelle 1b) Parameterbias, Standardfehlerbias und Coverage bei 500 Fällen

		hyp. Para-meter-wert	durchschn. Parameter	Parame-terverzer-rung (1)	Standard-abwei-chung	durchschn. Standard-fehler	Standard-fehlerver-zerrung (2)	95%-Cover-age
F1 (AA) BY	y1 (Ausl1)	1,000						
	y2 (Ausl2)	0,982	0,9832	0,00	0,0389	0,0357	-0,08	0,932
	y3 (Ausl3)	0,964	0,9666	0,00	0,0386	0,0358	-0,07	0,932
	y4 (Ausl5)	0,833	08386	0,01	0,0420	0,0386	-0,08	0,928
F2 (AU) BY	y5 (Auto2)	1,000						
	y6 (Auto3)	1,149	1,1447	0,00	0,0488	0,0446	-0,09	0,918
	y7 (Auto4)	1,040	1,0385	0,00	0,0451	0,0414	-0,08	0,928
F1 ON	F2	0,748	0,7594	0,02	0,0480	0,0437	-0,09	0,920
	X1 (Bildung)	0,037	0,0392	0,06	0,0352	0,0333	-0,05	0,935
	X2 (Schicht)	-0,202	-0,2005	-0,01	0,0348	0,0328	-0,06	0,934
F2 ON	X1 (Bildung)	-0,195	-0,1952	0,00	0,0400	0,0377	-0,06	0,938
	X2 (Schicht)	0,074	0,0742	0,00	0,0386	0,0372	-0,04	0,941
Residual-varianzen	F1	0,335	0,3306	-0,01	0,0350	0,0323	-0,08	0,921
	F2	0,572	0,5862	0,02	0,0407	0,0370	-0,09	0,905

(1) berechnet als: (durchschn. Parameterschätzung – hypothetischer Parameterwert)/hypothetischer Parameterwert
(2) berechnet als: (durchschnittliche Standardfehlerschätzung – Standardabweichung)/Standardabweichung

Als weiteres Kriterium zur Festlegung einer ausreichenden Fallzahl wird im folgenden Abschnitt die Teststärke (Power) betrachtet.

4.4 Vierter Schritt: Bestimmung von Teststärken

Im vierten und letzten Verfahrensschritt soll die Teststärke für verschiedene Modellparameter ermittelt werden. Sie wird über eine bewusste Modell-Fehlspezifikation von jeweils einem Effektparameter berechnet.[27] Als „empirischer Powerwert" kann dann der in einer Monte-Carlo-Simulation korrekterweise zurückgewiesene Anteil fehlspezifizierter Modelle bezeichnet werden.

Die Wahl der Parameter, die fehlspezifiziert werden, sollte bei einer SEM-Poweranalyse zunächst auf die kleinsten bedeutsamen Parameter (Mindesteffekte) des Analysemodells fallen.[28] Um dies am Beispiel zu veranschaulichen, wird hier der empirisch signifikante Effekt der Bildung (x1) auf den latenten Mediator „Autoritarismus" (F2/AU) fälschlicherweise auf einen Wert von 0.00 fixiert (in bewusster Abweichung vom ursprünglichen Parameterwert von -0.195).[29] Dies ermöglicht, den im Rahmen einer Monte-Carlo-Simulation ermittelten Anteil von Zurückweisungen des fehlspezifizierten Modells als geschätzten empirischen Powerwert (auf einem bestimmten Signifikanzniveau) zu definieren. Dieser Anteil kann dann auch als Wahrscheinlichkeit für das Zurückweisen einer falschen Null-Hypothese verstanden werden (vgl. Tabelle 2).

Im Beispiel ergibt sich für eine Fallzahl von 142 ein Powerwert von 0.466 (bei einem α von 0.05) sowie (bei einem α von 0.01) ein Powerwert von 0.258 (vgl. Tabelle 2a). Da diese Powerwerte den Grenzwert von 0.80 deutlich unterschreiten, sollte demnach die Fallzahl zur Schätzung unseres Beispielmodells unabhängig vom Ausmaß der Verzerrung aller Schätzwerte (vgl. Schritt 3) deutlich erhöht werden.

Eine schrittweise Anhebung der Fallzahl in mehreren Monte-Carlo-Simulationen zeigt, dass die Teststärke das Kriterium von 0.80 bei einem α von 0.05 erst bei einer Fallzahl von 400 mit einem Powerwert von 0.808 erreicht. Bei $N = 500$ liegt die

27 Eine alternative Möglichkeit zur Durchführung von Poweranalysen im Kontext von SEM-Analysen besteht darin, den Modellfit – und nicht einzelne Effekte – als Bezugseinheit für die Power-Bestimmung zu benutzen (vgl. ***MacCallum***/***Browne***/***Sugawara*** 1996).

28 Die Poweranalyse kann (und sollte) aber auch für jeden anderen theoretisch interessierenden Parameter durchgeführt werden.

29 Bis auf diese Fehlspezifikation bleibt die Syntax des Inputfiles dieselbe wie in Schritt 3 (vgl. Anhang).

Teststärke bereits bei 0.903 (vgl. Tabelle 2b). Für ein α von 0.01 liegt die Teststärke sogar erst bei 550 Fällen mit einem Wert von 0.816 über dem Grenzwert.

Tabelle 2 Monte-Carlo-Resultat zur Bestimmung der Teststärke des Effektes von Bildung (x1) auf Autoritarismus (F2/AU) (vierter Verfahrensschritt)

2a) Mplus-Output bei N=142

CHI-SQUARE P-VALUES	
Expected	Observed
0.990	1.000
0.980	0.999
0.950	0.997
0.900	0.993
0.800	0.981
0.700	0.963
0.500	0.902
0.300	0.807
0.200	0.721
0.100	0.587
0.050	0.466
0.020	0.334
0.010	0.258

2b) Mplus-Output bei N=500

CHI-SQUARE P-VALUES	
Expected	Observed
0.990	1.000
0.980	1.000
0.950	1.000
0.900	1.000
0.800	0.999
0.700	0.999
0.500	0.996
0.300	0.987
0.200	0.977
0.100	0.948
0.050	0.903
0.020	0.820
0.010	0.756

Zu beachten ist allerdings auch, dass nach den von ***Muthén/Muthén*** vorgeschlagenen Kriterien (in Abschnitt 3 beschrieben) die Verzerrung der Standardfehler für den Parameter, dessen Teststärke interessiert, nicht größer als 5% sein sollte. Im vorliegenden Beispiel wird dieser Wert erst bei 600 Fällen erreicht. Hingegen liegt die Standardfehlerverzerrung mit 400 Fällen bei 7% und mit 500 Fällen bei 6% (vgl. Tabelle 1b), was nur noch knapp über dem geforderten Grenzwert ist.

Führt man die vorgestellte Poweranalyse statt für einen Effekt mit schwacher Stärke auch für einen mit hoher Stärke durch, ergeben sich andere Resultate. Dies lässt sich im vorliegenden Analysemodell am Beispiel des kausalen Effektes vom latenten Mediator „Autoritarismus“ (F2/AU) auf das Konstrukt „Ausländerablehnung“ (F1/AA) zeigen (der standardisierte Regressionskoeffizient liegt hier bei 0.709, während er für den schwachen Effekt bei -0.286 lag). Nunmehr ergibt sich schon bei einer Fallzahl von 142 ein Powerwert von 1.00 (α=0.01). Allerdings legt die Verzerrungsfreiheit der Schätzwerte, die in Schritt 3 evaluiert wurde, auch in diesem Falle einen Stichprobenumfang von mindestens 400 Fällen nahe.

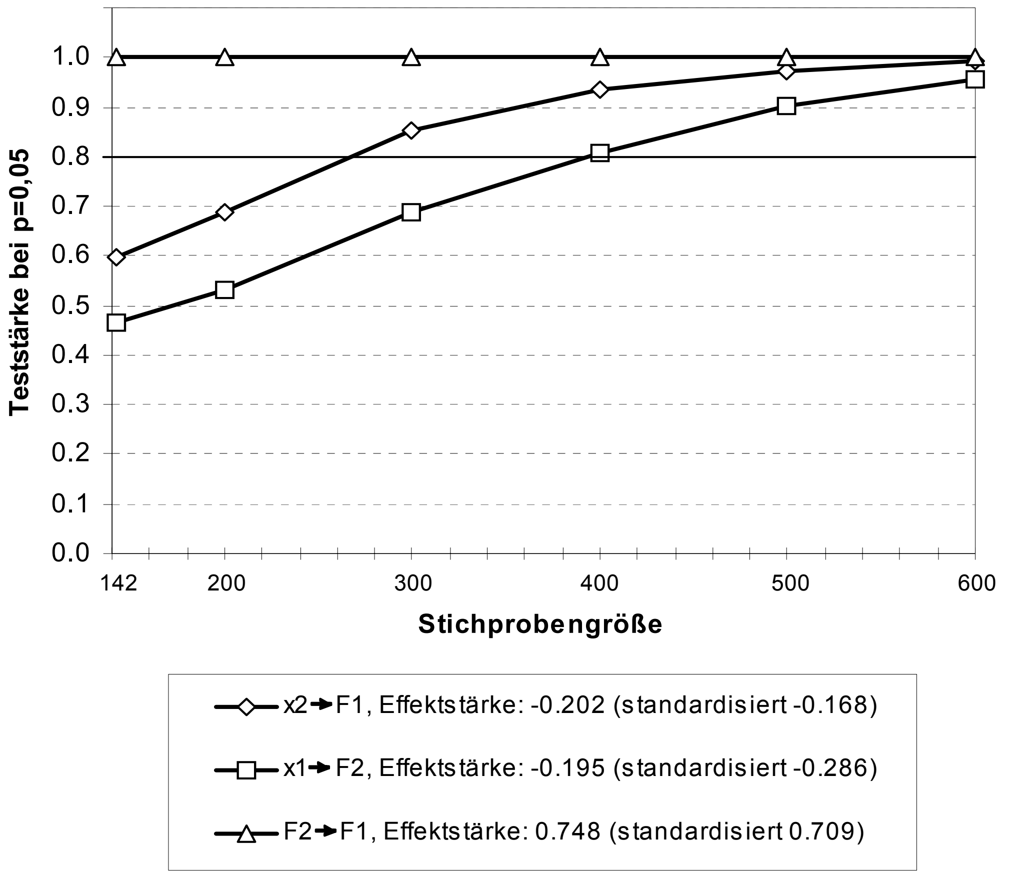
Keineswegs sollte aber davon ausgegangen werden, dass, wie oben geschehen, bei höherer Effektstärke der Powerwert automatisch ansteigt. Dies lässt sich unschwer an folgendem Beispiel erkennen: Oben haben wir für den Effekt von Bildung (x1) auf Autoritarismus (F2/AU) mit einer standardisierten Effektstärke von -0.286 eine Teststärke von 0.466 kalkuliert ($\alpha = 0.05$). Die Teststärke für den im Vergleich dazu schwächeren Effekt von subjektiver Schicht (x2) auf Ausländerablehnung (F1/AA) mit einer standardisierten Effektstärke von -0.168 liegt jedoch mit 0.598 ($\alpha = 0.05$) sehr deutlich über dem Wert des schwächeren Effekts, so dass hier bereits eine Fallzahl von 300 ausreichte, um eine Teststärke von 0.852 zu erreichen. Die Teststärke hängt eben bei Konstanz der Stichprobengröße nicht nur von der Effektstärke, sondern auch noch von vielen anderen Faktoren ab.[30] Folgerichtig sollte die Poweranalyse stets für alle interessierenden Parameter durchgeführt werden, u.a. auch dann, wenn alle Effektstärken als (fast) gleich groß oder schwach geschätzt werden.[31]

Die folgende Abbildung 6 zeigt, in welcher Weise sich die Teststärke (bzw. der Powerwert) für die drei hier diskutierten Effekte verändert, wenn die Stichprobengröße über die von uns beobachteten 142 Fälle anwächst. Deutlich ist zu erkennen, dass die Teststärke des „F2→F1“-Effektes von schwankenden Fallzahlen nicht betroffen ist, und dass die beiden anderen Effekte deutlich mehr als 142 Fälle benötigen, um akzeptable Powerwerte aufzuweisen. Dabei erreicht der relativ schwache „x2→F1“-Effekt schon bei ca. 260 Fällen eine Teststärke oberhalb von 0.80, während der relativ stärkere „x1→F2“-Effekt dies erst ab ca. 400 Fällen schafft.

30 Vgl. dazu die Ausführungen in Abschnitt 3.

31 Die Teststärke/power von Parametern mit gleichen „wahren“ Werten kann in Abhängigkeit von der Platzierung der betreffenden Variablen im Modell variieren, weil sie je nach Lokalität unterschiedliche statistische Beziehungen zu anderen Parametern im Modell haben können (vgl. ***Saris***/***Satorra***/***Sörbom*** 1987).

Abbildung 6 Der Zusammenhang zwischen Teststärke und Stichprobengröße bei drei als unterschiedlich stark geschätzten und unterschiedlich positionierten Modelleffekten



5 Resümee

Kategoriale SEM-Analysen werden von Horrorwarnungen hinsichtlich der dafür erforderlichen Fallzahlen begleitet. Oftmals werden bis zu 5000 Fälle gefordert (s.o.), um mit einer WLS-Schätzung und kategorialen Indikatorvariablen unabhängig von der Modellkomplexität zu halbwegs zuverlässigen und stabilen Parameterschätzwerten gelangen zu können.

Auch werden zur Bestimmung der notwendigen Fallzahl mehr oder weniger gut begründete Daumenregeln vorgeschlagen, die stets aber nur ganz wenige Modell- und Datenstrukturen der jeweiligen SEM-Analyse berücksichtigen und zudem auch noch zu weit streuenden Fallzahlempfehlungen führen. So würden die in Abschnitt 2 vorgestellten Daumenregeln von ***Hoogland*** (1999), die er für den Fall einer WLS-Schätzung aufstellte, bei dem von uns beispielhaft analysierten Modell zu Mindest-

Fallzahlen von 450 (zur Schätzung robuster Effektparameter), 900 (zur Schätzung robuster Standardfehler) oder 1035 (zur Erreichung robuster χ^2-Teststatistiken) führen.[32] Und auch die Universal-Daumenregel „Fälle pro Modellparameter", die den Typ des jeweiligen Schätzverfahrens und die Messqualität der Daten erst gar nicht berücksichtigt, könnte zu äußerst unterschiedlichen Fallzahlforderungen führen, die zwischen 110 und 440 Beobachtungen lägen.[33]

Im Unterschied dazu berücksichtigt das hier für die kategoriale SEM-Analyse vorgestellte Verfahren nach ***Muthén/Muthén*** (2002a) alle Spezifika des jeweils zu analysierenden Modells, alle Spezifika der zur Verfügung stehenden Daten und alle Spezifika der WLS-Schätzmethode. Und es kommt dabei zu recht eindeutigen Empfehlungen hinsichtlich der einzusetzenden Mindestzahl von Beobachtungsfällen anhand von fünf Bewertungskriterien, die von der Vermeidung eines allzu starken Schätzbias bis zur Erreichung einer ausreichenden Teststärke (für den Test einzelner Effektparameter) reichen.

Im Falle unseres Beispielmodells, für das nach den üblichen Daumenregeln, wie oben skizziert, Mindest-Fallzahlen von 110 bis 5000 empfohlen worden wären, berechnet das hier vorgeschlagene Verfahren in methodisch kontrollierter, nachvollziehbarer Weise bei Einsatz einer WLS-Schätzung mit kategorialen Indikatorvariablen einen sinnvollerweise anzustrebenden Stichprobenumfang von mindestens 400, besser noch von 500 Fällen. Und es begründet auch mit statistischen Kenngrößen, warum in diesem Falle einer substantiell interpretierten SEM-Analyse, die allein auf den 142 beobachteten Fällen beruhte, mit großer Skepsis begegnet werden müsste.

Die im Verfahren erzielten Ergebnisse decken sich mit den Resultaten von ***Algina*** und ***Olejnik***, wonach „sample sizes needed for accurate estimation are likely to be substantially larger than sample sizes needed for powerful hypothesis tests" (dies. 2000: 132). Wie gesehen, sollte jedoch hinzugefügt werden, dass es auch von der Effektgröße und dem in der Poweranalyse angesetztem α-Niveau abhängt, ob die Teststärke oder die Akkuratesse höhere Anforderungen an den Stichprobenumfang stellen. Folgerichtig ist es generell ratsam, wie hier gezeigt und wie es der Logik des dargestellten Verfahrens entspricht, sowohl die Akkuratesse der Schätzung als auch die Teststärke unter verschiedenen Bedingungen zur Bestimmung des notwendigen Stichprobenumfangs heranzuziehen und zu simulieren. Denn jedes einzelne

32 N=450 bei (50 * k, wenn K zwischen -1.0 und 0.00) mit k=9, K=-0.45; N=900 bei (10k(k+1), wenn K=0) mit k=9, K=-0.45; N=1035 bei (45(df) mit df=23.

33 N=110 bei 5 Fällen pro Modellparameter (q=22), N=220 bei 10 Fällen pro Modellparameter, N=440 bei 20 Fällen pro Modellparameter.

Kriterium könnte, für sich alleine genommen, falsche Fallzahlempfehlungen abgeben (z.B. könnte eine Poweranalyse geringere Stichprobenumfänge vorschlagen, als zur Erreichung einer ausreichend großen Verzerrungsfreiheit notwendig wären).

Das hier vorgestellte, vierstufige Verfahren kann für eine Apriori- oder eine Posthoc-Analyse eingesetzt werden. Bei der Apriori-Analyse ginge es um den notwendigen Stichprobenumfang einer geplanten SEM-Analyse. Im weniger idealistischen – aber häufig realistischeren – Falle einer Posthoc-Analyse mit gegebener Fallzahl kann das Verfahren zur Beurteilung der Qualität von Parameterschätzwerten und der Teststärke eingesetzt werden, selbst wenn dies "more like an autopsy than a diagnostic procedure" (***Kline*** 1998: 308) erscheinen mag.

Mit welchem Erkenntnisinteresse das Verfahren auch immer eingesetzt wird: Es liefert eine rationale, modell- und datenspezifische Grundlage für die Bewertung von Stichprobenumfängen hinsichtlich der Schätzung von unverzerrten Effektparametern und unverzerrten Standardfehlern sowie hinsichtlich der Möglichkeit, einflussstarke und einflussschwache Effekte im spezifizierten Modell zu entdecken.

Literatur

Algina, J.**/**Olejnik**, **S., 2000: Determining Sample Size for Accurate Estimation of the Squared Multiple Correlation Coefficient. Multivariate Behavioral Research 35: 119-136.

Bollen, K.A., 1989: Structural Equations with Latent Variables. New York et al.: Wiley.

Boomsma, A., 1982: The Robustness of LISREL Against Small Sample Sizes in Factor Analysis Models. S. 149-174 in: ***Jöreskog***, ***K.G.**/**Wold, H***. (Hrsg.): Systems Under Indirect Observation: Causality, Structure, Prediction (Band 1). Amsterdam: North-Holland.

Boomsma, A., 1983: On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality. Amsterdam: Sociometric Research Foundation.

Cohen, J., 1988: Statistical Power Analysis for the Behavioral Sciences. 2nd Edition. London: Erlbaum.

Curran, P.J.**/**West, S.G./Finch, J., 1996: The Robustness of Test Statistics to Non-normality and Specification Error in Confirmatory Factor Analysis. Psychological Methods 1: 16-29.

Hoogland, J.J., 1999: The Robustness of Estimation Methods for Covariance Structure Analysis. Groningen: Dissertation.

Hoyle, R.H.**/**Kenny, D.A., 1999: Sample Size, Reliability, and Tests of Statistical Mediation. S. 195-222 in: ***Hoyle, R.H.*** (Hrsg.): Statistical Strategies For Small Sample Research. Thousand Oaks et al.: Sage Publications.

Jackson, D.L., 2001: Sample Size and Number of Parameter Estimates in Maximum Likelihood Confirmatory Factor Analysis: A Monte Carlo Investigation. Structural Equation Modeling 8: 205-223.

Jackson, D.L., 2003: Revisiting Sample Size and Number of Parameter Estimates: Some Support for the N:q Hypothesis. Structural Equation Modeling 10: 128-141.

Kaplan, D., 1997: Statistical Power in Structural Equation Modeling. S. 110-117 in: ***Hoyle, R.H.*** (Hrsg.): Structural Equation Modeling: Concepts, Issues, and Applications. Thousand Oaks et al.: Sage Publications.

Kaplan, ***D.***, 2000: Structural Equation Modeling. Foundations and Extensions. Thousand Oaks et al.: Sage Publications.

Kline, ***R.E.***, 1998: Principles and Practice of Structural Equation Modeling. New York/London: Guilford Press.

MacCallum, R.C./Browne, M.W./ Sugawara, H.M., 1996: Power Analysis and Determination of Sample Size for Covariance Structure Modeling. Psychological Methods 2: 130-149.

Marsh, H.W./Hau, K.-T., 1999: Confirmatory Factor Analysis: Strategies for Small Sample Sizes. S. 251-284 in: ***Hoyle, R.H.*** (ed.), Statistical Strategies for Small Sample Research. London: Sage.

Muthén, B.O., 1983: Latent Variable Structural Equation Modeling With Categorical Data. Journal of Econometrics 22: 43-65.

Muthén, B.O., 1984: A General Structural Equation Model with Dichotomous, Ordered Categorial, and Continuous Latent Variable Indicators. Psychometrika 49: 115-132.

Muthén, B.O., 1993: Goodness of Fit With Categorical and Other Nonnormal Variables. S. 205-234 in: ***Bollen, K.A./Long, J.S.*** (Hrsg.), Testing Structural Equation Models. Newbury Park: Sage.

Muthén, B.O., 2002: Using Mplus Monte Carlo-Simulations in Practice: A Note on Assessing Estimation Quality and Power in Latent Variable Models. Vers. 2.0 vom 22.3.2002. Mplus Web Notes 1: 1-9 (www.statmodel/Mplus/examples/webnote.html)

Muthén, B.O./Curran, P.J., 1997: General Longitudinal Modeling of Individual Differences in Experimental Designs. A Latent Variable Framework for Analysis and Power Estimation. Psychological Methods 2: 371-402.

Muthén, B.O./Satorra, A., 1995: Technical Aspects of Muthén's LISCOMP Approach to Estimation of Latent Variable Relations with a Comprehensive Measurement Model. Psychometrika 60: 489-503.

Muthén, L.K./Muthén, B.O., 2001: Mplus. Statistical Analysis with Latent Variables. User´s Guide. Version 2. Los Angeles: Muthén & Muthén.

Muthén, L.K./Muthén, B.O., 2002a: How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. Structural Equation Modeling 9: 599-620.

Muthén, L.K./Muthén, B.O., 2002b: How to Calculate the Power to Detect that a Parameter is Different From Zero. www.statmodel.com/power.html, S. 1-6.

Olsson, U.H. et al., 2000: The Performance of ML, GLS and WLS Estimation in Structural Equation Modeling Under Conditions of Misspecification and Nonnormality. Structural Equation Modeling 7: 557-595.

Saris, W.E./Stronkhorst, L.H., 1984: Causal Modelling in Nonexperimental Research. Amsterdam: Sociometric Research Foundation.

Saris, W.E./Satorra, A., 1993: Power Evaluations in Structural Equation Models. S. 181-204 in: ***Bollen, K.A./Long, J.S.*** (Hrsg.), Testing Structural Equation Models. Newbury Park: Sage.

Saris, W.E./Satorra, A./Sörbom, D., 1987: The Detection and Correction of Specification Errors in Structural Equation Models. S. 105-129 in: ***Clogg, C.C.*** (Hrsg.), Sociological Methodology. Washington: Jossey-Bass.

Satorra, A./Bentler, P., 1986: Some robustness properties of goodness of fit statistics in covariance structure analysis. American Statistical Association: Proceedings of the Business and Economic Statistics Section, S.549-554

Satorra, A./Saris, W.E., 1985: The Power of the Likelihood Ratio Test in Covariance Structure Analysis. Psychometrika 50: 83-90.

Stone, C.A./Sobel, M.E., 1990: The Robustness of Estimates of Total Indirect Effects in Covariance Structure Models Estimated by Maximum Likelihood. Psychometrika 55: 337-352.

Tanaka, J.S., 1987: „How big is big enough?“: Sample Size and Goodness of Fit in Structural Equation Models with Latent Variables. Child Development 58: 134-146.

Urban, D./Mayerl, J., 2003: Autoritarismus und Ausländerablehnung. Ein vergleichender Hypothesentest. Universität Stuttgart, Institut für Sozialwissenschaften, unv. Manuskript.

Xie, Y., 1989: Structural Equation Models for Ordinal Variables: An Analysis of Occupational Destination. Sociological Methods and Research 17: 325-352.

Yuan, K.-H./Bentler, P.M., 1994: Bootstrap-corrected ADF Test Statistics in Covariance Structure Analysis. British Journal of Mathematical and Statistical Psychology 47: 63-64.

Anhang

```
TITLE:           Mplus-Inputfile für Schritt 1
DATA:            FILE IS daten.dat;
ANALYSIS:        ESTIMATOR= WLS;
VARIABLE:        NAMES ARE y1-y7 x1-x2;
                 USEVARIABLES ARE y1 y2 y3 y4 y5 y6 y7 x1 x2;
                 MISSING ARE y1-y7 x1 x2(99);
                 CATEGORICAL ARE y1-y7;
MODEL:           F1 BY y1 y2 y3 y4;
                 F2 BY y5 y6 y7;
                 F1 on F2 x1 x2;
                 F2 on x1 x2;
OUTPUT:          SAMPSTAT STANDARDIZED RESIDUAL TECH2 TECH4;

TITLE:           Mplus-Inputfile für Schritt 2
MONTECARLO: FILE IS identity.dat;
                 NAMES = y1-y7 x1-x2;
                 NOBSERVATIONS = 100000;
                 NREPS = 1;
                 SEED = 53487;
ANALYSIS:        ESTIMATOR = ML;
MODEL:           f1 BY y1@1 y2@.982 y3@.964 y4@.833;
                 f2 BY y5@1 y6@1.149 y7@1.040;
                 f1 on f2@.748 x1@.037 x2@-.202;
                 f2 on x1@-.195 x2@.074;
                 y1@0.345 y2@0.368 y3@0.391 y4@0.545;
                 y5@0.428 y6@0.244 y7@0.381;
                 f1@.335;
                 f2@.572;
OUTPUT:          RESIDUAL;

TITLE:           Mplus-Inputfile für Schritt 3
MONTECARLO: FILE IS pop.dat;
                 NAMES = y1-y7 x1-x2;
                 CUTPOINTS = y1-y7(-1.5 -1 -0.5 0 0.5 1);
                 CATEGORICAL = y1-y7(6);
                 NOBSERVATIONS = 142; !hier kann die Fallzahl variiert werden
                 NREPS = 5000;
                 SEED = 53487;
ANALYSIS:        ESTIMATOR = WLS;
MODEL:           f1 BY y1@1 y2*.982 y3*.964 y4*.833;
                 f2 BY y5@1 y6*1.149 y7*1.040;
                 f1 on f2*.748 x1*.037 x2*-.202;
                 f2 on x1*-.195 x2*.074;
                 f1*.335;
                 f2*.572;

TITLE:           Mplus-Inputfile für Schritt 4
MONTECARLO: FILE IS pop.dat;
                 NAMES = y1-y7 x1-x2;
                 CUTPOINTS = y1-y7(-1.5 -1 -0.5 0 0.5 1);
                 CATEGORICAL = y1-y7(6);
                 NOBSERVATIONS = 142;!hier kann die Fallzahl variiert werden
                 NREPS = 5000;
                 SEED = 53487;
ANALYSIS:        ESTIMATOR = WLS;
MODEL:           f1 BY y1@1 y2*.982 y3*.964 y4*.833;
                 f2 BY y5@1 y6*1.149 y7*1.040;
                 f1 on f2*.748 x1*.037 x2*-.202;
                 f2 on x1@0 x2*.074;
                 f1*.335;
                 f2*.572;
```