Multiple Imputation by Chained Equations: Eine Leistungsevaluation bei Schätzung von Strukturgleichungsmodellen mittels Monte-Carlo-Simulationen.

Von der Fakultät für Wirtschafts- und Sozialwissenschaften der Universität Stuttgart zur Erlangung der Würde eines Doktors der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.) genehmigte Abhandlung

Vorgelegt von

Andreas Wahl

aus Reutlingen

Hauptberichter:	Prof. Dr. Dieter Urban
Mitberichter:	Prof. Dr. Jochen Mayerl

Tag der mündlichen Prüfung: 30.10.2020

Institut für Sozialwissenschaften der Universität Stuttgart

2020

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis[1]

AME	Average Marginal Effect
Bias	Bias/Verzerrung der Parameter
CFI	Comparative Fit Index
Direct-ML	Direct Maximum Likelihood, auch: Full Information Maximum Likelihood (FIML)
EM	Expectation-Maximization
EMB	Multiple Imputation durch EM in Kombination mit Bootstrapping
FCS	Conditional Modeling; Fully Conditional Specification, auch: Multiple Imputation by Chained Equations (MICE)
FMI	Fraction of Missing Information
H0	Multiple Imputation anhand der Modellstruktur des Analysemodells durch Bayes-Schätzer
IM(s)	Imputationsmodell(e)
JM	Joint Modeling; Multiple Imputation unter Annahme einer multivariaten Normalverteilung
MAR	Missing At Random
MC	Monte-Carlo
MCAR	Missing Completely At Random
MCMC	Markov-Chain-Monte-Carlo
MDT(s)	Missing Data Technik(en)
MI	Multiple Imputation
ML	Maximum Likelihood
MLR	Maximum Likelihood Robust
MNV	Multiple Imputation durch Data Augmentation

[1] Das Abkürzungsverzeichnis beinhaltet Akronyme, die in der vorliegenden Arbeit durchgängig und über verschiedene Kapitel hinweg benutzt werden und die nicht nur in einzelnen Abbildungen/Tabellen vorkommen. Die Bezugnahme auf diese Akronyme könnte für die Leserschaft teilweise fremd wirken, weil für die Missing Data Techniken auf eine korrekte grammatikalische Verwendung der Akronyme verzichtet wird. Für die Ausfallmechanismen der Missing Values oder die SEM-Fit-Indices ist dies zum Teil auch der Fall. Beispielhaft seien folgende Sätze aufgeführt: ‚Wenn MAR gegeben ist, dann…' bzw. ‚FCS wird in vielen Softwarepaketen verwendet'. Korrekterweise müssten die Sätze ‚Wenn der MAR-Ausfallmechanismus gegeben ist, dann…' bzw. ‚Die FCS-Technik wird in vielen Softwarepakten verwendet' lauten. Da die Akronyme im Verlauf der Arbeit oft benutzt werden, wird, auch weil es in der zitierten Literatur meist so gehandhabt wird, darauf verzichtet. Stattdessen soll an dieser Stelle verdeutlicht werden, wie die Akronyme der Missing Data Techniken zu klassifizieren sind: Bei Direct-ML, MI und EM handelt es sich um spezifische Methoden/Verfahren von Missing Data Techniken. Bei EMB, FCS, H0, JM, MNV und PMM handelt es sich um Varianten/Techniken der Multiplen Imputation.

MSE	Mean Squared Error
NI	Non Ignorable, auch Not Missing At Random (NMAR) oder Missing Not At Random (MNAR)
PML	Penalized Maximum Likelihood
PMM	Predictive Mean Matching
PSR	Potential Scale Reduction Factor
RMSEA	Root Mean-Square Error of Approximation
SE-Bias	Bias/Verzerrung des Standardfehlers
SEM/SE-Modell	Strukturgleichungsmodellierung/Strukturgleichungsmodell
SRMR	Standardized Root Mean-Square Residual

Simulationskonfigurationen/Bezeichnung der Variablen in den statistischen Analysen

250	Fallzahl von 250
750	Fallzahl von 750
skew1	Symmetrische Konfiguration
skew2	Asymmetrische Konfiguration
skew3	Stark asymmetrische Konfiguration
mar1	Konfiguration ohne Missing Values (Referenzmodell)
mar2	Konfiguration mit 5 % Missing Values
mar3	Konfiguration mit 20 % Missing Values
mar4	Konfiguration mit 35 % Missing Values

Abstract – deutsch

Fehlende Werte sind ein omnipräsentes Phänomen der empirisch-quantitativ arbeitenden Sozialforschung. Da die meisten empirischen Datensätze fehlende Werte aufweisen, müssen möglichst geeignete Wege und Verfahren gefunden werden, um diese vor der Analyse angemessen zu behandeln. Das gilt auch für Analysen mit der Strukturgleichungsmodellierung (SEM). Zur Behandlung der fehlenden Werte, werden aktuell zwei Methoden präferiert: Direct-ML (ein Maximum Likelihood-Schätzverfahren) und die Multiple Imputation (MI). Während Direct-ML im SEM-Kontext in einigen Studien systematisch evaluiert wurde, gilt das für die MI nicht. Das kann darin begründet sein, dass für die MI verschiedene Varianten existieren, mit welchen die Ersetzung der fehlenden Werte jeweils unterschiedlich verläuft. Aufgrund ihrer Flexibilität ist die MI mittels *conditional modeling* (FCS) eine beliebte und oft eingesetzte Variante. Jedoch zeigt sich, dass deren Leistungsfähigkeit im SEM-Kontext kaum evaluiert wurde.

In der vorliegenden Monte-Carlo-Studie wird diese Forschungslücke geschlossen, indem eine umfangreiche systematische Evaluation von FCS durchgeführt wird und verschiedene Spezifikationen von FCS untersucht werden (darunter eine mit *predictive mean matching*; PMM). Zusätzlich werden zum gegenüberstellenden Vergleich weitere Missing Data Techniken (MDTs) evaluiert. Das sind Direct-ML, eine Einfachimputation mit Expectation-Maximization (EM), zwei MI-Varianten mittels *joint modeling* Ansatz (EMB und MNV) sowie eine Variante, die bei der Imputation der fehlenden Werte die Modellstruktur des Analysemodells berücksichtigt (H0). Alle sieben MDTs werden im Kontext von drei verschiedenen SEM-Populationsmodellen unter unterschiedlichen Simulationskonfigurationen getestet. Zu den variierten Testbedingungen gehören: die Fallzahl, die Variablenskalierungen und -verteilungen sowie der Anteil an fehlenden Werten. Die Performanz der MDTs wird hinsichtlich verschiedener SEM-Fit-Indices, der geschätzten Parameter, deren Effizienz und der geschätzten Standardfehler bewertet.

Im Ergebnis können zwei MDTs identifiziert werden, die unter allen Bedingungen zuverlässig arbeiten: Direct-ML und H0. Mit beiden Verfahren gehen für alle Performanzkriterien zufriedenstellende Ergebnisse einher. Die anderen MDTs schneiden dagegen etwas schlechter ab. Zwar liefern auch diese gute Ergebnisse für die Parameter und die Standardfehler (Letzteres mit Ausnahme von EM), allerdings nicht für die Fit-Indices (mit Ausnahme für das SRMR). In vielen Fällen kommt es nach dem Einsatz dieser MDTs zur fehlerbehafteten Modellbewertung. Deswegen werden die Ergebnisse der Arbeit in Handlungsempfehlungen übersetzt, die der Praxis als Orientierungshilfen dienen sollen, da sie angeben, unter welchen Bedingungen, mit welcher MDT zufriedenstellende Ergebnisse zu erwarten sind.

Abstract – englisch

In the empirical-quantitative (social) sciences, missing values are an omnipresent phenomenon. Because most empirical data sets include missing values, suitable ways and procedures must be found to treat them appropriately before the data can be analyzed. This also applies to statistical analyses with structural equation models (SEM). For the treatment of missing values, two methods are currently preferred: Direct-ML (a maximum likelihood estimation method) and Multiple Imputation (MI). While some studies evaluate the performance of Direct-ML in SEM analysis, this is less the case for MI. That might be because there are different variants for MI and each of those imputes the missing values in different ways. Because of its flexibility, the *chained equations* approach (FCS) for MI is a popular and frequently used variant. However, it turns out that in the context of analysis with SEM, its performance has hardly been evaluated.

The present Monte Carlo study expands on the current research through a comprehensive evaluation of FCS and several of its specifications (including a specification with *predictive mean matching*; PMM). For comparative reasons, additional Missing Data Techniques (MDTs) are evaluated. These are Direct-ML, single imputation with Expectation-Maximization (EM), two variants of MI using a *joint modeling* approach (EMB and MNV) and a variant that imputes missing values by taking into account the structure of the analysis model (H0). All MDTs are tested under three different SEM-population models and various configurations for simulation. The varied test conditions include: the number of cases, the scaling of the variables and their distribution, and the proportion of missing values. The performance of the MDTs is evaluated with respect to different fit indices used most prominently in SEM analysis, the estimated parameters, their efficiency, and the estimated standard errors.

As a result, two MDTs can be identified that work satisfactorily under all configurations: Direct-ML and H0. Both methods produce good results for all criteria examined, whereas the other MDTs perform slightly worse. Although they also provide good results for the parameters and the standard errors (the latter with the exception of EM), they do not deliver acceptable results for the fit indices (with the exception of SRMR). In many cases, the use of these MDTs leads to erroneous model evaluation and therefore to false rejections of SEMs. For this reason, the results of this study are translated into recommendations for action, which are intended to serve as guidelines for practitioners, since they show under which conditions satisfying results can be expected with each of the tested MDTs.

1 Einleitung

Viele statistische Methoden zur Analyse von empirischen Datensätzen sind darauf angewiesen, dass die zu analysierenden Daten vollständig sind. In nahezu allen empirischen Datensätzen, und damit auch in der empirischen Sozialforschung, lassen sich aber fehlende Werte (Missing Values bzw. Missings) beobachten: Der zu analysierende Datensatz ist damit unvollständig. Um dennoch die gewünschten statistischen Analysen durchführen zu können, kommen deshalb sogenannte Missing Data Techniken (kurz: MDTs) zum Einsatz. Diese ermöglichen es die fehlenden Werte entsprechend zu handhaben. Infolgedessen können die gewünschten statistischen Methoden eingesetzt und deren Schätzergebnisse analysiert werden.[2]

Für den Umgang mit fehlenden Werten kommen vielfach Methoden zur Anwendung, die als ad-hoc Methoden bezeichnet werden. Zu diesen zählen Ausschlussverfahren, wie der listenweise oder paarweise Fallausschluss, bei denen die fehlenden Werte von der Analyse ausgeschlossen werden, oder Einfachimputationen wie die Mittelwert- oder (stochastische) Regressionsimputation, bei welchen die fehlenden Werte durch Schätzwerte ersetzt bzw. *imputiert* werden. Problematisch an diesen Methoden ist, dass durch deren Einsatz die interessierende statistische Analyse verzerrte Schätzergebnisse produziert. Das kann sowohl die Parameterschätzwerte als auch die Standardfehler und damit die inferenzstatistischen Schlüsse betreffen. Alle diese ad-hoc-Methoden können daher nicht oder nur eingeschränkt empfohlen werden.[3]

Im Gegensatz zu den ad-hoc-Methoden, können sogenannte moderne MDTs in weit mehr Anwendungsfällen eingesetzt werden. Zu diesen Methoden zählen die Multiple Imputation (MI) aber auch Maximum Likelihood-Schätzverfahren (ML-Schätzverfahren) wie Direct-ML (auch Full Information Maximum Likelihood; FIML) oder Expectation-Maximization (EM). Diese MDTs schaffen es, die fehlenden Werte so zu handhaben, dass nach deren Einsatz die Parameterschätzwerte und die Standardfehler der statistischen Analyse unverzerrt sind. Aufgrund ihrer vielfältigen Einsatzmöglichkeiten sind sie den ad-hoc-Methoden in vielen Fällen vorzuziehen. Gleichzeitig sind diese Methoden auch in vielen Statistikpaketen implementiert (z. B. in EQS 6.4, M*plus* 7.31, SPSS 25 oder R 3.4.3)[4], sodass deren Einsatz auch ohne spezielle Programmier- oder Softwarekenntnisse möglich ist.

[2] Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 370960346.

[3] Diese Methoden sind nicht Gegenstandsbereich der Arbeit und werden nicht weiter vorgestellt. In der Literatur werden Gründe für den Verzicht dieser Techniken aber auch Bedingungen vorgestellt, wann sie dennoch eingesetzt werden können: Little/Rubin (2002); Allison (2003); Marsh (1998); Pigott (2001); Urban/Mayerl (2018).

[4] Aussagen zu diesen Statistikpaketen beziehen sich, wenn nicht anders angegeben, auf diese Versionen. Aussagen zu SAS beziehen sich auf die Version 15.1, Aussagen zu Stata auf die Version 16.

Neben den ML-Schätzverfahren stellt die Multiple Imputation den aktuellen *state of the art* in Sachen fehlende Werte dar. Grundsätzlich lässt sich das Vorgehen der MI in drei Phasen unterteilen, welche als Imputations-, Analyse- und als Poolingphase bezeichnet werden. In der Imputationsphase werden die fehlenden Werte im Originaldatensatz durch neu geschätzte Werte imputiert; es ergeben sich *m* vollständige Datensätze. Es folgt die Analysephase. Hierbei wird für jeden *m*-ten Datensatz die gewünschte statistische Analyse separat durchgeführt, was in *m* Ergebnissen resultiert. Das Zusammenführen dieser Analyseergebnisse geschieht in der Poolingphase, worin aus mehreren Ergebnissen ein durchschnittliches Ergebnis wird. Weil der Imputationsprozess unabhängig vom Analyseprozess ist, ist die MI nicht an ein bestimmtes Analyseverfahren gebunden. Damit können die vollständigen Datensätze auch mit jeglichen statistischen Methoden analysiert werden.

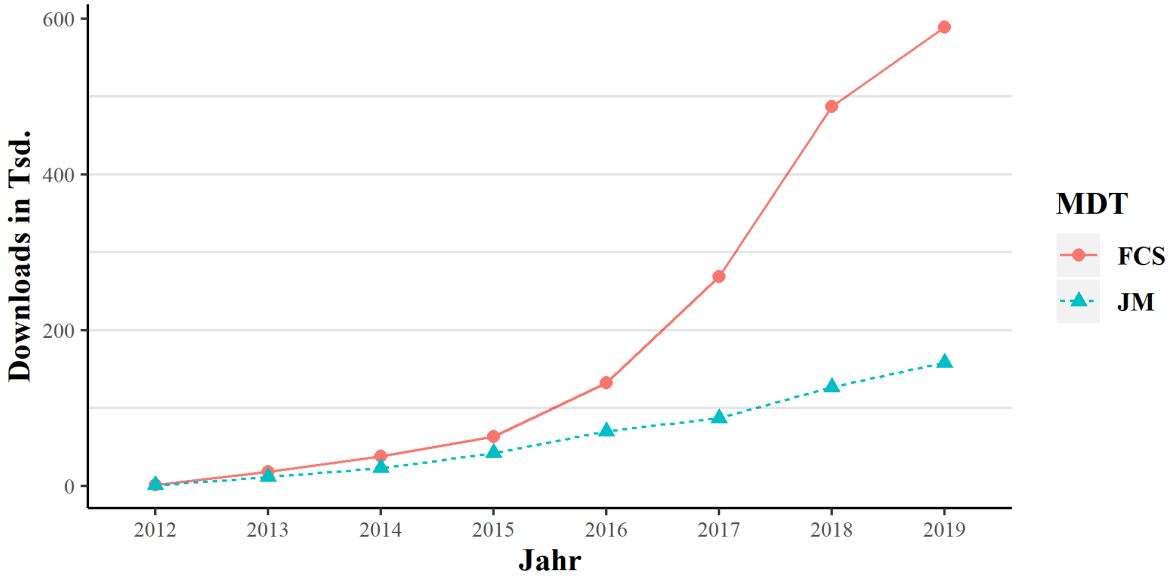
Während die Analyse- und die Poolingphase immer gleich ablaufen, gibt es Unterschiede in der Imputationsphase. Ziel dieser Phase ist es, Schätzwerte für die Missing Values zu generieren, die plausibel sind. Das bedeutet in diesem Zusammenhang, dass die Imputationen Eigenschaften aufweisen sollten, die es möglich machen, dass die Schätzergebnisse der Analysemethoden unverzerrt sind und sich korrekte inferenzstatistische Schlüsse ergeben (also unverzerrte Parameterschätzungen und Standardfehler) (vgl. Schafer 1997: 90; Kleinke 2017: 371 f.). Im Grunde gibt es zwei Möglichkeiten solche Imputationen zu generieren: zum einen die Möglichkeit des *joint modeling* (in dieser Arbeit JM), zum anderen die des *conditional modeling* (in dieser Arbeit FCS). In der ersten wird die Annahme einer gemeinsamen multivariaten Verteilung aller Variablen getroffen. Bei der Verteilung selbst, handelt es sich meist um eine multivariate Normalverteilung (Allison 2000; 2003; Honaker/King 2010; Schafer 1997; Tanner/Wong 1987). Problematisch ist dabei, dass deren Einsatzmöglichkeiten eigentlich auf metrisch skalierte Variablen eingeschränkt ist. Denn nur, wenn annähernd normalverteilte Variablen vorliegen und die multivariate Normalverteilung einigermaßen approximiert werden kann, ist davon auszugehen, dass die Imputationen plausibel sind. Sollten aber Daten vorliegen, die eine solche Annahme nicht rechtfertigen, wie es bei diskreten Variablen oder bei metrischen aber gleichzeitig schiefverteilten Variablen der Fall sein kann, ist es möglich, dass die imputierten Werte zu verzerrten Schätzergebnissen führen.

Liegen demnach Gründe vor, wonach die Annahme einer multivariaten Normalverteilung verletzt ist, sollte eher eine MI-Variante eingesetzt werden, welche dieser Annahme nicht unterliegt. Das ist mit der zweiten Möglichkeit zur Generierung der Imputationen gegeben. Hierbei wird für jede Variable ein an das Skalenniveau angepasstes, eigenes Schätzverfahren im

Imputationsmodell spezifiziert. Die Idee dahinter ist, dass bei einer binären Variable eine logistische, bei einer ordinalen eine multinominal logistische und bei einer metrischen eine lineare Regression spezifiziert wird, um die fehlenden Werte zu ersetzen. Weiterhin kann auch die *predicitive mean matching*-Schätzung (PMM) eingesetzt werden. Letztlich dürfte die Flexibilität dieser Variante gegenüber JM von Vorteil sein, da in empirischen Datensätzen nicht nur metrisch skalierte Variablen vorhanden sind, sondern oftmals auch diskrete (Raghunathan u. a. 2001; van Buuren 2007; 2012). Diese Technik wird als Multiple Imputation by Chained Equations[5] oder Fully Conditional Specification (kurz: MICE, bzw. FCS) bezeichnet.

Werden im Statistikprogramm R die Downloadzahlen einzelner Pakete für beide MI-Varianten herangezogen und diese als Indikator für den Einsatz der jeweiligen Technik verwendet, wird deutlich, dass FCS weitaus häufiger zum Einsatz kommt als JM (Abbildung 1). Gleichzeitig ist FCS auch in den großen Statistikpaketen implementiert (SPSS, SAS, Stata; für SPSS ist nur diese Variante der MI verfügbar).

Abbildung 1: Downloadzahlen der MI-Pakete in R



Anmerkungen: Downloadzahlen für FCS (Paket *mice*) und für JM (Pakete: *norm*, *norm2* und *Amelia*) zwischen 2012 (Beginn der Erfassung der Downloadzahlen für diese Techniken) und dem 31.12.2019.[6]

1.1 Relevanz der Arbeit

Trotz der weiten Verbreitung von FCS gibt es bisher nur wenig methodische Literatur, die sich mit ihr auseinandersetzt und ihre Leistungsfähigkeit systematisch evaluiert. Systematische Stu-

[5] Auch: Multivariate Imputation by Chained Equations.
[6] Bei eigenen Tabellen und Abbildungen wird auf eine Quellenangabe verzichtet. Tabellen und Abbildungen aus anderen Quellen werden entsprechend zitiert.

dien zur Evaluation statistischer Analysemethoden sind allerdings unerlässlich, weil die Verhaltenseigenschaften dieser Methoden nur bekannt sind, wenn die Samplegröße unendlich groß wird und die, für die Methoden notwendigen, Verteilungseigenschaften der Daten eingehalten werden (bspw. multivariate Normalverteilung). Liegen aber endliche Samplegrößen vor und/oder werden die Voraussetzungen der Methoden verletzt, dann darf davon ausgegangen werden, dass dies zu Schätzproblemen führt. Jedoch lässt sich nicht bestimmen, ab wann ein Einfluss zu erwarten ist und wie stark dieser sein wird. Deshalb ist es notwendig, den Einfluss empirischer Verteilungseigenschaften zu analysieren; dies geschieht durch Monte-Carlo-Simulationen[7] (MC-Simulationen) (vgl. Mooney 1997: 1 ff.; Paxton u. a. 2001: 289). Da für FCS aber nur wenige Studien vorliegen, die mit MC-Simulationen FCS untersuchen, ist auch noch weitestgehend unbekannt, wie FCS unter empirisch anzutreffenden Bedingungen zu bewerten ist.[8]

Exemplarisch lassen sich einige Studien anführen, welche FCS anhand ihrer Simulationsergebnisse bewerten.[9] Lang/Wu (2017) testen die Plausibilität der Imputationen mit FCS, wenn diese mit multinomial logistischen Regressionen oder mit *classification and regressions trees* (CART) produziert werden. Sie zeigen, dass die Handhabung der fehlenden Werte durch die multinomial logistische Regression innerhalb von FCS zu den besten Ergebnissen führt, wohingegen CART unzuverlässig arbeitet. Kropko u. a. (2013) testen drei FCS-Spezifikationen im Vergleich zu JM. Für FCS werden zum einen multinomial logistische Regressionen angelegt, um die Imputationen zu generieren und zum anderen *renormalized* logistische Regressionen. In der dritten Spezifikation verwenden sie lineare Regressionen. Es liegen bessere Ergebnisse für die FCS-Spezifikationen als für JM vor.

Einen ähnlichen Test führt McNeish (2017) durch. FCS wird unter zwei Spezifikationen getestet, wobei in der ersten Spezifikation PMM implementiert wird, in der zweiten lineare Regressionen. Beide Spezifikationen werden mit JM verglichen. Im Ergebnis zeigt sich, dass für JM unter allen Bedingungen bessere Ergebnisse vorliegen. FCS liefert dagegen in beiden

[7] Studien, die mit MC-Simulationen arbeiten, werden als MC-Studien bezeichnet.

[8] Weiterhin gibt es für FCS auch keine statistisch-theoretische Rechtfertigung, wonach die Imputationen plausibel sind. Denn auch FCS geht davon aus, dass die Imputationen aus einer gemeinsamen Verteilung der Variablen stammen. Während diese aber für JM bekannt ist und explizit vorgegeben wird (eine multivariate Normalverteilung) und deshalb auch die Bedingungen bekannt sind, wann die Imputationen plausibel sind (wenn für die empirischen Variablen eine multivariate Normalverteilung vorliegt), bleibt diese gemeinsame Verteilung für FCS unbekannt, da sie nur implizit angenommen und nicht vorgegeben wird. Da nun aber unbekannt bleibt, aus welcher gemeinsamen Verteilung die Imputationen stammen (bzw. ob eine solche überhaupt vorliegt), kann auch keine Aussage getroffen werden, wie die gemeinsame Verteilung der empirischen Variablen auszusehen hat, damit die Imputationen plausibel sind (siehe dazu Kapitel 3.4) (vgl. van Buuren u. a. 1999: 690; Raghunathan u. a. 2001: 88). Zwar scheint dies für die Anwendenden weniger ein Problem zu sein (vgl. van Buuren 2018: 120 f.), allerdings werden dadurch MC-Studien umso wichtiger.

[9] Weitere Ausführungen finden sich in Kapitel 5 dieser Arbeit.

Fällen eher unzuverlässige Ergebnisse; von deren Einsatz wird vor allem bei kleinen Fallzahlen abgeraten. Auch Yu u. a. (2007) beschäftigen sich mit PMM. Hierbei wird diese Spezifikation von FCS in einen Vergleich zur Spezifikation mit linearen Regressionen und zu JM gesetzt. Für die getesteten Varianten der MI lassen sich sehr gute Ergebnisse erwarten, solange die Daten nur wenig von einer Normalverteilung abweichen. Bei Verletzungen der Normalverteilungsannahme, wird die Performanz aller Varianten schlechter, wobei PMM leicht bessere Ergebnisse erzielt als JM und die FCS-Spezifikation mit linearen Regressionen. Dass PMM in vielen Szenarios eingesetzt werden kann, zeigt auch Kleinke (2017; 2018). In vielen Fällen zeigt sich, dass PMM eine sehr zuverlässige Methode zur Handhabung von fehlenden Werten ist.

Pritikin u. a. (2018) vergleichen die Performanz von FCS (mit *proportional odds model* und PMM) mit Direct-ML. Sie zeigen, dass FCS weniger gute Ergebnisse im Hinblick auf die Genauigkeit der Parameterschätzung erbringt als Direct-ML. Ein ähnliches Ergebnis liefert auch die Studie von Jia/Wu (2019). Als MDTs werden Direct-ML, JM und zweimal FCS herangezogen. FCS wird einmal mit logistischen Regressionen spezifiziert und einmal mit *random forests*. Sowohl Direct-ML als auch JM arbeiten unter den meisten Bedingungen zufriedenstellend. FCS ist dagegen weitaus weniger geeignet, wobei die *random forest*-Spezifikation die schlechteren Ergebnisse liefert. Demnach zeigt sich auch hier, dass FCS weniger gute Ergebnisse produziert als JM oder Direct-ML.

Wie aus diesen Ausführungen ersichtlich wird, liegen für FCS widersprüchliche Ergebnisse vor. Vor allem in denjenigen Studien, in denen FCS-Spezifikationen mit JM oder mit Direct-ML verglichen werden, gibt es durch FCS gegenüber den anderen Techniken, trotz der vermeintlichen Eignung für alle Skalenniveaus, keine Zugewinne im Hinblick auf die Plausibilität der Imputationen (McNeish 2017). FCS schneidet sogar in Fällen schlechter ab, in denen die Technik gegenüber JM oder Direct-ML eigentlich im Vorteil sein sollte, weil die vorliegenden Simulationsdesigns die Annahme einer gemeinsamen multivariaten Normalverteilung verletzen (Jia/Wu 2019; Pritikin u. a. 2018). Gleichzeitig zeigen aber Kropko u. a. (2013), Kleinke (2017; 2018) und Lang/Wu (2017), dass FCS dennoch auch sehr gute Ergebnisse hervorbringen kann. Bemerkenswert erscheint dabei vor allem die FCS-Spezifikation mit PMM, die in vielen Fällen gute Ergebnisse liefert und gegenüber JM zumindest gleichwertig ist, wenn nicht sogar die vorzuziehende Wahl zur Imputation darstellt.

Für die Bewertung von FCS gilt grundsätzlich, dass bisher nur wenig methodische Literatur vorliegt. Die Bewertung einzelner Spezifikationen erfolgt dabei bisher nur unter diskreten oder unter metrisch skalierten Variablen. Letzteres ist empirisch aber nicht immer zutreffend, denn

meist werden Variablen mit fünf Skalenpunkten bereits als metrische Variablen behandelt, weil sie oftmals die Voraussetzungen erfüllen, wonach sie als quasi-metrisch definiert werden können.[10] Wie die FCS-Spezifikationen für metrische Variablen (lineare Regressionen und PMM) quasi-metrische handhaben, und wie sie unter solchen Bedingungen abschneiden, bleibt unbekannt und kann nicht beurteilt werden, da hierzu keine Studien vorliegen. Zudem erfolgt die Bewertung von FCS bisher nur unzureichend in Kombination mit der in der empirischen Sozialforschung sehr verbreiteten Strukturgleichungsmodellierung. Auffallend dabei ist aber nicht die schlechtere Performanz von FCS gegenüber Direct-ML in Bezug auf die Parameterschätzungen und Standardfehler, sondern das nahezu komplette Außerachtlassen der Bewertung von FCS im Hinblick auf die Fit-Indices. Die Fit-Indices stellen für diese Analysemethode allerdings einen integralen Bestandteil der statistischen Analyse dar. Die Nicht-Berücksichtigung derselben führt letztlich dazu, dass keine Information darüber vorliegt, ob sich FCS für diesen Analysekontext überhaupt eignet.

Aufgrund der Tatsache, dass die Spezifikation der Imputationsmodelle bei FCS jeweils den Anwendenden überlassen ist und aus der bisherigen Forschung nicht deutlich wird, welche Spezifikation von FCS unter welchen Bedingungen am robustesten ist und die besten Ergebnisse liefert, kann den Anwendenden auch keine Empfehlung ausgesprochen werden, wann FCS geeignet ist. Das liegt auch daran, dass die betrachteten Studien bisweilen nur unzureichende (keine als quasi-metrisch definierten Variablen) und wenig realitätsnahe (kaum Variationen der Variablenverteilungen) Bedingungen testen sowie wenige Erkenntnisse über mögliche Analyseverfahren (Strukturgleichungsmodellierung) in Kombination mit FCS liefern. Das ist aber aufgrund der breiten Verfügbarkeit und vor allem auch der Beliebtheit dieser Technik problematisch. Diese Arbeit möchte zur Verbesserung des Forschungsstandes zu FCS beitragen. Aus diesem Grund sollen mit der Arbeit folgende forschungsleitenden Fragen beantwortet und die dazugehörigen Ziele erreicht werden:

1. Eignet sich FCS im Rahmen der Strukturgleichungsmodellierung als Alternative zur Ersetzung von fehlenden Werten? Zur Beantwortung muss eine dezidierte Leistungsevaluation von FCS bei der Schätzung von Strukturgleichungsmodellen und hinsichtlich deren Abschneiden in Bezug auf die Fit-Indices erfolgen.

[10] Variablen können als quasi-metrisch definiert und mit Methoden analysiert werden, die eigentlich für metrische Daten ausgelegt sind, wenn sie fünf Voraussetzungen erfüllen. Die Variablen sollten 1) sich ordnen lassen, sollten 2) numerische und semantisch äquidistante Kategorien aufweisen, sollten 3) fünf oder mehr Kategorien aufweisen, 4) für die jeweils genügend Fälle vorliegen und sollten 5) in Bezug zu einer latenten, metrischen Hintergrundvariablen stehen (vgl. Schnell u. a. 2013: 38; Urban/Mayerl 2018: 13 f.).

2. Sind mit FCS unter möglichst realitätsnahen und empirisch vorzufindenden Bedingungen plausible Imputationen zu erwarten? Eine breit gestreute Evaluation fehlt bislang, weshalb eine Untersuchung der Performanz von FCS bei diskreten als auch quasi-metrischen Variablen unter variierten Verteilungen zu erfolgen hat.
3. Welche der möglichen FCS-Spezifikationen ist unter möglichst vielen Bedingungen am robustesten und liefert die besten Ergebnisse? Dazu hat eine Evaluation der Technik unter verschiedenen Spezifikationen stattzufinden. Zumindest ein Vergleich der Möglichkeiten, metrische Variablen handzuhaben, sollte unter den angestrebten Bedingungen erfolgen.
4. Wie gut schneidet FCS im Vergleich zu den konkurrierenden Verfahren JM und Direct-ML ab, da sich beide Verfahren unter verschiedensten Simulationsbedingungen zumindest als gleichwertig, wenn nicht sogar besser erwiesen als FCS? Ziel hierbei muss es sein, herauszufinden, wann FCS gegenüber JM und Direct-ML vorzuziehen ist, oder wann FCS nicht mehr eingesetzt werden sollte.
5. Die Evaluationsergebnisse zu den einzelnen MDTs sollen in Handlungsempfehlungen für die empirische Praxis resultieren. Diese sollen es ermöglichen, einzuschätzen, ob sich die gewählte MDT für die vorliegenden empirischen Daten eignet.

1.2 Aufbau der Arbeit

Nachdem die Ziele für die vorliegende Arbeit definiert wurden, folgen in den nächsten drei Kapiteln Ausführungen zur Verfahrenslogik der hier untersuchten Methoden. Das zweite Kapitel befasst sich zunächst mit der Klassifikation der fehlenden Werte und deren Ausfallmechanismen. Das dritte Kapitel beschäftigt sich dann intensiv mit der MI. Im vierten Kapitel wird Direct-ML vorgestellt, da sie als Vergleichsparameter für FCS dienen soll. Im darauffolgenden fünften Kapitel wird der Forschungsstand diskutiert. Hierbei stehen bereits vorhandene Ergebnisse aus MC-Studien im Fokus. Unter anderem werden auch die bereits in diesem Kapitel aufgeführten Studien differenzierter betrachtet. Die Darstellung des Forschungsstandes soll weiterhin dazu dienen, die Bedingungen darzulegen, unter welchen FCS bereits getestet wurde. Zusätzlich dient der Forschungsstand dazu, weitere Desiderate aufzuzeigen, die in den oberen Ausführungen unerwähnt geblieben sind, um daraus eine Definition der zu füllenden Forschungslücke vorzunehmen. In Kombination mit den vorangegangenen Kapiteln, erfolgt im fünften Kapitel auch die Ableitung der zu testenden Hypothesen. Im Anschluss daran wird im sechsten Kapitel das Forschungsdesign vorgestellt. Neben der Rechtfertigung der MC-Simula-

tionsgrößen, wird darin auch die Definition der Bewertungsparameter vorgenommen, mit welchen die Performanz der einzelnen MDTs evaluiert wird. Zusätzlich wird die rechentechnische Umsetzung des Designs vorgestellt.

Die Kapitel sieben bis neun befassen sich schließlich mit der Analyse der Simulationsergebnisse im Hinblick auf die ausgewählten Bewertungsparameter. In Kapitel zehn werden die abgeleiteten Hypothesen getestet. Im elften Kapitel werden die gewonnenen Erkenntnisse anhand eines empirischen Beispiels exemplifiziert. Die Arbeit schließt mit dem zwölften und letzten Kapitel. Darin werden die Ergebnisse im Hinblick auf die forschungsleitenden Fragen und Ziele sowie im Hinblick auf die zuvor definierte Forschungslücke betrachtet. Die Ergebnisse werden weiterhin in Handlungsempfehlungen übersetzt, die der empirischen Forschungspraxis die Möglichkeit bieten soll, für die jeweils vorliegende empirische Datenbasis, die robusteste MDT zu identifizieren. Zudem werden die Schwächen und Einschränkungen der Arbeit diskutiert sowie Ausblicke auf mögliche weitere Forschungsvorhaben gegeben.

2 Fehlende Werte

Missing Values lassen sich grob in ‚Unit-Nonresponses' und in ‚Item-Nonresponses' einteilen. Bei Unit-Nonresponses wird die Beantwortung eines Fragebogens von einzelnen befragten Personen komplett verweigert, sodass im Datensatz keinerlei gültige Antworten verbleiben. Dies ist insbesondere bei Panelstudien relevant, wenn Befragte bei Nachbefragungen nicht mehr teilnehmen. Hingegen werden bei Item-Nonresponses manche Fragen beantwortet, andere bleiben hingegen unbeantwortet. Dies ist der Normalfall in der empirischen Realität, denn kaum ein Datensatz weist für alle Fälle auf allen Variablen gültige Werte auf. Für die vorliegende Arbeit stehen deshalb Item-Nonresponses im analytischen Fokus. Zur Bearbeitung von Unit-Nonresponses können neben Imputationsmethoden verschiedene Gewichtungsansätze verwendet werden, welche die Ausfälle kompensieren können (siehe: Brick/Montaquila 2009; Brick 2013).

Neben der Unterscheidung in Unit- und Item-Nonresponses lassen sich Missing Values zudem in drei verschiedene Muster einordnen sowie anhand ihres Ausfallmechanismus (Rubin 1976) klassifizieren. Das Muster der fehlenden Werte kann univariat, monoton oder willkürlich/generell sein (siehe Little/Rubin 1987; vgl. van Buuren 2012: 96 f.). Ein univariates Muster liegt vor, wenn nur eine Variable im kompletten Datensatz fehlende Werte aufweist. Bei einem monotonen Muster lassen sich die Variablen im Datensatz nach ihrem Anteil an fehlenden Werten ordnen. Sind die Variablen entsprechend geordnet, nimmt der Missinganteil mit jeder nach-

folgenden Variablen zu (vgl. Schafer/Graham 2002: 150). Liegt kein univariates und kein monotones Muster vor, dann handelt es sich um ein generelles Muster. Die Verteilung der Missings im Datensatz weist hierbei keinerlei Systematik auf und jeder Wert für jedes Set an Variablen könnte fehlen. Dieses Muster stellt auch die höchsten Anforderungen an die MDTs dar (vgl. van Buuren 2012: 96). Während eine Unterteilung der fehlenden Werte in ihr jeweiliges Muster aber eher nebensächlich ist, ist die Bestimmung des vorliegenden Ausfallmechanismus unerlässlich, denn er bestimmt darüber, welche Methode zur Behandlung der fehlenden Werte geeignet ist und welche nicht.

2.1 Die drei Ausfallmechanismen: MCAR, MAR und NI

Die Ausfallmechanismen beschreiben die Ursachen dafür, warum fehlende Werte vorliegen. Fehlen Werte im Datensatz rein zufällig, dann werden diese als *missing completely at random* (MCAR) bezeichnet. Das bedeutet, dass die fehlenden Werte in einer Variablen *Y* sowohl unabhängig von *Y* selbst sind als auch von den anderen, im Analysemodell befindlichen Variablen *X*. Wenn die Werte zufällig fehlen, dann können die entsprechenden Fälle aus der statistischen Analyse ausgeschlossen werden, weil es sich dann, bei der Gesamtheit aller Fälle mit fehlenden Werten, um eine zufällige Stichprobe aus dem ursprünglichen Sample handelt. Die Annahme wonach die Missing Values rein zufällig entstanden sind, ist in der Empirie allerdings wohl nur selten, wenn überhaupt, anzutreffen.

Missing at random (MAR) ist ein weiterer Ausfallmechanismus, der eine weniger restriktive Annahme trifft als MCAR. Unter einem MAR-Mechanismus wird verstanden, dass die Missing Values in einer Variablen *Y* abhängig von anderen, im Analysemodell befindlichen Variablen *X* sind, nicht aber von *Y* selbst. Die fehlenden Werte in *Y* lassen sich demnach auf beobachtete Werte (*X*) zurückführen (siehe Little/Rubin 1987; vgl. Schafer 1997: 10). Die ‚at random'-Formulierung des Ausfallmechanismus ist dabei etwas irreführend: Wenn die fehlenden Werte in *Y* abhängig von anderen Variablen (*X*) sind, dann können sie nicht mehr zufällig sein. Die Bezeichnung ‚at random' zielt in diesem Fall auf eine konditionale Zufälligkeit ab. Nur unter der Voraussetzung, dass die verursachenden Variablen berücksichtigt werden, sind die fehlenden Werte zufällig verteilt. Das liegt daran, dass die Fälle mit Missing Values keine zufällige Stichprobe des kompletten Datensatzes mehr darstellen, wie es bei MCAR der Fall ist. Stattdessen sind sie nur noch zufällige Stichproben innerhalb der Ausprägungen der verursachenden Variablen. Der dritte Ausfallmechanismus wird als *non ignorable* (NI), *not missing at random* (NMAR) oder als *missing not at random* (MNAR) bezeichnet. Im Gegensatz zu MAR, sind bei

NI die fehlenden Werte in *Y* nicht nur von anderen Variablen *X* abhängig, sondern auch von *Y* selbst (vgl. Spieß 2010: 119).

Sowohl Missing Values unter einem MCAR- als auch MAR-Ausfallmechanismus sind mit den dafür bestimmten MDTs gut handhabbar; darunter Direct-ML und die MI. Diese beiden Ausfallmechanismen werden in der Literatur deshalb auch als ignorierbar angesehen. Das wiederum bedeutet, dass es nicht notwendig ist, den Ausfallmechanismus explizit zu modellieren. Für Missing Values mit dem NI-Mechanismus müssen dagegen spezielle Methoden eingesetzt werden, welche eben dies erlauben.[11]

In der empirischen Praxis besteht allerdings das Problem, dass nur auf den MCAR-Mechanismus getestet werden kann. Der dafür vorgesehene Little´s Test prüft, ob MCAR vorliegt oder nicht. Liegt kein MCAR vor, dann ist zumindest von MAR auszugehen. Die Unterscheidung zwischen MAR und NI hat schlussendlich analytisch zu erfolgen, denn um zu prüfen, ob die fehlenden Werte in *Y* nur durch andere Variablen *X* zustande kommen und nicht zusätzlich durch *Y* selbst bedingt sind, müssten die fehlenden Werte bekannt sein; das ist aber meist nicht der Fall. Der empirische Praxiseinsatz, der hier vorgestellten MDTs muss demnach zwingend mit einer Diskussion des vorliegenden Ausfallmechanismus einhergehen. Denn erst wenn es nachvollziehbar ist, dass in dem jeweils vorliegenden Fall zumindest von einem MAR-Ausfallmechanismus auszugehen ist, lassen sich die Ergebnisse der statistischen Analyse entsprechend bewerten. Bleibt aber ungeklärt, ob MAR vorliegen könnte, dann können die eingesetzten MDTs und damit auch die Schätzergebnisse nicht als adäquat betrachtet werden.

3 Die Multiple Imputation

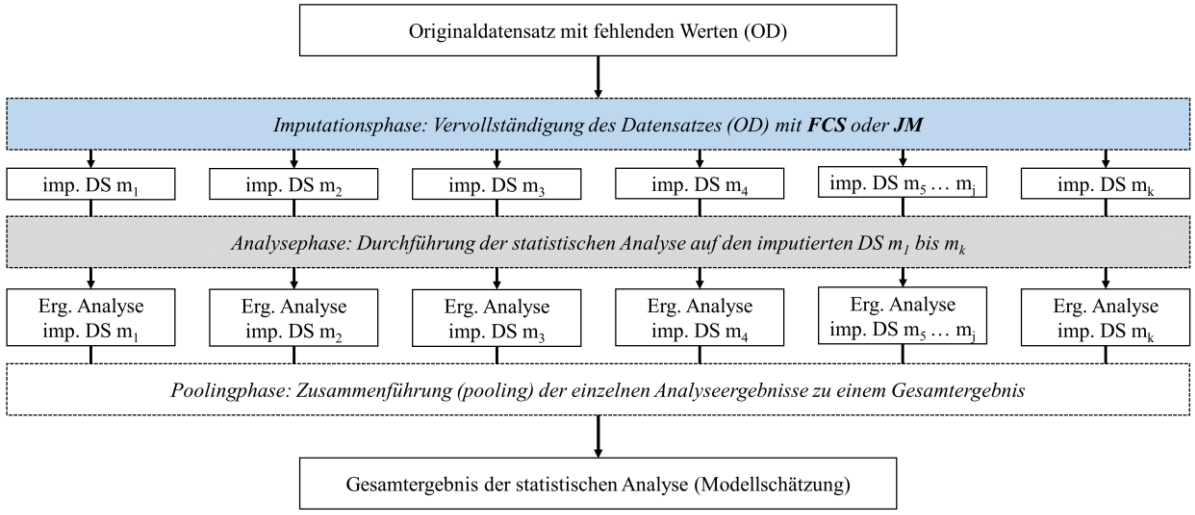
Bereits in der Einleitung dieser Arbeit wurde auf die drei Phasen der MI eingegangen: die Imputations-, Analyse- und die Poolingphase.[12] Im vorliegenden Kapitel steht nun zunächst der Ablauf der MI (Abbildung 2) im Vordergrund, bevor in den nächsten Unterkapiteln die Voraussetzungen für die Imputationsphase und die Verfahrenslogik derselben erläutert werden.

[11] Das kann mit Heckman´s Selection Model oder mit Pattern Mixture Models erreicht werden (vgl. Carpenter/Kenward 2013: 229 ff.; Little/Rubin 2002: 312 ff.). Problematisch ist allerdings, dass dabei nur dann von unverzerrten Ergebnissen auszugehen ist, wenn die nicht testbare Annahme gegeben ist, wonach das Modell zur Berücksichtigung des Ausfallmechanismus korrekt spezifiziert ist. Damit sind die Ergebnisse der interessierenden statistischen Analysen von diesem Modell abhängig und können sich, je nachdem wie es spezifiziert wird, unterscheiden. Aus diesem Grund müssen bei NI-Missings auch gute Gründe vorliegen, warum das Modell zur Berücksichtigung des Ausfallmechanismus korrekt ist und es müssen Sensitivitätsanalysen durchgeführt werden, bei welchen die Spezifikationen dieses Modells variiert werden (vgl. Allison 2009: 74 f.).

[12] Siehe dazu auch die Standardwerke zur Multiplen Imputation: Rubin (1987); Little/Rubin (2002); Schafer (1997); Allison (2000); van Buuren (2012).

Zunächst liegt ein Datensatz mit fehlenden Werten vor (OD). Er muss nun vervielfältigt werden, sodass *m* Datensätze vorliegen; das geschieht in der Imputationsphase. Dies ist notwendig, da mit den fehlenden Werten Unsicherheiten verbunden sind, die durch die Vervielfältigung der Datensätze berücksichtigt werden können. Die Unsicherheit, die mit den fehlenden Werten einhergeht, bezieht sich darauf, dass die ‚wahren' Werte, also diejenigen Werte die beobachtet worden wären, hätten die befragten Personen die zugrundeliegenden Fragen beantwortet, nicht bekannt sind. Liegt nur ein Datensatz vor, bei welchem die Missing Values mit Schätzwerten imputiert werden, dann werden diese Schätzwerte als wahre Werte begriffen, die keinerlei Unsicherheiten mehr aufweisen. Das kann zu verzerrten Parameterschätzungen führen, betrifft vor allem aber die Standardfehler. Diese werden in der Regel unterschätzt, was in weniger restriktiven inferenzstatistischen Schlüssen resultiert (die Wahrscheinlichkeit den alpha-Fehler zu begehen erhöht sich). Dadurch, dass aber *m* verschiedene Datensätze mit *m* Schätzwerten für die fehlenden Werte vorliegen, geht mit den Imputationen ein gewisser Grad an Varianz einher, der der Tatsache Rechnung trägt, dass die imputierten Werte nicht den wahren Werten entsprechen. Infolgedessen können auch angemessenere Parameterschätzungen vorliegen, vor allem aber werden die Standardfehler korrekt wiedergegeben.

Abbildung 2: Ablauf einer Multiplen Imputation



Nachdem der Datensatz mit fehlenden Werten (OD) in der Imputationsphase mit FCS oder JM entsprechend behandelt wurde, liegen *m* imputierte Datensätze vor (imp. DS m_1 bis m_k). Diese können dann so behandelt werden, als ob keine Missing Values vorgelegen hätten. Somit können auch jegliche statistischen Analysen damit durchgeführt werden. Dazu erfolgt in der Analysephase für jeden der *m* Datensätze die interessierende Analyse separat. Es liegen dann *m* verschiedene Analyseergebnisse vor (Erg. Analyse imp DS m_1 bis m_k). Infolgedessen müssen diese

Analyseergebnisse zu einem gesamten Ergebnis zusammengeführt werden: Das geschieht in der letzten Phase und betrifft neben der Parameterschätzung auch die Standardfehler (Poolingphase). Um dabei die Unsicherheit zu bewahren, die mit den fehlenden Werten einhergeht, hat Rubin bestimmte Regeln aufgestellt (*Rubin´s rules*).

Die Parameter des interessierenden Analysemodells sind dabei nichts anderes als die Durchschnittswerte des jeweiligen Parameters $\bar{\theta}$ über alle *m* Datensätze hinweg, mit $\hat{\theta}_i$ für den jeweiligen *i*-ten Datensatz (die Darstellung der Pooling-Regeln orientiert sich an Little/Rubin 2002):

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i. \tag{3.1}$$

Die Gesamtvarianz V_T des jeweiligen Parameters wird mit der between-Varianz V_B und der within-Varianz V_W berechnet:

$$V_W = \frac{1}{m}\sum_{i=1}^{m} SE_i^2. \tag{3.2}$$

$$V_B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \bar{\theta})^2. \tag{3.3}$$

$$V_T = V_W + V_B + \frac{V_B}{m}. \tag{3.4}$$

Wird danach die Wurzel aus V_T gezogen, erhält man den Standardfehler des betreffenden Parameters. Durch die Berücksichtigung der between- und within-Varianz (der *m* imputierten Datensätze) fallen die Standardfehler einer MI etwas konservativer aus, als wenn nur ein einziger Datensatz imputiert wird.

3.1 Voraussetzungen

Damit die Ergebnisse der statistischen Analyse auf MI-imputierten Datensätzen unverzerrt sind, unabhängig ob FCS oder JM eingesetzt wird, muss wenigstens ein MAR-Ausfallmechanismus vorliegen. Zudem muss ein genügend großes *m* an Datensätzen gewählt werden und das Imputationsmodell (IM) muss mit dem Modell der statistischen Analyse übereinstimmen. Eine ausreichend große Anzahl von Datensätzen ist notwendig, um mit der MI sicherzustellen, dass verlässliche statistische Inferenz daraus hervorgeht. Rubin (1987: 114 f.) zeigt, dass dabei ein *m* von drei bis fünf Datensätzen durchaus ausreichend ist, um gute Ergebnisse zu erzielen, und verweist darauf, dass es in vielen Fälle auch unnötig ist, mehr Datensätze zu generieren. Allerdings zeigen Graham u. a. (2007) anhand ihrer MC-Studie, dass die gewählte Anzahl der Datensätze einen erheblichen Einfluss auf die Standardfehler der Parameter, auf die p-Werte oder auch auf die Teststärke hat. Im Ergebnis weisen sie nach, dass die Anzahl der Datensätze, die nach der Berechnung durch Rubin notwendig wäre, um eine bestimmte Teststärke zu erhalten, nicht genug ist. Im Fazit kommen sie zum Schluss, dass *m* umso höher sein sollte, je höher der Anteil an fehlenden Werten ist. Sie empfehlen, dass bei einer *fraction of missing information* (FMI) von .5 mindestens ein *m* von 40 gewählt werden sollte (vgl. ebd.: 212). Da aber in den

meisten Fällen die FMI[13] geringer ist, als der Anteil der Missing Values, kann durchaus davon ausgegangen werden, dass *m* ungefähr diesem entsprechen sollte.

Bei der Übereinstimmung zwischen dem IM und dem Analysemodell geht es nicht darum die Struktur des späteren Analysemodells zu spezifizieren, sondern darum, die im Analysemodell benötigten Variablen zu berücksichtigen, denn das „imputation model is not intended to provide a parsimonious description of the data, nor does it represent structural or causal relationships among variables. The model is merely a device to preserve important features of the joint distribution (means, variances, and correlations) in the imputed values" (Schafer/Graham 2002: 167). Wenn im IM alle Variablen des Analysemodells vorhanden sind, dann entspricht das IM dem Analysemodell. Sollten nicht alle Variablen im IM vorhanden sein, die im Analysemodell von Bedeutung sind, liegt keine Übereinstimmung beider Modelle mehr vor (vgl. Meng 1994: 553 f.). In einem solchen Fall geht das IM davon aus, dass die nicht berücksichtigten Variablen unabhängig von den berücksichtigten Variablen sind. Damit wird der Einfluss, den bestimmte Variablen im Analysemodell hätten, im IM nicht berücksichtigt. Das bedeutet, dass die ersetzten Werte einen Zusammenhang nicht berücksichtigen, wie ihn das Analysemodell vorsieht. Das führt dazu, dass der spätere Zusammenhang im Analysemodell gegen null tendiert. Wie stark diese Verzerrung gegen null ausfällt, hängt dann aber davon ab, wie groß der ersetzte Anteil an fehlenden Werten ist (vgl. Graham 2012: 62; Spieß 2008: 61 ff.).[14]

Während also das Weglassen wichtiger Variablen im IM zu verzerrten Ergebnisschätzungen führen kann, kann das Hinzunehmen weiterer Variablen, obwohl diese im Analysemodell nicht von Interesse sind, in verbesserten Imputationen und damit besseren Schätzungen für die Parameter und Standardfehler resultieren. Solche Variablen werden als Hilfsvariablen bezeichnet. Insbesondere bei relativ hohen Korrelationen zwischen Hilfs- und Analysevariablen sollten sie

[13] Die FMI ist ein Maß dafür, wie viel Information eines Parameters aufgrund von fehlenden Werten nicht vorhanden ist (vgl. Allison 2002: 42). Sollten fehlende Werte nur in einer einzigen Variablen vorkommen, so entspricht die FMI deren Anteil. In multivariaten Datensätzen mit Missing Values auf mehreren Variablen, ist die FMI allerdings nicht mehr gleich deren Anteil (vgl. Longford 2005: 55). Stattdessen ist die FMI davon abhängig, ob Variablen vorhanden sind, die im IM stark mit der/den Variablen korrelieren, die Missing Values aufweist/aufweisen. In einem solchen Fall ist der Anteil an Missings in einem Datensatz ein konservativeres Maß als die FMI; denn in der Regel ist die FMI kleiner als der Anteil an fehlenden Werten (vgl. Rubin 1987: 114). Da die FMI nicht einfach zu berechnen ist, wird dazu geraten, die benötigte Anzahl an *m* Datensätzen mithilfe des Anteils an fehlenden Werten zu berechnen (siehe Bodner 2008 oder neu: von Hippel 2018) oder *m* eben gemäß diesem entsprechend zu wählen (vgl. von Hippel 2009: 278). Mit dem Anteil an Missing Values wird der Gesamtanteil in einem Datensatz gemeint, nicht der prozentuale Anteil an Fällen, die Missing Values aufweisen.

[14] Nicht nur sollten alle Variablen, die im Analysemodell analysiert werden, im IM berücksichtigt sein, sondern es müssen auch alle Variablenbeziehungen darin vorhanden sein. Das betrifft mögliche Interaktionen, Terme höherer Ordnung aber auch verschiedene Gruppen oder Zeitpunkte (Enders u. a. 2014; Enders/Gottschall 2011; von Hippel 2009). Für die Berücksichtigung von Interaktionen in Strukturgleichungsmodellen in Kombination mit der MI siehe außerdem: Chen u. a. (2011); Murray/Reiter (2016); Si/Reiter (2013).

in das IM aufgenommen werden (siehe Collins u. a. 2001; Yoo u. a. 2007; Yoo 2009). Gleichzeitig könnten diese Variablen zusätzliche Informationen zum Fehlen der Werte besitzen, was die Wahrscheinlichkeit erhöht, dass tatsächlich MAR vorliegt (vgl. Li 2010: 22).[15]

Generell sind für die Imputationsphase demnach eine dem Anteil an fehlenden Werten berücksichtigende Anzahl an *m* Datensätzen zu wählen sowie IMs anzustreben, die ein Maximum an Information über die fehlenden Werte und deren Ursachen enthalten, damit neben der Plausibilität der Imputationen auch die Wahrscheinlichkeit eines MAR-Ausfallmechanismus erhöht wird.

3.2 Die Imputationsphase

Im vorherigen Abschnitt wurden die allgemeinen Voraussetzungen für die MI vorgestellt. Sie liegen allen Varianten der Methode zugrunde. Bevor nun in den nächsten Kapiteln der Ablauf dargestellt werden kann, wie plausible Imputationen[16] mit FCS und JM erzeugt werden können, muss folgender Punkt angesprochen werden: Damit die Imputationen überhaupt plausibel sein können, müssen sie voneinander unabhängig sein. Dies wird bei der MI dadurch erreicht, dass die Parameter, welche für die Generierung der Schätzwerte für die Imputationen notwendig sind, aus der, auf den beobachteten Werten basierenden ‚wahren' posterior-Verteilung gezogen werden (vgl. Spieß 2008: 55).[17] Das Problem bei fehlenden Werten ist aber, dass diese posterior-Verteilung nicht direkt spezifiziert werden kann, weil eben unvollständige Informationen vorliegen. Ziel der MI ist es deshalb, die wahre posterior-Verteilung über *Markov-Chain-Monte-Carlo*-Techniken (MCMC) iterativ zu approximieren. Wenn sich die Markov-Kette (die

[15] Mit der Hinzunahme möglichst vieler Hilfsvariablen können aber Multikollinearitäts- und Konvergenzprobleme in der Imputationsphase auftreten. Aus diesem Grund werden Möglichkeiten diskutiert, welche die Anzahl der Hilfsvariablen reduzieren, aber gleichzeitig deren Informationsgehalt beibehalten können (bspw. durch Hauptkomponentenanalysen oder Summenscores). Siehe dazu: Eekhout u. a. (2014); Gottschall u. a. (2012); Howard u. a. (2015); Plumpton u. a. (2016).

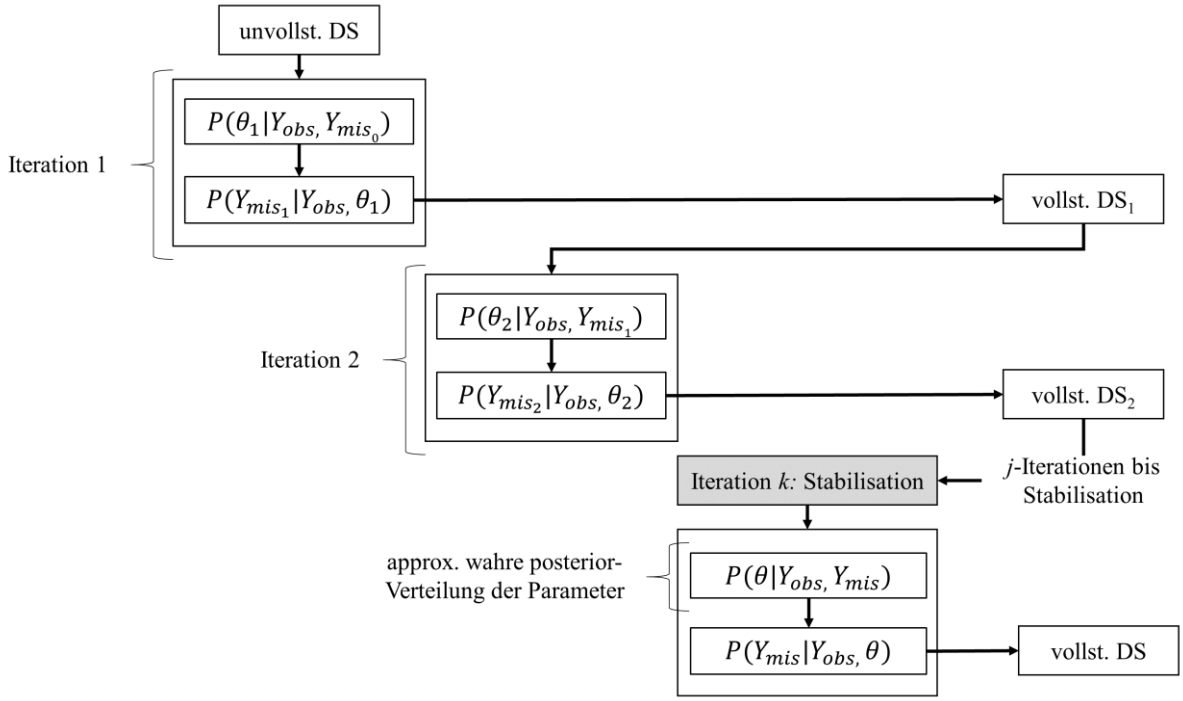
[16] Nochmals sei daran erinnert, dass ‚plausibel' in diesem Zusammenhang bedeutet, dass die Imputationen Eigenschaften aufweisen, die es möglich machen, dass die Parameter und Standardfehler der Analysemethoden unverzerrt sind, sodass sich korrekte inferenzstatistische Schlüsse ergeben.

[17] Die MI ist eine Methode, die auf dem Theorem von Bayes fußt. Danach ist eine posterior-Verteilung (auch a posteriori Verteilung) eine Verteilung der Parameter, welche die Wahrscheinlichkeit maximieren die beobachteten Daten zu erhalten, *wenn* die Vorstellungen bzgl. dieser Parameter, die vor der Beobachtung der Daten vorgelegen haben, miteinbezogen werden. Nach dem Bayes-Theorem geht die posterior-Verteilung aus einer prior-Verteilung (auch a priori Verteilung; angenommene Verteilung der Parameter, bevor die Daten erfasst worden sind) und der Likelihood-Funktion der vorliegenden Daten (Parameterwerte mit denen es am wahrscheinlichsten ist, die vorliegenden Daten zu beobachten) hervor (vgl. Held/Bové 2014: 167 ff.; Jackman 2009: 8 ff.). Da die posterior-Verteilung bei der MI aber im Normalfall keine prior-Informationen miteinbezieht, wird diese zum Großteil von den vorliegenden Daten bestimmt. Damit liefert die posterior-Verteilung der Parameter nichts anderes als eine Parameterverteilung, welche (ähnlich der Likelihood-Funktion) die Wahrscheinlichkeit maximiert, die beobachteten Daten zu erhalten (vgl. Allison 2002: 35; Enders 2010: 185 f.). Bei der ‚wahren' posterior-Verteilung der Parameter, handelt es sich um diejenige, die in einem Datensatz ohne Missing Values vorgelegen hätte.

Iterationskette) stabilisiert hat und konvergiert ist, dann ist davon auszugehen, dass die Approximation an die posterior-Verteilung erfolgreich ist (vgl. Gelman u. a. 2014: 275 f.).[18]

Wie diese Approximation erfolgt, soll an dieser Stelle etwas ausgeführt werden. Die Ausführungen und die Notationen orientieren sich an den Arbeiten von Schafer (1997: 105 f.) und van Buuren u. a. (2006: 1051 f.). Die angestrebte posterior-Verteilung ist eine gemeinsame Verteilung der Daten Y, die durch die Parameter θ (die Parameter welche die Wahrscheinlichkeit maximieren die beobachteten Daten zu erhalten) beschrieben werden kann: Sie wird als $P(\theta|Y)$ definiert. Problematisch an einem Datensatz mit Missing Values ist nun aber, dass Y aus einem beobachteten Teil Y_{obs} und einem unbeobachteten Teil Y_{mis} besteht. Die wahre posterior-Verteilung ist demnach $P(\theta|Y_{obs}, Y_{mis})$. Wie bereits erwähnt, kann diese Verteilung nicht direkt spezifiziert werden, sondern sie kann nur über eine bestimmte Anzahl an Iterationen hinweg approximiert werden. Dieser Prozess ist in Abbildung 3 dargestellt.

Abbildung 3: Markov-Chain-Monte-Carlo bei der Multiplen Imputation



Zunächst wird in der ersten Iteration die posterior-Verteilung für die Parameter auf Grundlage der beobachteten und der unbeobachteten Daten spezifiziert, wobei die Missing Values in diesem Schritt durch Platzhalter (deshalb auch Y_{mis_0}) ersetzt werden: $P(\theta_1|Y_{obs}, Y_{mis_0})$. Aus dieser Verteilung gehen die Paramater hervor, welche die vorliegenden Daten in der ersten Iteration am wahrscheinlichsten werden lassen (θ_1). Mithilfe dieser gezogenen Parameter wird dann eine

[18] Die Phase vor der Stabilisation wird als *burn-in*-Phase bezeichnet.

weitere posterior-Verteilung[19] spezifiziert: $P(Y_{mis_1}|Y_{obs}, \theta_1)$. Damit können die fehlenden Werte durch die tatsächlich beobachteten Daten und der zu diesem Zeitpunkt vorliegenden Parameter ersetzt werden. Im Grunde werden in diesem Schritt mithilfe von Modellschätzungen, wie lineare Regressionen, die Schätzwerte für die fehlenden Werte bestimmt, um diese dann zu imputieren. Am Ende der ersten Iteration liegt dann ein vollständiger Datensatz vor (DS_1), dessen Imputationen aber nicht aus der wahren posterior-Verteilung stammen. Infolgedessen ist eine erneute posterior-Verteilung für die Parameter zu spezifizieren, die nunmehr die aktualisierten Missing Values berücksichtigt (Y_{mis_1}). Durch die neu gezogenen Parameter (θ_2) können dann wieder die fehlenden Werte imputiert werden. Es liegt erneut ein vollständiger Datensatz vor (DS_2). Wieder basieren die Imputationen aber nicht auf den Parametern der wahren posterior-Verteilung.

Erst wenn sich nach einigen Wiederholungen die Iterationskette stabilisiert hat (Iteration *k*), ist davon auszugehen, dass die wahre posterior-Verteilung für die Parameter vorliegt: $P(\theta|Y_{obs}, Y_{mis})$. Werden bereits vorher Ziehungen für die Parameter getätigt, dann sind dies keine Ziehungen aus der gewünschten Verteilung, denn in jedem Schritt der Iterationskette beruht die posterior-Verteilung der Parameter *nur* auf den zu diesem Zeitpunkt vorliegenden Daten. Erst nachdem die Verteilung stationär geworden ist und sich die Iterationskette stabilisiert hat, handelt es sich bei den Ziehungen für die Parameter um die gewünschten Zufallsauswahlen aus der wahren posterior-Verteilung. Dann basieren auch alle folgenden Imputationen auf Parametern, die aus der wahren posterior-Verteilung gezogen wurden.

Mit den Zufallsziehungen wird gewährleistet, dass die imputierten Werte innerhalb eines Datensatzes unabhängig voneinander sind, dass die *m* Datensätze niemals identisch sind und dass die Unsicherheit, die mit den fehlenden Werten einhergeht, bewahrt wird.[20] Allerdings kann damit nicht sichergestellt werden, dass die Imputationen zwischen den *m* Datensätzen unabhängig sind. Denn unter Umständen können diese stark miteinander korrelieren. Das ist vor allem dann der Fall, wenn Datensätze aus direkt aufeinanderfolgenden Iterationen ausgewählt werden. Aus diesem Grund sollten die gewünschten *m* Datensätze entweder zufällig aus der Iterationskette ausgewählt werden (nachdem sie stabil ist), oder es sollten zwischen den

[19] Es handelt sich dabei um eine prädiktive posterior-Verteilung (auch a posteriori prädiktive Verteilung).
[20] Das heißt auch, dass dieselbe Analyse auf zwei verschiedenen Sets an *m* Datensätzen (wenn die MI zweimal, unter sonst gleichen Konfigurationen, auf demselben Ausgangsdatensatz angewendet wird), etwas unterschiedliche Ergebnisse liefern wird. Eine Ausnahme liegt dann vor, wenn für die MCMC-Ketten jeweils derselbe Startwert für den computergesteuerten Zufallsprozess gewählt wird.

ausgewählten Datensätzen genügend Iterationsschritte durchlaufen sein, damit keine Korrelationen zwischen den Imputationen der ausgewählten Datensätze vorliegen (vgl. Brown 2015: 412; Schafer 1997: 106). Da die Erfassung der Konvergenz und die Auswahl der *m* Datensätze Anwendungsprobleme darstellen, werden diese erst später thematisiert (siehe Kapitel 6.4).

Im Verlauf dieses Kapitels wurde erwähnt, dass eine gemeinsame Verteilung der Daten existiert, die durch die Parameter θ beschrieben werden kann. Gleichzeitig wurde dargelegt, wie die Approximation an diese Verteilung erfolgt, wenn fehlende Werte vorliegen. Vor allem bei multivariaten Daten besteht allerdings das Problem, dass nicht klar ist, *wie* die gemeinsame Verteilung der Daten aussehen soll. In den nachfolgenden Kapiteln werden unterschiedliche Herangehensweisen zur Lösung dieses Problems vorgestellt. Dies ist der zentrale Punkt, in dem sich FCS und JM methodisch unterscheiden.

3.3 Multiple Imputation unter einer multivariaten Normalverteilung (JM)

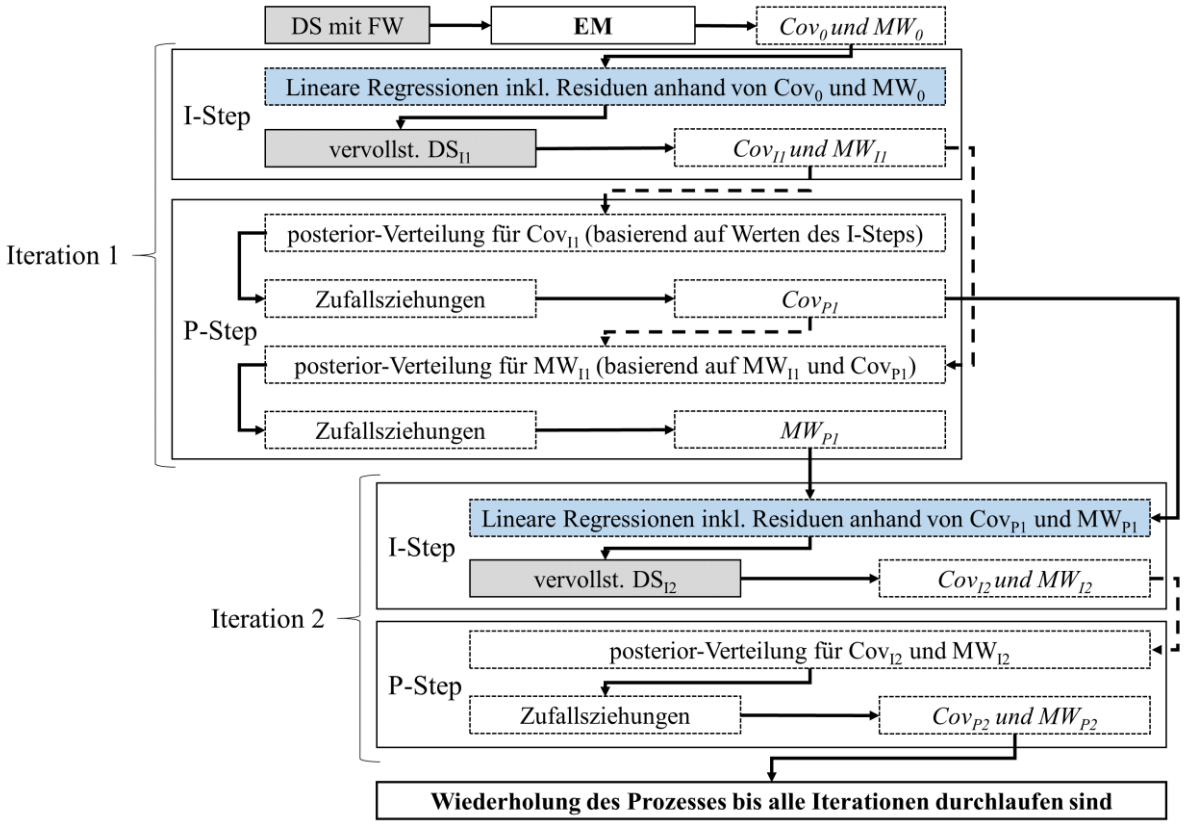
Für die Generierung von plausiblen Imputationen unter einer multivariaten Normalverteilung stehen zwei Möglichkeiten zur Verfügung: Das ist zum einen *data augmentation* (in dieser Arbeit MNV) und zum anderen eine Variante, die mit Hilfe von EM[21] in Kombination mit Bootstrapping die fehlenden Werte ersetzt (in dieser Arbeit EMB).

MNV ist ein zweistufiger, iterativer Prozess, der aus einem Imputation-Step (I-Step) und einem Posterior-Step (P-Step) besteht (siehe Abbildung 4).[22] Die Annahme hinter MNV ist, dass die gemeinsame Verteilung der Daten *Y* (sowohl die beobachteten als auch die unbeobachteten) eine multivariate Normalverteilung ist, die durch eine Kovarianzmatrix und einen Mittelwertvektor beschrieben werden kann (die Parameter θ). Zu Beginn des MNV-Prozesses werden für den ersten I-Step zunächst ein erster Mittelwertvektor und eine erste Kovarianzmatrix benötigt. Mithilfe von EM werden beide geschätzt (Cov_0 und MW_0). An dieser Stelle beginnt der erste Iterationsschritt mit dem ersten I-Step. Darin werden mit Cov_0 und MW_0 die Koeffizienten für die linearen Regressionen berechnet (Iteration 1; I-Step; Lineare Regression), um damit die Schätzwerte für die Missing Values zu generieren, die, nachdem ihnen zufällig Residuen zugeordnet wurden, imputiert werden. Infolgedessen liegt ein Datensatz vor, der vollständig ist und ersetzte Werte enthält (vervollst. DS_{I1}). Auf diesem werden erneut eine Kovarianzmatrix (Cov_{I1}) und ein Mittelwertvektor (MW_{I1}) berechnet. Im anschließenden P-Step nutzt

[21] Für einen Einblick in die Funktionsweise von EM, siehe Kapitel 4.2.
[22] MNV geht auf Tanner/Wong (1987) und Schafer (1997) zurück; dort finden sich auch um einiges detailliertere Informationen. Abbildung 4 und die nachfolgenden Ausführungen zum Ablauf von MNV sind von den Ausführungen in Allison (2003: 551 f.) und Enders (2010: 187-254) inspiriert.

MNV MC-Simulationen, um aus den posterior-Verteilungen der Kovarianzmatrix und des Mittelwertvektors (deshalb auch Posterior-Step) einen neuen Mittelwertvektor und eine neue Kovarianzmatrix zufällig zu ziehen (Iteration 1; P-Step; Zufallsziehung). Danach liegen eine neue Kovarianzmatrix und ein neuer Mittelwertvektor vor (Cov_{P1} und MW_{P1}), die sich von Cov_{I1} und MW_{I1} unterscheiden; MNV hat eine *erste* Iteration durchlaufen.

Abbildung 4: Data Augmentation (MNV)



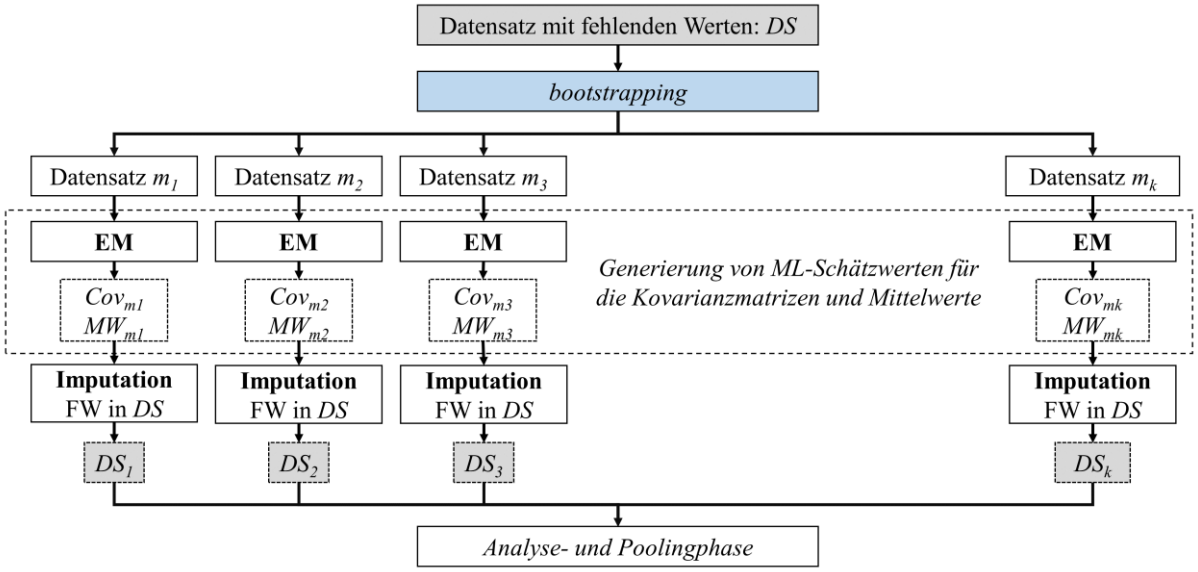
Anmerkungen: Die grau hinterlegten Felder in den Abbildungen 4 bis 8 zeigen an, zu welchem Zeitpunkt ein vollständiger Datensatz vorliegt. Die blau hinterlegten Felder deuten den Start einzelner Iterationen bzw. den Start des Prozesses zur Ersetzung der fehlenden Werte an.

Mit einem erneuten I-Step beginnt anschließend die zweite Iteration. Darin werden die neue Kovarianzmatrix und der neue Mittelwertvektor als Grundlage für die Regressionsgleichungen benutzt, um die fehlenden Werte vorherzusagen (Iteration 2; I-Step; Lineare Regression). Nachdem erneut Residuen zugeordnet wurden, werden neue Mittelwerte und Kovarianzen berechnet (Cov_{I2} und MW_{I2}). Infolgedessen gehen in einem neuen P-Step, durch Zufallsziehungen mittels MC-Simulationen, eine neue Kovarianzmatrix und ein neuer Mittelwertvektor hervor (Cov_{P2} und MW_{P2}). Beide dienen dann dem I-Step der dritten Iteration als Grundlage usw. Der Wechsel zwischen I- und P-Step wird über viele Male wiederholt, bis sich die Iterationskette stabilisiert hat und die gewünschte wahre posterior-Verteilung der Parameter vorliegt ($P(\theta|Y)$). Durch das Sampling der Mittelwertvektoren und Kovarianzmatrizen wird mit MNV die Vorgabe gemacht,

dass beides ausreichend ist, um die gemeinsame Verteilung der Daten zu beschreiben (wie es für eine multivariate Normalverteilung der Fall ist).

Alle Datensätze, die nach der Stabilisation der Iterationskette imputiert werden, können für die Analysephase ausgewählt werden. Zu beachten ist aber, dass aufeinanderfolgende Iterationen zum Teil stark miteinander korrelieren können. Es gilt demnach sicherzustellen, dass keine Abhängigkeiten zwischen den ausgewählten Datensätzen vorliegen. Im Gegensatz zu MNV liegt eine solche Problematik für EMB nicht vor, weil die *m* Datensätze mit einem Bootstrapping-Verfahren hergestellt werden. Dabei werden aus dem Originaldatensatz, *m* verschiedene Datensätze zufällig gezogen (siehe Abbildung 5). Aufgrund dieses mehrfachen und zufälligen Resamplings mit Zurücklegen, kann zwischen den Datensätzen Variation hergestellt werden, um die, mit den fehlenden Werten einhergehende, Unsicherheit zu berücksichtigen. Bootstrapping ersetzt bei EMB die Zufallsziehungen der Kovarianzmatrizen und Mittelwertvektoren aus deren posterior-Verteilungen (vgl. Honaker/King 2010: 564 f.; Honaker u. a. 2011: 4).

Abbildung 5: Ablauf EMB



Nachdem die *m* Datensätze (Datensatz m_1 … m_k) vorliegen, werden mit EM für jeden dieser Datensätze die Punktschätzungen der Kovarianzmatrix und des Mittelwertvektors generiert.[23] Schlussendlich liegen dann *m* Kovarianzmatrizen und Mittelwertvektoren vor (Cov_{m1} ... Cov_{mk} und MW_{m1} ... MW_{mk}). Diese werden im Anschluss dazu verwendet, um im *ursprünglichen*

[23] Wie MNV liegt auch EMB die Annahme zugrunde, dass Mittelwertvektoren und Kovarianzmatrizen (demnach die Parameter θ) die gemeinsame Verteilung der Daten beschreiben können – dass die gemeinsame Verteilung der Daten eine multivariate Normalverteilung ist.

Datensatz (DS) die Missings zu imputieren (Imputation FW in DS). Es liegen dann *m* ursprüngliche Datensätze vor (DS_1 ... DS_k), die sich in den ersetzten fehlenden Werten unterscheiden. Wurden die fehlenden Werte in den *m* Datensätzen mit EMB oder mit MNV ersetzt, können sie mit der gewünschten Analysemethode analysiert werden.

Sowohl mit EMB als auch mit MNV gehen alle Vorteile der MI einher: Unter der MAR-Annahme sind unverzerrte Parameter und Standardfehler zu erwarten und die Unsicherheit der fehlenden Werte wird berücksichtigt. Liegt zudem eine multivariate Normalverteilung der Daten vor, dann ist davon auszugehen, dass beide Techniken relativ ähnliche Analyseergebnisse liefern werden. Sollten aber Daten vorliegen, welche diese Annahme nicht rechtfertigen, dann kann es der Fall sein, dass beide Varianten unplausible Imputationen generieren, die zu verzerrten Schätzergebnissen führen. Das Ausmaß der Verzerrungen kann aber davon abhängen, wie stark die einzelnen Abweichungen von der Normalverteilung sind. Eine klare Aussage, wann und ob überhaupt mit Verzerrungen zu rechnen ist, kann an dieser Stelle nicht getroffen werden. Dafür können MC-Studien herangezogen werden (siehe Kapitel 5).

3.4 Multiple Imputation by Chained Equations (FCS)

Um bereits vorab auszuschließen, dass eine Verletzung der multivariaten Normalverteilung das Ergebnis beeinflusst, können die Imputationen, anstatt mit EMB oder MNV, mit FCS generiert werden. FCS beruht nicht auf der Annahme einer gemeinsamen multivariaten Normalverteilung, sondern legt für jede Variable mit fehlenden Werten ein eigenes Schätzmodell an, um die fehlenden Werte mit Hilfe aller anderen im Datensatz befindlichen Variablen zu schätzen. Die Idee dahinter ist, dass bei einer binären Variablen eine logistische, bei einer ordinalen eine multinominal logistische und bei einer metrischen eine lineare Regression spezifiziert wird, um die fehlenden Werte zu ersetzen. Weiterhin kann auch PMM eingesetzt werden.

Während JM eine gemeinsame multivariate Normalverteilung annimmt und nur die Parameter θ sampelt, welche diese beschreiben (die Kovarianzmatrix und die Mittelwertvektoren), trifft FCS keine solche Restriktion. FCS geht davon aus, dass eine gemeinsame Verteilung der Daten – wie immer diese aussehen mag – implizit vorhanden ist, „and that draws from it can be generated by Gibbs sampling the conditional distributions“ (van Buuren u. a. 1999: 690). Das bedeutet, dass Ziehungen aus der gewünschten posterior-Verteilung der Parameter ($P(\theta|Y)$) durch iteratives Sampling aus den konditionalen, univariaten Verteilungen jeder Variablen Y_k möglich werden ($P(\theta_k|Y_{-k})$). Damit wird $P(\theta|Y)$ implizit durch diese einzelnen Verteilungen jeder Variablen approximiert (vgl. van Buuren 2007: 227). Die gemeinsame Verteilung der Daten, die durch die Parameter θ beschrieben werden kann, wird mit FCS also erst

‚gesucht' und nicht wie bei JM bereits vorgegeben. Die nachfolgende Abbildung 6 stellt diesen Ablauf dar. Daran orientiert sich auch die Beschreibung des Vorgangs.[24]

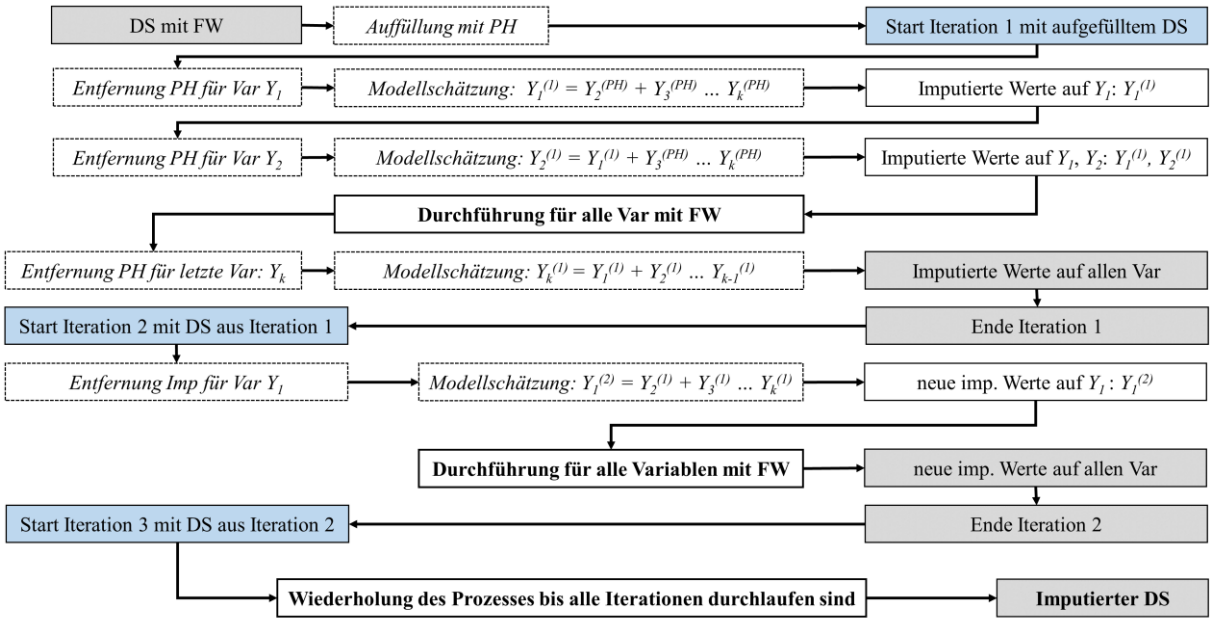
FCS benötigt für jede unvollständige Variable in einem Datensatz ein eigenes Schätzverfahren im IM, welches von der Skalierung der jeweiligen Variablen mit fehlenden Werten abhängig ist (Modellschätzung). Damit diese Schätzverfahren eingesetzt werden können, muss der Datensatz mit fehlenden Werten (DS mit FW) zunächst vervollständigt werden. Zuerst werden aus den univariaten Verteilungen der beobachteten Werte jeder Variablen zufällig Werte gezogen, die die fehlenden Werte im Datensatz ersetzen. Der Datensatz ist nach dieser Prozedur vollständig (Auffüllung mit Platzhaltern; PH). Im Folgenden werden für die Variable mit dem geringsten Anteil an Missing Values (Y_1) diese Platzhalter wieder entfernt. Es liegt damit ein Datensatz vor, der nur in einer Variablen fehlende Werte aufweist. Es folgt die Modellschätzung.

Im ersten Schritt wird ein Modell spezifiziert, das der Skalierung der Variablen gerecht wird (z. B. ein lineares Regressionsmodell für eine metrische Variable) und auf alle Informationen im Datensatz zurückgreift, also auf alle Variablen einschließlich ihrer Platzhalter (Modellschätzung: Y_1[25]). Die Modellschätzung liefert im Falle einer linearen Regression die Regressionskoeffizienten, Residualvarianzen sowie die Varianz-Kovarianzmatrix. Damit lässt sich die posterior-Verteilung der Parameter für diese Variable spezifizieren ($P(\theta_1|Y_{-1})$). Mittels MC-Simulationen werden daraus die Parameter für eine weitere lineare Regression gezogen (konzeptuell äquivalent zum Posterior-Step für MNV), um damit die fehlenden Werte vorherzusagen; es liegen dann Imputationen für Y_1 vor. Durch die zufällig gezogenen Parameter, die im zweiten Schritt der Modellschätzung dazu benutzt werden die Werte zu imputieren, wird erneut sichergestellt, dass die imputierten Werte voneinander unabhängig sind und dass die Unsicherheit, die mit den fehlenden Werten einhergeht, bewahrt wird.

[24] FCS geht auf die Autorenteams um van Buuren u. a. (2006) und Raghunathan u. a. (2001) zurück. Die vorliegenden Ausführungen basieren auf diesen Arbeiten.

[25] Die Hochzahl in der Abbildung nimmt Bezug auf die Iteration, in welcher sich FCS befindet. Darauf wird im Fließtext nicht nochmals verwiesen.

Abbildung 6: Fully Conditional Specification (FCS)



Wurde im ersten Schritt die erste Variable imputiert, folgt in einem zweiten Schritt die Imputation für die zweite Variable. Bei der zweiten Variablen handelt es sich wiederum um diejenige Variable, die nach der ersten Variablen, den geringsten Anteil an Missing Values aufweist (Y_2). Nachdem die fehlenden Werte der Variablen Y_1 ersetzt wurden, dient diese ersetzte Variable dann mit allen anderen Variablen als unabhängige Variable. Für Y_2 wird das für das Skalenniveau angebrachte Modell zugrunde gelegt (z. B. eine logistische Regression; Modellschätzung: Y_2). Daraus gehen erneut die Koeffizienten der Modelschätzung hervor (für die binäre und multinomiale logistische Regression sind das die Regressionskoeffizienten und deren Varianz-Kovarianz-Matrix), woraus sich dann deren posterior-Verteilung ergibt: $P(\theta_2|Y_{-2})$. Mittels Zufallsziehungen werden daraus die Parameter für eine weitere logistische Regression gezogen, um damit dann die fehlenden Werte zu imputieren. Nach diesem Schritt sind Y_1 und Y_2 vollständig und FCS fährt mit der dritten Variablen fort usw.

FCS hat eine Iteration durchlaufen, wenn die Missing Values in allen Variablen ersetzt wurden. Für eine zweite Iteration wird der Datensatz, der nun ersetzte Werte enthält, zugrunde gelegt (Start Iteration 2 mit DS aus Iteration 1) und das Ganze beginnt von vorne. Dieser Vorgang wird so lange wiederholt, bis sich die Kette stabilisiert hat. Ist das der Fall, dann können die Ziehungen für die jeweiligen Parameter als Zufallsziehungen aus der wahren posterior-

Verteilung gelten und alle nachfolgend imputierten Datensätze können zur Analyse herangezogen werden (Imputierter DS).[26] Im Gegensatz zu MNV werden demnach bei FCS die Parameter, welche die gemeinsame Verteilung der Daten beschreiben ($P(\theta|\, Y)$), durch iteratives Sampling aus den konditionalen (durch die Daten bedingten), univariaten Verteilungen approximiert ($P(\theta_1|Y_{-1})$, $P(\theta_2|Y_{-2})$ ... $P(\theta_k|Y_{-k})$). Und selbst wenn die einzelnen Verteilungen nicht kompatibel sein sollten und die gemeinsame Verteilung der Daten aufgrund dessen nicht zufriedenstellend approximiert werden kann, funktioniert FCS (vgl. Horton/Kleinman 2007: 83; van Buuren 2012: 111 f.)

Im Gegensatz zu JM sind die Annahmen, die über die Daten getroffen werden, mit FCS weniger restriktiv, denn je nach Skalenniveau wird das adäquate Schätzverfahren gewählt. Das hat zur Folge, dass sich FCS für verschiedenste Skalenniveaus eignet und das Problem wird umgangen, wonach eine multivariate Normalverteilung vorliegen sollte. Nachteilig ist allerdings, dass jedes Schätzmodell der fehlenden Werte eigene Probleme mit sich bringt. Denn die Methoden für metrische Variablen (die lineare Regression) beruhen weiterhin auf der Annahme, dass für die betreffenden Variablen metrische, normalverteilte Daten vorliegen und diese eben nicht allzu sehr davon abweichen (ob durch schiefe Verteilungen oder durch quasi-metrisch definierte Variablen). Zusätzlich ist es bei den Modellschätzungen mit den logistischen Regressionen möglich, dass keine stabile Schätzung erzielt werden kann, wenn vollständige Separation oder unvollständige Informationen aufgrund von gering besetzten Zellen vorliegen (vgl. van Buuren 2018: 67 ff.).

Besteht demnach die Gefahr, dass mit den linearen Regressionen die Beschaffung der Daten nicht beibehalten werden kann oder besteht die Gefahr von ungünstigen Werteverteilungen und wenig besetzten Zellen, wie es bei kleinen Fallzahlen und Variablen mit vielen Kategorien vorkommen kann, dann wird empfohlen eher PMM innerhalb von FCS einzusetzen.

3.4.1 Predictive Mean Matching (PMM)

PMM ist die Alternative für lineare Regressionen innerhalb von FCS, sollten metrische Variablen Missing Values aufweisen. Ein Vorteil von PMM liegt vor allem darin, dass dadurch die Beschaffenheit der Daten, wie sie vor den Imputationen vorliegt, intakt bleibt. Dies trifft vor allem dann zu, wenn es sich bei den Skalenniveaus der Variablen mit fehlenden Werten nicht mehr um metrische, sondern nur noch um quasi-metrische Skalenniveaus handelt. Denn unter solchen Umständen kommt es mit den linearen Regressionen zwangsläufig zu Imputationen,

[26] Auch hier gilt es wieder die Abhängigkeiten der einzelnen Iterationen zueinander zu beachten.

die außerhalb der Variablenskalierungen liegen, weil die Schätzwerte einer linearen Regression eben nicht an die Skalierung der Variablen gebunden sind. Solche ‚nicht-realistischen' Werte werden in der Forschung bereits diskutiert. Gegenstand der Diskussion ist die Frage, ob es nicht sinnvoller ist, diese Werte auf die jeweils naheliegende Kategorie zu runden. In mehreren Studien zeigt sich aber, dass es für die Analyse besser ist, die Werte so zu belassen wie sie geschätzt werden (Bernaards u. a. 2007; Horton u. a. 2003; Wu u. a. 2015).[27]

Ist es allerdings gewünscht, dass keine Werte außerhalb der Kategorien einzelner Variablen verbleiben, bietet sich PMM an.[28] Grundsätzlich läuft PMM innerhalb von FCS genauso ab, wie die Modellschätzungen mit den linearen Regressionen. Zunächst werden die Parameter der Regressionsschätzung ermittelt, um danach die posterior-Verteilung zu spezifizieren. Daraus werden per Zufallsziehungen neue Parameter gezogen, die dann erneuten linearen Regressionen als Grundlage für die Ermittlung der Schätzwerte dienen. Allerdings werden diese ermittelten Schätzwerte mit PMM nicht direkt imputiert. Stattdessen dienen sie dazu, ein Set mit sogenannten Spendern (*donors*) zu identifizieren. Bei diesen handelt es sich um tatsächliche Beobachtungswerte, deren vorhergesagten Werte demjenigen am nächsten liegen, der für den fehlenden Wert vorhergesagt wird. Wurde das Spenderset identifiziert, wird daraus per Zufallsziehung ein Spender ausgewählt; dieser wird für den betreffenden Missing Value imputiert. Mit PMM werden also unbeobachtete Werte, durch beobachtete Werte ersetzt, was in ‚realistischen' Imputationen resultiert (vgl. van Buuren 2018: 77 ff.).

Ein weiterer Vorteil von PMM ist die größere Robustheit gegenüber missspezifizierten IMs. Das ist zum Beispiel dann der Fall, wenn die Normalverteilungsannahme in Frage steht (bspw. bei Variablen mit wenigen Kategorien) oder wenn nicht-lineare Beziehungen im IM nicht berücksichtigt werden (vgl. Gaffert u. a. 2016: 2).[29] Das liegt daran, dass bei PMM die linearen Regressionen lediglich dafür verwendet werden, möglichst ähnliche, tatsächlich beobachtete Werte zu finden. Diese Eigenschaft bevorteilt PMM dahingehend, als dass keine Werte vorliegen können, die nicht auch beobachtet wurden. Das führt dazu, dass die ursprüngliche Verteilung der Daten gut erhalten bleibt und dass auch deren Beziehungen besser abgebildet werden.

---

[27] Zumal es kein Ziel der MI ist, Werte zu imputieren, die beobachtbar gewesen wären, also in die Kategorien fallen. Das Ziel einer MI ist es, trotz fehlender Werte, Aussagen über die zugrundeliegende Population zuzulassen und unverzerrte inferenzstatistische Schlüsse zu ermöglichen. Imputierte Werte, die nicht in die Kategorien der Variablen fallen und demnach auch nicht hätten beobachtet werden können sind demnach unproblematisch, sofern sie die Schlüsse auf die Population nicht verfälschen.

[28] Da das Verfahren Missing Values mit tatsächlich beobachteten Werten ersetzt, kann PMM zu den Hot-Deck-Verfahren gezählt werden (siehe Andridge/Little 2010).

[29] Vorsicht ist allerdings geboten, wenn die nicht berücksichtigten Beziehungen relativ stark sind. In einem solchen Fall ist es notwendig, diese in das IM aufzunehmen (vgl. Morris u. a. 2014: 11).

Gleichzeitig führt diese Eigenschaft auch dazu, dass mit PMM die Möglichkeit besteht, Variablen zu imputieren, die kein metrisches Messniveau aufweisen (binäre, ordinale und quasimetrische Variablen) (siehe Allison 2015; vgl. Horton/Kleinman 2007: 85; Vink 2015: 38).

Neben diesen Vorteilen geht mit PMM allerdings ein implizites Problem einher: Das betrifft das Set der Spender. Schlussendlich kann PMM nur sinnvolle Werte imputieren, wenn auch geeignete Spender gefunden werden. Sollten keine geeigneten Spender vorhanden sein, könnte die Performanz von PMM beeinträchtigt sein. Fällt das Set der Spender zu klein aus, könnte immer wieder derselbe Spender imputiert werden. Das wiederum könnte zu größeren Korrelationen zwischen den Imputationen führen und zu verringerter between-Varianz der *m* Datensätze (was wiederum zu unterschätzten Standardfehlern führt) (vgl. Schenker/Taylor 1996: 430). Auch kann es problematisch sein, wenn zu viele Spender ausgewählt werden. Denn dann finden sich im Spenderset Schätzwerte für die tatsächlichen Beobachtungswerte wieder, deren Distanz zum Schätzwert für den fehlenden Wert sehr groß ist. In einem solchen Fall könnten die Imputationen dieser Spender zu Verzerrungen führen, da sie den betreffenden fehlenden Wert nur unzulänglich widerspiegeln. In Bezug auf das Spenderset zeigt sich, dass sich mit drei bis fünf Spendern zufriedenstellende Ergebnisse erwarten lassen (vgl. Kleinke 2017: 399 f.).

Wird die Größe des Spendersets als potentielle Ursache für Verzerrungen berücksichtigt und enthält das IM alle Beziehungen, die im Analysemodell untersucht werden sollen (auch wenn leichtere Fehlspezifikationen für PMM unproblematisch erscheinen), dann könnte mit PMM, auch weil damit verschiedenste Skalierungen berücksichtigt werden können, ein ‚Allrounder' vorliegen, der für sehr viele verschiedene Datensatzstrukturen geeignet ist.

3.5 Multiple Imputation unter Berücksichtigung des Analysemodells (H0)

Den bereits vorgestellten MI-Varianten ist gemein, dass das IM nur in der Auswahl der Variablen dem Analysemodell entspricht und dabei nicht die Modellstruktur des Analysemodells berücksichtigt. Eine Variante, welche die Modellstruktur bereits in das IM einbezieht, ist die eigens für M*plus* implementierte H0. Wie FCS ist H0 flexibel gegenüber unterschiedlich skalierten Variablen, denn damit ist es möglich, sowohl diskrete als auch metrische Variablen zu berücksichtigen. Gleichzeitig arbeitet H0 für die metrisch definierten Variablen mit fehlenden Werten, wie auch FCS, mit konditionalen, univariaten Verteilungen. Weil H0 die Modellparameter explizit vorgibt, welche die Daten beschreiben können und die Imputationen der fehlenden Werte aus bedingten, univariaten Verteilungen generiert werden, kann H0 als eine Art Kombination aus FCS und JM angesehen werden.

Im Gegensatz zu den anderen Varianten der MI erfolgt die Behandlung der fehlenden Werte mit Hilfe eines Bayes-Schätzers direkt im Rahmen der Schätzung des Analysemodells. Wiederum ist das Ziel die Approximation der posterior-Verteilung ($P(\theta|Y)$), wobei die Daten Y wieder durch die Parameter θ zu beschreiben sind. Schlussendlich wird das spezifizierte Analysemodell wiederholt geschätzt, bis sich die Schätzung stabilisiert hat. Wenn die Konvergenz der MCMC-Kette eingetreten ist, dann sind die vorliegenden Parameter unabhängige Zufallsziehungen aus der angestrebten posterior-Verteilung (vgl. Asparouhov/Muthén 2010a: 16 f.). Wie bei den zuvor vorgestellten MI-Varianten enthält θ auch bei H0 die Parameter, welche die vorliegenden Daten am wahrscheinlichsten werden lassen. Der Unterschied besteht allerdings darin, dass explizit angegeben wird, *welche Variablenbeziehungen* dafür verantwortlich sind, um diese Daten zu erhalten. Dementsprechend werden zur Beschreibung der Daten auch nur die Variablenbeziehungen notwendig, die spezifiziert werden. Variablenbeziehungen, die nicht explizit vorgegeben sind, werden damit auch nicht beachtet. Im Gegensatz dazu geht JM davon aus, dass *alle* Variablenbeziehungen[30] zwischen den im IM befindlichen Variablen notwendig sind, um die Daten zu beschreiben.

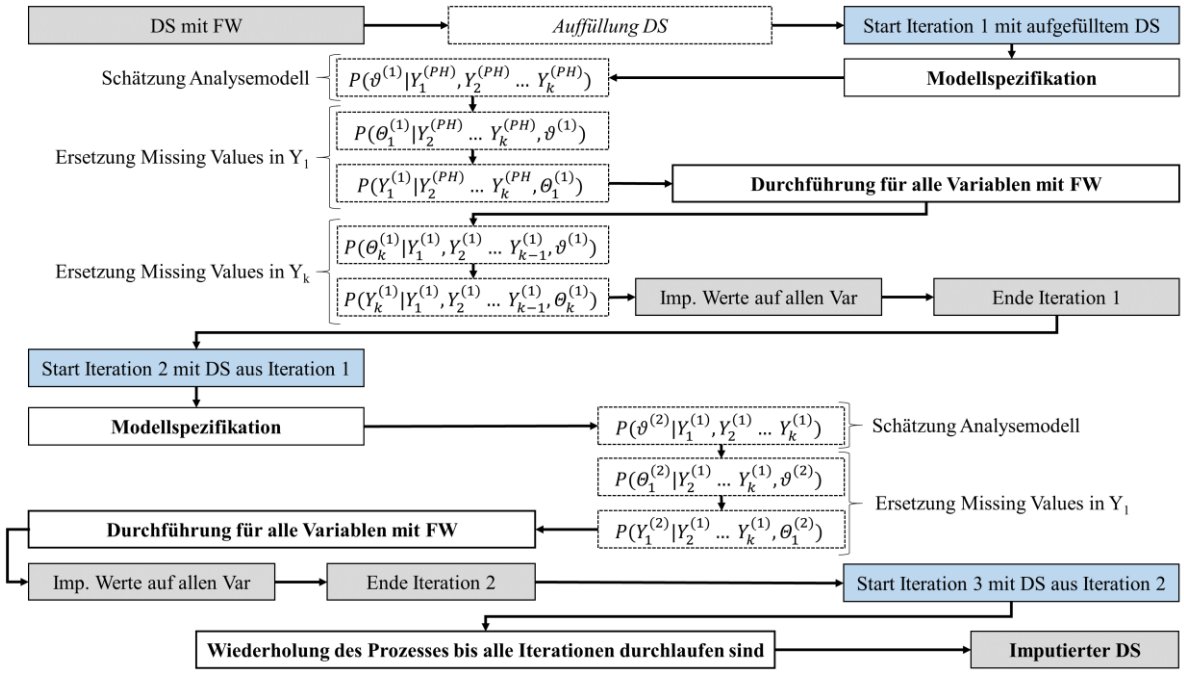
In Abbildung 7 ist, unter der Bedingung, dass Missing Values vorliegen, der Schätzprozess mit H0 konzeptionell dargestellt.[31] Anzumerken ist, dass θ neben den Parametern zur Ersetzung der Missing Values auch die Parameter des Analysemodells enthält. Aus diesem Grund soll θ die gesamten Parameter abbilden, während ϑ die Parameter des Analysemodells und Θ die Parameter für die Missing Values enthält.

Zu Beginn des Prozesses werden die Modellparameter des Analysemodells benötigt. Anhand eines vorläufig vervollständigten Datensatzes wird das Analysemodell geschätzt. Die Komplexität dieses Modells bestimmt dabei die Schritte, die notwendig sind, damit die Modellparameter aus dem MCMC-Prozess hervorgehen. Liegen binäre oder ordinale Variablen vor, müssen dem Prozess zur Bestimmung der Modellparameter zwei weitere Schritte hinzugefügt werden. Im ersten Schritt werden diese Variablen (hier: Z) in unbeobachtete metrische Variablen überführt (Z^*). Im zweiten Schritt müssen dann die Schwellenwerte für diese generiert werden. Z^* inklusive der berechneten Schwellenwerte bilden dann die binären oder ordinalen Variablen ab. Weisen solche Variablen Missing Values auf, werden diese direkt bei der Generierung der metrischen Variablen (Z^*) berücksichtigt (vgl. Asparouhov/Muthén 2010a: 9 ff.). Nachdem das

[30] Es werden die Varianzen, Kovarianzen und Mittelwerte aller Variablen im IM gesampelt.
[31] Die Ausführungen basieren auf den Arbeiten von Asparouhov/Muthén (2010a), Asparouhov/Muthén (2010c) und Kaplan/Depaoli (2012: 657 f.).

Modell geschätzt wurde und dabei die diskreten Variablen (ggf. mit Missing Values) berücksichtigt wurden, liegen die Modellparameter ϑ in erster Iteration vor.[32]

Abbildung 7: H0-Imputation



Diese Parameter werden dazu benutzt, die Missing Values auf den metrischen Variablen zu ersetzen. Hierzu werden zusätzliche Schritte notwendig, deren Anzahl davon abhängig ist, wie viele metrische Variablen fehlende Werte aufweisen (in Abbildung 7 sind es *k* Variablen und damit *k* Schritte). Die Missing Values auf diesen Variablen, werden, wie bei FCS, mittels linearer Regressionsschätzungen ersetzt. Auch hierbei ist der Prozess zweistufig: Im ersten Schritt wird mithilfe der vorliegenden Modellparameter und den Daten die posterior-Verteilung der Parameter für diese Variable spezifiziert ($P(\Theta_1|Y_{-1}, \vartheta)$, die konditionale univariate Verteilung für die Variable). Mittels Zufallsziehungen gehen daraus die Parameter für die linearen Regressionen hervor. Diese werden dann dafür benutzt, um die fehlenden Werte vorherzusagen; es liegen danach Imputationen für Y_1 vor. Im nachfolgenden Schritt werden die Missings in Y_2, dann in Y_3 usw. imputiert. Der Vorgang ist abgeschlossen, nachdem alle Missing Values ersetzt wurden (Ende Iteration 1). Wiederum wird durch die Zufallsziehungen sichergestellt, dass die imputierten Werte voneinander unabhängig sind und dass die Unsicherheit, die mit den fehlenden Werten einhergeht, bewahrt wird.

Basierend auf dem aktualisierten Datensatz erfolgt ein zweiter Iterationsschritt. Darin werden zunächst die Informationen für die einzelnen Modellparameter aktualisiert, (das heißt, das

[32] Der Einfachheit halber wird die Modellschätzung in Abbildung 7 ohne diese Zwischenschritte dargestellt.

Analysemodell inklusive der Berücksichtigung der diskreten Variablen, ob mit oder ohne Missings, wird erneut geschätzt), bevor die Missing Values in den metrischen Variablen ersetzt werden. Der Prozess wird so lange wiederholt, bis die Iterationskette konvergiert ist. Ist die Kette stabil, sind alle Ziehungen für die Parameter θ, in diesem und den nachfolgenden Iterationsschritten, unabhängige Zufallsziehungen aus der gewünschten posterior-Verteilung ($P(\theta|\, Y)$). Wie bei FCS wird diese Verteilung der Parameter demnach durch iteratives Sampling aus den konditionalen (durch die Daten bedingten) Verteilungen approximiert ($P(\vartheta|Y)$, $P(\Theta_1|Y_{-1},\vartheta)$, $P(\Theta_2|Y_{-2},\vartheta)$ ... $P(\Theta_k|Y_{-k},\vartheta)$).

Was H0 gegenüber den anderen MI-Varianten benachteiligt, ist zum einen, dass damit Hilfsvariablen nur durch Einbezug in die Modellstruktur berücksichtigt werden können und zum anderen, dass das spezifizierte Analysemodell korrekt sein und mit dem Populationsmodell übereinstimmen sollte.[33] Weiterhin gilt auch, dass die Ersetzung der Missing Values bei metrischen Variablen auf der Annahme beruht, dass für die betreffenden Variablen metrische Daten vorliegen, welche nicht zu sehr von der Normalverteilung abweichen (da die bedingten Verteilungen aus denen die Imputationen hervorgehen Normalverteilungen sind). Ob und inwiefern die Berücksichtigung der Missing Values von diskreten Variablen durch die Generierung einer metrischen Hintergrundvariablen erfolgreich ist, ist zudem davon abhängig, ob die zugrundeliegende Normalverteilung, von welcher der dafür zuständige Algorithmus startet, um die Schwellenwerte zu generieren, einen vernünftigen Startpunkt darstellt (vgl. Asparouhov/Muthén 2010a: 13 ff.).

3.6 Zusammenfassung: Einflussfaktoren auf die Performanz der Methode

Wird zur Imputation der fehlenden Werte die MI herangezogen, müssen folgende, grundlegende Punkte beachtet werden:

1. Es muss zumindest ein MAR-Ausfallmechanisums realistisch sein.
2. Die Anzahl der Datensätze (m) muss in etwa dem Anteil der Missings entsprechen.
3. Die IMs müssen mindestens alle inhaltlich relevanten Variablen und inhaltlich relevanten Variablenbeziehungen berücksichtigen (die Hinzunahme von Hilfsvariablen in das IM ist optional, kann aber die Plausibilität der Imputationen erhöhen und MAR wahrscheinlicher werden lassen).

[33] Asparouhov/Muthén (2010c) zeigen zwar, dass leichte Fehlspezifikationen kein Problem darstellen sollten: „While the H0 imputations relies to some extent on the correct specification of the imputation model that specification has only a limited impact on the final data analysis and it appears that minor misspecifications of the imputation model are harmless" (vgl. ebd.: 22). Was sie aber genau unter ‚leichten Fehlspezifikationen' verstehen, wird nicht ausgeführt. Es ist deshalb sinnvoll davon auszugehen, dass Fehlspezifikationen einen Einfluss nehmen.

4. Es ist sicherzustellen, dass Konvergenz vorliegt. Erst dann entsprechen die Imputationen zufälligen Werten, die unabhängig voneinander sind und die die Unsicherheit, die mit den fehlenden Werten einhergeht, berücksichtigen.
5. Aufeinanderfolgende Datensätze in der MCMC-Sequenz sollten nicht für die Analyse ausgewählt werden, da die Imputationen zwischen diesen Datensätzen unter Umständen sehr stark miteinander korrelieren können (gilt nicht für EMB).

Wird all dies sichergestellt, entscheidet die ausgewählte MI-Variante darüber, ob nach der Imputationsphase plausible Imputationen vorliegen und aus der Analyse- und Poolingphase unverzerrte Parameter und Standardfehler hervorgehen. Problematisch ist jedoch, dass jede MI-Variante zusätzliche Annahmen über die vorliegenden Daten mit in die Imputationsphase einbringt.

Bis auf PMM gehen alle vorgestellten Varianten davon aus, dass entweder eine gemeinsame multivariate Normalverteilung der Daten vorliegt (EMB, MNV), oder dass die als metrisch definierten Variablen normalverteilt sind (FCS, H0). In empirischen Datensätzen können diese Annahmen bisweilen aber nicht eingehalten werden und meist liegen Gründe vor, die diese Annahmen nicht rechtfertigen. Das liegt daran, dass eigentlich metrische Variablen oftmals auf wenige Kategorien zugeschnitten sind und damit nur noch als quasi-metrisch zu definieren sind, was bereits im Vorhinein eine (multivariate) Normalverteilung nur approximativ möglich macht. Zudem können solche Variablen häufig bereits univariat aufgrund von schiefen, endlastigen oder bimodalen Verteilungen nicht mehr als normalverteilt gelten.

Gleichzeitig liegen in empirischen Datensätzen auch diskrete Variablen vor, sodass EMB oder MNV eigentlich anhand ihrer Verfahrenseigenschaften nicht oder nur bedingt für empirische Datensätze geeignet sind. Infolgedessen wären FCS oder H0 angebracht, um die Imputationen zu generieren. Problematisch für FCS könnte aber sein, dass gering besetzte Zellen das Schätzergebnis beeinflussen, was vor allem bei kleinen Fallzahlen und vielen Kategorien in den betreffenden Variablen der Fall sein kann. Für H0 stellt sich dagegen die Frage, inwiefern die Berücksichtigung der Missing Values auf diskreten Variablen direkt im Rahmen der statistischen Analyse plausible Imputationen möglich macht. Zusätzlich liegt für H0 das Problem vor, dass das IM mit dem Populationsmodell übereinstimmen muss, aus dem die Daten hervorgegangen sind. Da für empirische Daten aber nicht mit Sicherheit davon ausgegangen werden kann, dass das Analysemodell dem Populationsmodell entspricht, stellt sich die Frage, was passiert, wenn H0 ein falsches IM anlegt.

All diese Restriktionen treffen für PMM nicht zu. Stattdessen sind die Imputationen bei PMM von dem zugrundeliegenden Spenderset abhängig. Je kleiner dieses ist, desto weniger potentielle Spender stehen für jeden fehlenden Wert zur Verfügung und je größer es ist, desto eher stehen Spender zur Verfügung, die eine große Distanz zum fehlenden Wert aufweisen. Zwar legt die Literatur zu PMM nahe, dass es sich dabei um einen ‚Allrounder' handeln könnte, aber wenn nur wenige Kategorien in den Variablen mit fehlenden Werten vorliegen, wie es bei diskreten oder quasi-metrischen Variablen der Fall sein kann, und der Anteil an fehlenden Werten hoch ist, dann reduziert sich auch das Set der möglichen Spender für diese Variablen. Liegen zudem noch wenige Fälle vor, könnte sich die Auswahl der Spender als problematisch gestalten, was wiederum in unplausiblen Imputationen resultieren könnte.

Festzuhalten ist demnach, dass es bei empirischen Daten, trotz sorgfältiger Auswahl und Anwendung einzelner MI-Varianten möglich ist, dass diese unplausible Imputationen generieren, weil die Annahmen einer jedweden Technik immer etwas verletzt werden; somit werden auch verzerrte Parameter und Standardfehler vorliegen. Welche Verletzungen der Annahmen einzelner MI-Varianten wirkmächtig werden und in unplausiblen Imputationen resultieren, kann aus diesen Ausführungen allerdings nicht abgeleitet werden. Zudem können auch nur die Skalierung der Variablen sowie deren Verteilungseigenschaften und für FCS oder PMM die Fallzahl als mögliche Einflussfaktoren identifiziert werden, jedoch nicht, ob auch bestimmte Anteile an Missing Values mögliche Einflussfaktoren darstellen. Um Hinweise zu erlangen, ob bei einer Verletzung der Annahmen tatsächlich mit unplausiblen Imputationen und verzerrten Schätzungen für die Parameter und Standardfehler zu rechnen ist, müssen MC-Studien herangezogen werden. Auf diese wird bei der Darstellung des Forschungsstandes in Kapitel 5 eingegangen. Zuvor werden im nächsten Kapitel noch Direct-ML und EM vorgestellt. Das ist notwendig, da Direct-ML zur Erreichung eines der Ziele der Arbeit benötigt wird und EM einen integralen Bestandteil von EMB darstellt.

4 Maximum Likelihood-Schätzverfahren bei fehlenden Werten

Wie die MI versuchen auch die ML-basierten modernen MDTs diejenigen Parameter θ zu finden, welche die Wahrscheinlichkeit maximieren, die beobachteten Daten Y zu erhalten. Das Ziel ist dementsprechend auch mit ML dasselbe: die Approximation von $P(\theta|Y)$. Während allerdings mit der MI die posterior-Verteilung der Parameter angestrebt wird, die vorgelegen hätte, wenn keine Missing Values vorhanden gewesen wären, versuchen die ML-Schätzverfahren diese Parameter durch die Maximierung der Likelihood-Funktion zu approximieren. Bei

fehlenden Werten kann diese Approximation durch Direct-ML oder EM erfolgen. Mit Direct-ML erfolgt die Behandlung der fehlenden Werte direkt bei der Parameterschätzung des Analysemodells, wohingegen die Behandlung der Missings mit EM von der eigentlichen statistischen Analyse losgelöst ist (wie auch bei der MI). Bei EM werden zunächst die fehlenden Werte berücksichtigt, bevor im zweiten Schritt die gewünschte Analyse durchgeführt wird. Im Grunde ist das zweistufige Verfahren mit EM äquivalent zu Direct-ML, denn Analysen, die auf ML-Schätzwerten basieren, ob sie durch Direct-ML oder EM generiert werden, werden sich nur gering voneinander unterscheiden. Aus diesem Grund wird EM auch als Indirect-ML[34] bezeichnet. Auch werden die Analyseergebnisse von Direct-ML und EM immer gleich sein, sofern auf denselben Daten und unter sonst gleichen Bedingungen, dieselbe Analyse mehrmals durchgeführt wird. Das liegt daran, dass sowohl Direct-ML als auch EM immer zu *einer* ML-Kovarianzmatrix und zu *einem* ML-Mittelwertvektor konvergieren von denen angenommen wird, dass sie denjenigen entsprechen, die ohne Missing Values beobachtet worden wären. Im Unterschied dazu unterscheiden sich die Analyseergebnisse mit der MI leicht voneinander, da die Imputationen auf Zufallsziehungen basieren (vgl. Enders 2010: 199; King u. a. 2001: 55). Zunächst wird Direct-ML vorgestellt, bevor auf EM eingegangen wird.

4.1 Direct-ML (FIML)

Direct-ML bzw. Full Information Maximum Likelihood (FIML) basiert auf derjenigen Schätzmethode, welche in der empirischen (Sozial-)Forschung und im Analysekontext der Strukturgleichungsmodellierung wohl am häufigsten eingesetzt wird: dem ML-Schätzverfahren. Dieses weist Eigenschaften auf, mit welchen Schätzwerte möglich sind, die asymptotisch konsistent (je größer das Sample ist, desto größer ist die Wahrscheinlichkeit, dass der geschätzte Parameter nicht vom Populationsparameter abweicht), asymptotisch effizient (in einem großen Sample ist die Varianz der Schätzwerte minimal) und asymptotisch normalverteilt sind, vorausgesetzt, die zugrundeliegenden Annahmen werden eingehalten. Darunter fallen ein metrisches Messniveau für die Daten, eine multivariate Normalverteilung der Daten, ein genügend großes Sample, ein korrekt spezifiziertes statistisches Analysemodell und die Unabhängigkeit der Beobachtungen (vgl. Finney/DiStefano 2013: 443). Während die letzten beiden Punkte oftmals als gegeben angenommen werden, sind die ersten drei häufig Ursachen für verzerrte Schätzungen der Parameter und Standardfehler.

[34] Sofern die Kovarianzmatrix und die Mittelwerte als Grundlage für die Modellanalyse dient (siehe Kapitel 4.2).

Bei einer ML-Schätzung werden die Populationsparameter gesucht, die dafür verantwortlich sind, die vorliegenden Daten zu erhalten. Dazu werden wiederholt Parameter geschätzt, bis eine Kombination gefunden wird, welche die Wahrscheinlichkeit maximiert, dass die Daten daraus hervorgegangen sind. Diese maximierte Wahrscheinlichkeit wird in einem Log-Likelihood-Wert ausgedrückt: der Sample-Log-Likelihood (S-LL). Sie zeigt die Abweichung der momentan geschätzten Parameter von den beobachteten Daten an. Ändert sie sich zwischen wiederholten Schätzungen nicht mehr, liegt eine Parameterkombination vor (θ), mit welcher es am wahrscheinlichsten ist, die vorliegenden Daten zu produzieren. Die ML-Schätzung ist konvergiert und es kann davon ausgegangen werden, dass die gefundene Parameterkombination des Analysemodells den Populationsparametern nahekommt/entspricht.

Die S-LL ist die Summe der individuellen LL-Werte der einzelnen Fälle im Datensatz. Hierin liegt auch der Unterschied zwischen einer normalen ML-Schätzung und Direct-ML. Denn eine normale ML-Schätzung benötigt zur Berechnung der fallweisen LL immer die gleiche Anzahl an Beobachtungen, während Direct-ML dabei alle Fälle miteinbeziehen kann, unabhängig davon, ob sie alle die gleiche Anzahl an gültigen Werten aufweisen. Damit werden nicht mehr alle Variablen zur Berechnung der LL notwendig, sondern nur noch diejenigen, die auch tatsächlich gültige Werte aufweisen. Durch die Berücksichtigung der Missing Values während der Berechnung der S-LL werden mit Direct-ML auch keine fehlenden Werte imputiert (siehe Arbuckle 1996; Enders 2001a; Enders/Bandalos 2001).

Soll bspw. ein Modell einer konfirmatorischen Faktorenanalyxse mit vier Indikatoren für einen Faktor (CFA-Modell) gerechnet weden, verlangt die ML-Schätzung, dass alle vier Indikatoren Beobachtungswerte für jeden Fall aufweisen. Es wird die LL pro Fall, basierend auf allen vier Variablen, berechnet. Bei Direct-ML hingegen kann die berechnete LL für den betreffenden Fall entweder nur auf einer, auf zwei, drei oder allen vier Variablen beruhen. Damit stehen im Umkehrschluss auch mehr Fälle für die statistische Analyse zur Verfügung als bei einer einfachen ML-Schätzung (sofern die Missing Values ausgeschlossen, anstatt imputiert wurden). Das kann sich im Ergebnis widerspiegeln: Mit Direct-ML sind effizientere Schätzungen möglich, da sie alle beobachtete Information in ihrer Schätzung berücksichtigen kann.[35]

Da Direct-ML nur eine andere Berechnung für die S-LL zugrunde legt, können ML und Direct-ML in Bezug auf den Schätzprozess und in Bezug auf die zugrundeliegenden Annahmen

[35] Für die Formeln zur Berechnung der LL und S-LL sei auf Enders (2010) verwiesen.

als äquivalent betrachtet werden. Damit besitzen die Schätzergebnisse von Direct-ML alle gewünschten Eigenschaften einer normalen ML-Schätzung. Auch sind unter MCAR und MAR[36] unverzerrte Parameter und Standardfehler zu erwarten. Zusätzlich ist Direct-ML in den meisten, für die Analyse von Strukturgleichungsmodellen verfügbaren Softwarepaketen implementiert (in M*plus* ist Direct-ML das Standardschätzverfahren, wenn Missing Values vorliegen).

Sollten die Annahmen allerdings nicht zutreffen, kann nicht mehr davon ausgegangen werden, dass die Schätzergebnisse die gewünschten Eigenschaften aufweisen. Stattdessen ist davon auszugehen, dass diese verzerrt sind und dass die daraus gezogenen Schlüsse falsch sein könnten. Vor allen Dingen trifft dies auf Verletzungen der Annahme einer multivariaten Normalverteilung zu.[37] Aus diesem Grund wurden, auch für Direct-ML, verschiedene Korrekturverfahren entwickelt, mit denen es möglich ist, unverzerrte Modellschätzungen zu generieren.[38] Zwar kann damit auf nicht normalverteilte Daten, auch wenn Missing Values vorliegen, reagiert werden, fraglich bleibt aber, wie lange diese Korrekturmaßnahmen wirksam sind. Denn während die Parameterschätzungen recht robust gegenüber der Verletzung der Normalverteilungsannahme sind (sofern diese Verletzungen nicht aufgrund diskreter Variablen zustande kommen), kann es dabei zu Verzerrungen in den Standardfehlern kommen (vgl. Finney/DiStefano 2013: 475 f.; Urban/Mayerl 2014: 142). Liegen im Kontext von Missing Values Variablen vor, die nur bedingt ein metrisches Messniveau aufweisen und zusätzlich noch schief verteilt sind, sind sehr wahrscheinlich auch die Parameterschätzungen davon betroffen. In solchen Fällen ist also nicht immer davon auszugehen, dass diese unverzerrt sind. Allerdings kommt es dann darauf an, wie viele Skalenpunkte diese Variablen aufweisen, und wie groß die Abweichung von der multivariaten Normalverteilung dadurch ist (vgl. Jia 2016: 85 ff.).

4.2 Expectation-Maximization (EM)

Neben Direct-ML liegt mit EM eine zweite Möglichkeit vor, ML-Schätzwerte zu generieren, wenn fehlende Werte vorliegen. EM ist ein zweistufiges Verfahren, das aus einem Expectation-

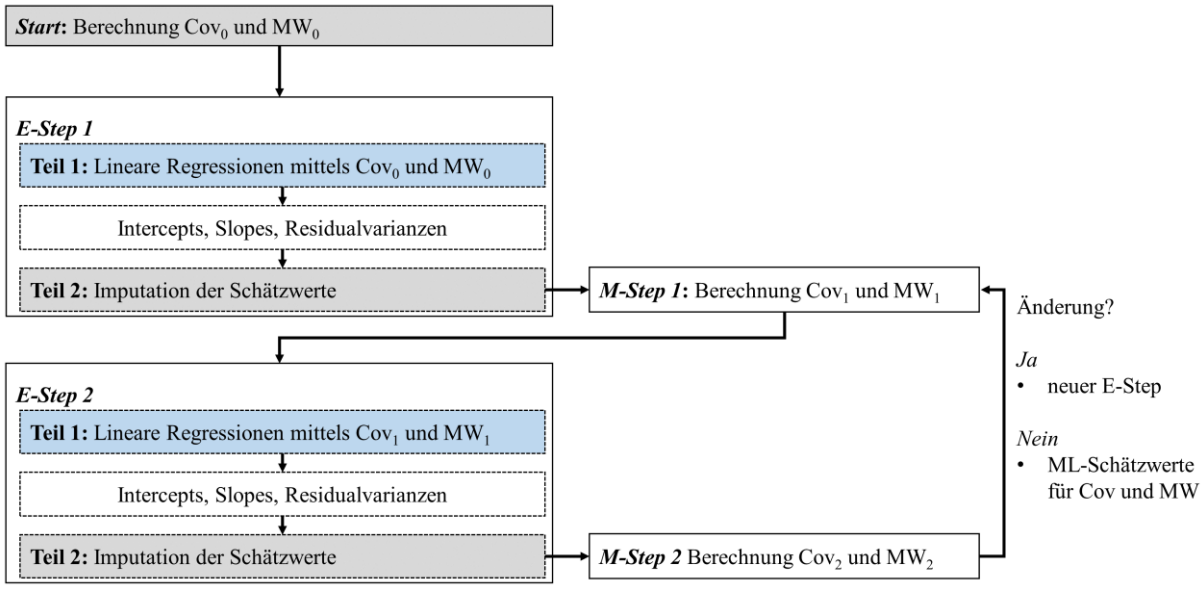
[36] Direct-ML hat gegenüber zweistufigen Verfahren den Nachteil, dass die Hinzunahme von Hilfsvariablen schwieriger ist, weil diese in das Analysemodell eingebunden werden müssen. Für die Berücksichtigung von Hilfsvariablen mit Direct-ML sei Graham (2003) empfohlen.

[37] Im Gegensatz zu Verletzungen der Normalverteilungsannahme stellen kleinere Fallzahlen eher keine Probleme für die ML-Schätzung dar (vgl. Anderson/Gerbing 1984: 171; Curran u. a. 1996: 25).

[38] Zum einen gibt es die Möglichkeit von Bootstrap-Verfahren und zum anderen von direkten Korrekturen während der Schätzung des Analysemodells. Beim Boostrapping werden zwei Verfahren notwendig. Das ist erstens, der naive Bootstrap für die Standardfehler und zweitens, der Bollen-Stine-Bootstrap für die Chi²-Statistik (vgl. Bentler 2006: 314). Bei den Korrekturverfahren werden dagegen die Standardfehler und die Chi²-Statistik direkt korrigiert (bspw. indem das Ausmaß der Verletzung der multivariaten Kurtosis berücksichtigt wird, wie es bei der Satorra-Bentler-Korrektur der Fall ist; vgl. Finney/DiStefano 2013: 443 f.).

Step (E-Step) und aus einem Maximization-Step (M-Step) besteht. Wie Direct-ML setzt auch EM eine multivariate Normalverteilung voraus, da nur die ML-Schätzwerte der Kovarianzmatrix und der Mittelwerte generiert werden, von denen ausgegangen wird, dass sie die Daten beschreiben können (θ; Dempster u. a. 1977). In der folgenden Abbildung 8 ist der Prozess zur Generierung der ML-Schätzwerte mit EM abgetragen, der zudem auch EMB zugrunde liegt. Die Ausführungen zur Funktion von EM basieren auf den Arbeiten von Allison (2002: 19 ff.) und Little/Rubin (2002: 167 ff.).

Abbildung 8: Expectation-Maximization (EM)



Vor dem ersten E-Step müssen dem Verfahren Startwerte zugewiesen werden (Start). Hierzu werden die fehlenden Werte entweder paar- oder fallweise ausgeschlossen. Daraus resultieren die Mittelwerte und eine erste Kovarianzmatrix (Cov_0 und MW_0). Aus diesen lassen sich dann im ersten E-Step die Parameter für die linearen Regressionen berechnen (Intercepts, Slopes und Residuen; E-Step, Teil 1) und es können Werte für die Missing Values geschätzt werden. Diesen Schätzwerten werden zufällig Residuen zugeordnet, um sie danach zu imputieren (E-Step 1, Teil 2). Mit den linearen Regressionen werden bei EM zwar fehlende Werte imputiert, allerdings werden diese nur dazu benötigt, um im anschließenden M-Step eine neue Kovarianzmatrix und neue Mittelwerte berechnen zu können: Im E-Step werden demnach nur Informationen, die, aufgrund von fehlenden Werten, in der Kovarianzmatrix und in den Mittelwerten im M-Step fehlen würden, durch *erwartete* (*expected*) Informationen substituiert. Am Ende des E-Steps liegt demnach zwar ein vollständiger Datensatz vor, dieser wird aber lediglich dazu benutzt, um eine neue Kovarianzmatrix (Cov_1) und neue Mittelwerte (MW_1) berechnen zu können

(M-Step 1). An dieser Stelle hat EM eine erste Iteration durchlaufen. Die zweite Iteration folgt anschließend mit einem erneuten E-Step (E-Step 2).

In dieser zweiten Iteration dienen die Kovarianzmatrix und die Mittelwerte aus dem vorherigen M-Step (Cov_1 und MW_1) als Grundlage für die Regressionsgleichungen. Diese unterscheiden sich nun zum vorherigen E-Step und lassen ‚bessere Vorhersagen' für die Imputationen zu. Nachdem der Datensatz erneut imputiert wurde, kommt es zu einem neuen M-Step, in welchem wieder neue Kovarianzen und Mittelwerte berechnet werden (M-Step 2). Sowohl Cov_2 als auch MW_2 dienen dann in der dritten Iteration dem erneuten E-Step als Grundlage für die Regressionsgleichungen usw. Der Wechsel zwischen E- und M-Step dauert so lange an, bis sich die Kovarianzmatrizen und Mittelwerte zwischen den Iterationen nicht mehr ändern, bzw. bis sich der daraus ergebende Wert für die Sample-Log-Likelihood (S-LL) nicht mehr signifikant ändert. Wenn sich keine Änderung mehr feststellen lässt, ist EM konvergiert und die S-LL ist maximiert. Die resultierende Kovarianzmatrix und die Mittelwerte sind dann ML-Schätzwerte. Auf deren Basis lassen sich dann die Analysen durchführen.

Diese Analysen weisen alle Eigenschaften einer ML-Schätzung auf und sowohl unter MCAR als auch MAR[39] können in der Regel unverzerrte Parameter und Standardfehler erwartet werden. Tatsächlich lässt sich das Verfahren auch relativ einfach einsetzen, da es in großen Statistikpaketen, wie SPSS, implementiert ist und für die Anwendung deshalb keine größeren Vorkenntnisse notwendig sind.

Wie mit Direct-ML können auch mit EM die Parameter und Standardfehler verzerrt sein, sollte eine Verletzung der multivariaten Normalverteilung vorliegen und/oder diese Annahme aufgrund der Daten nicht gerechtfertigt sein (bspw. bei nicht metrischen Variablen). Zusätzlich gibt es mit EM zwei weitere Probleme, je nachdem, ob die gelieferten ML-Schätzwerte (die Kovarianzmatrix und die Mittelwerte) als Analysegrundlage verwendet werden oder ob der imputierte Datensatz als Grundlage verwendet wird. Werden die ML-Schätzwerte direkt verwendet, stellt sich die Frage, welche Fallzahl für die Analysen angegeben werden soll. Denn Analysen, die auf Kovarianzen beruhen, setzen voraus, dass eine Fallzahl mitgeteilt wird. Für die maximierte Kovarianzmatrix aus einem EM-Prozess gibt es aber keine einheitliche Fallzahl (vgl. Graham/Schafer 1999: 3). Das kann dazu führen, dass die Standardfehler durch die Wahl

[39] Im Gegensatz zu Direct-ML ist der Einbezug von Hilfsvariablen mit EM einfacher. Hierzu nimmt man zusätzlich zu den Analysevariablen die Hilfsvariablen in den Schätzprozess mit auf. Auf Grundlage dieses Datensatzes wird dann EM durchgeführt. Aus der entstandenen Kovarianzmatrix und den Mittelwerten können dann die Informationen entnommen werden, die für das Analysemodell nötig sind. Die Analyse beinhaltet dann die zusätzlichen Informationen der Hilfsvariablen (vgl. Li 2010: 25).

der Fallzahl beeinflusst sind, was wiederum die inferenzstatistischen Schlüsse erschwert (ein Problem, das bei Direct-ML nicht gegeben ist). Deshalb wird angeraten zusätzlich Bootstrapping einzusetzen, um die Güte der Standardfehler unter der gewählten Fallzahl zu prüfen.

Dem ‚Fallzahlproblem' kann entgegengewirkt werden, wenn der imputierte Datensatz für die Analysen verwendet wird, denn die Fallzahl desselben ist bekannt (wie es bei EMB der Fall ist). Bei den statistischen Analysen kommt es dann aber, obwohl der Datensatz auf ML-Schätzwerten basiert, zu ähnlich negativen Folgen wie es bei Datensätzen der Fall ist, die mit Hilfe von Einfachimputationen behandelt werden. Denn bei einem EM-Datensatz weist der imputierte Wert keinerlei Unsicherheit mehr auf. Er würde als wahrer Wert begriffen werden, was zu Verzerrungen führen könnte. Während aber die Parameterschätzungen durchaus unverzerrt sein könnten, würden die Standardfehler unterschätzt werden (siehe Kapitel 3.1 und die Gründe warum *m* Datensätze, wie sie bei der MI benutzt werden, notwendig sind). Allerdings wird in jedem E-Step eine Korrektur durch die Residualvarianz angebracht; damit könnte diese Unterschätzung unproblematisch sein, sodass die Verzerrungen die inferenzstatistischen Schlüsse nicht beeinflussen (vgl. Allison 2009: 79; Peters/Enders 2002: 84).

4.3 Zusammenfassung: Einflussfaktoren auf die ML-Schätzverfahren

Wie für die MI gelten auch für die ML-Schätzverfahren einige Voraussetzungen:

1. Es muss zumindest ein MAR-Ausfallmechanismus realistisch sein.
2. Alle inhaltlich relevanten Variablen und Variablenbeziehungen müssen im IM (EM) oder im Analysemodell (Direct-ML) berücksichtigt werden (die Hinzunahme von Hilfsvariablen in das IM für EM oder das Analysemodell für Direct-ML ist optional, kann aber die Schätzung der Parameter und Standardfehler verbessern und MAR wahrscheinlicher werden lassen).
3. Erst wenn die Konvergenz sichergestellt ist, liegen ML-Schätzwerte vor.
4. Beide Verfahren setzen eine multivariate Normalverteilung der Daten und damit metrisch skalierte Variablen voraus sowie eine genügend große Fallzahl.

- Direct-ML setzt zudem ein korrekt spezifiziertes Analysemodell voraus.
- Für EM besteht die Problematik der Fallzahl: Einerseits ist nicht klar, welche Fallzahl den statistischen Analysen mitgeteilt werden soll, sofern die Kovarianzmatrix als Analysegrundlage dient. Andererseits könnte der imputierte Datensatz verwendet werden, um der Fallzahlproblematik zu entgehen. Beide Vorgehensweisen können aber die Schätzung der Standardfehler beeinflussen. Im letzteren Fall ist mit einer Unterschätzung derselben zu rechnen.

Aufgrund der Annahme einer multivariaten Normalverteilung gilt für die ML-Verfahren ähnliches wie für EMB oder MNV (und bisweilen für FCS und H0; siehe Kapitel 3.6), denn in empirischen Datensätzen ist diese Annahme meist nicht haltbar, da eben nicht nur metrisch skalierte Variablen vorliegen, sondern auch diskrete. Meist sind die Variablen auch nicht normalverteilt, sodass eine Verletzung der multivariaten Normalverteilung immer in irgendeiner Weise vorliegt. Zudem liegen für empirische Datensätze oftmals auch kleinere Fallzahlen vor, was ein weiterer Faktor für verzerrte ML-Schätzwerte sein könnte. Vor allen Dingen eine Kombination aus multivariat nicht normalverteilten Daten (sei es aufgrund von schiefen Verteilungen oder diskreten Variablen) und kleinen Fallzahlen könnte sich als problematisch herausstellen; dann dürften sich auch mit ML nicht mehr unbedingt unverzerrte Parameter und Standardfehler einstellen. Wie für die MI gilt auch hierbei, dass nicht bekannt ist, bzw. nicht abgeleitet werden kann, ab wann (oder ob) die Verletzungen der Annahmen zu Verzerrungen führen. Weiterhin stellt sich auch hier die Frage, ob bestimmte Anteile an Missing Values mögliche Einflussfaktoren darstellen. Demnach sind auch für die ML-basierten Schätzverfahren Hinweise aus MC-Studien heranzuziehen, die eine Einordung hinsichtlich dieser Einflussfaktoren zulassen. Im fünften Kapitel wird deshalb der Forschungsstand im Hinblick auf die ML-Verfahren und die MI-Varianten präsentiert.

5 Forschungsstand

Im Folgenden werden MC-Studien präsentiert, die sich zumindest mit einer der besprochenen MDTs auseinandersetzen. Das Forschungsdesign in den einzelnen Studien unterscheidet sich grob darin, ob synthetische oder empirische Daten untersucht werden und ob die Datenanalyse im Kontext der Strukturgleichungsmodellierung (SEM), also mit Strukturgleichungsmodellen (SE-Modellen), erfolgt oder nicht. Tabelle 1 gibt einen Überblick über die betrachteten Studien, ordnet diese gemäß den Unterschieden im Forschungsdesign ein und stellt dar, welche MDTs untersucht werden.[40]

[40] In diesen Studien werden nicht nur die hier beschriebenen MDTs getestet, sondern es liegen auch Ergebnisse zu anderen MDTs vor (bspw. zu den Ausschlussmethoden). Da diese aber nicht von Interesse für diese Arbeit sind, werden deren Ergebnisse nicht diskutiert. Im Grunde gibt es mehr Studien, die einzelne MDTs miteinander vergleichen. So haben Goldstein u. a. (2009) oder Goldstein u. a. (2014) die Eignung von eigens entwickelten Verfahren untersucht. Auch lassen sich Studien identifizieren, die selbst programmierte Schätzverfahren im Rahmen von FCS anwenden (Lee/Mitra 2016; van Ginkel/Kroonenberg 2017). Für Anwendende, die auf Softwarepakete angewiesen sind, haben solche Spezialvarianten allerdings eher wenig Relevanz, da sie nicht immer zur Verfügung stehen. Deshalb werden nur Studien dargestellt, die zumindest diejenigen MDTs testen, die auch für alle Forschende verfügbar sind. Auch finden sich in dieser Tabelle keine Studien wieder, die bereits in Kapitel 3.1 diskutiert wurden.

Tabelle 1: Überblick über betrachtete MC-Studien

	Autor(en)	Varianten der Multiple Imputation					ML-Verfahren	
		JM		FCS		H0	FIML	EM
		MNV	EMB		PMM			
Studien zu FCS	McNeish (2017)	x		x	x		x	
	Yu u. a. (2007)	x		x	x			
	Jia/Wu (2019)		x	x			x	
	Kropko u. a. (2013)	x	x	x				
	Lang/Wu (2017)			x				
	Lee/Carlin (2010)	x		x				
	Raghunathan u. a. (2001)	x		x				
	van Buuren u. a. (2006)			x				
	White/Carlin (2010)			x				
	Wu u. a. (2015)		x	x		x		
	Zhang u. a. (2017)	x		x				
	Kleinke (2017)	x			x			
	Kleinke (2018)				x			
	Pritikin u. a. (2018)				x		x	
Studien zu JM und ML	Honaker/King (2010)	x	x					
	Graham u. a. (1996)	x					x	x/+
	Newman (2003)	x					x	x
	Olinsky u. a. (2003)	x					x	+
	Finch (2010)	x						
	King u. a. (2001)	x						
	Leite/Beretvas (2010)	x						
	Lin (2010)	x						+
	Asparouhov/Muthén (2010c)					x	x	
	Chen u. a. (2018)						x	
	Enders (2001c)						x	
	Gold/Bentler (2000)						x	x
Studien zu den Fit-Indices	Ferro (2014)*	x					x	
	Li (2010)*	x					x	+
	Li/Lomax (2017)*	x					x	+
	Shin u. a. (2017)*	x					x	
	Wang (2007)*	x						x
	Enders (2001b)*						x	
	Enders/Bandalos (2001)*						x	
	Peters/Enders (2002)*						x	x
	Savalei/Bentler (2005)*						x	
	Savalei/Falk (2014)*						x	
	Teman (2012)*						x	
	Gold u. a. (2003)*							x
	Summe	**20**	**4**	**11**	**5**	**2**	**20**	**6 / 5**
		22		**14**				

Anmerkungen: MNV: *joint modeling* (JM) mit *data augmentation*; EMB: JM mit EM und Bootstrapping; FCS: *conditional modeling*; PMM: FCS mit *predictive mean matching*; H0: MI unter Berücksichtigung des Analysemodells; FIML: Direct-ML; EM: Expectation-Maximization.
Zu FCS: Für die Generierung der Imputationen können mehrere verschiedene Modellschätzungen eingesetzt werden. In der linken Spalte sind Studien vermerkt, die FCS-Spezifikationen untersuchen aber PMM dabei *nicht* berücksichtigen; die rechte Spalte enthält Studien die mit PMM als FCS-Spezifikation arbeiten.
Zu EM: Es gilt zwischen einer Variante, bei welcher die Analysen auf den ML-Schätzwerten basieren (x), und einer Variante, bei der die Analysen auf einem imputierten Datensatz basieren (+), zu unterscheiden.
Weitere: Autoren*: Bewertung der MDTs bzgl. der Fit-Indices. Weiß: MC-Studien mit synthetischen Daten im SEM-Kontext. Blau: MC-Studien mit synthetischen Daten im Rahmen anderer Analysekontexte. Grau: MC-Studien mit empirischen Daten unabhängig vom Analysekontext.
Zur Tabellensortierung: Studien zu FCS: sortiert nach FCS, dann PMM; Studien zu JM und ML sowie Studien zu den Fit-Indices: sortiert nach MNV, dann EMB, danach Direct-ML.

Der erste Teil der Tabelle 1 (Studien zu FCS) legt offen, dass FCS im Rahmen von SEM kaum evaluiert wird: Es lassen sich hierzu nur zwei Studien identifizieren, weshalb auch nur wenige Informationen gegeben sind, wie FCS bei der Schätzung der Parameter und Standardfehler zu bewerten ist (weiß). Zudem untersucht keine dieser beiden Studien die Performanz von FCS im Hinblick auf die Fit-Indices. Um zumindest Aussagen über die Zuverlässigkeit von FCS für die Schätzungen der Parameter und Standardfehler machen zu können, können auch MC-Studien aus anderen Analysekontexten oder mit empirischen Daten (blau und grau) herangezogen werden. Das Ergebnis daraus lässt zumindest darauf schließen, ob FCS tendenziell unverzerrte Parameter und Standardfehler liefert.[41] Damit liegen insgesamt 14 Studien vor, darunter fünf Studien, die PMM betrachten und elf Studien, die verschiedenste Spezifikationen untersuchen. Da für FCS die Möglichkeit besteht, verschiedene Modellschätzungen zur Imputation der Missings anzulegen, kann es sein, dass sich diese Anzahl von elf Studien schnell weiter verringert auf diejenigen mit ähnlicher bzw. gleicher FCS-Spezifikation, was die Informationsbasis bzgl. der Zuverlässigkeit für einzelne FCS-Spezifikationen weiter reduziert (Tabelle 2 in Kapitel 5.1).

Für JM liegen im SEM-Kontext mehr Informationen vor (weiß). Es können acht Studien (sieben MNV; eine EMB) identifiziert werden, in denen teilweise auch die Fit-Indices untersucht werden. Da es sich dabei aber nur um wenige Studien handelt (fünf Studien; Autoren mit einem Asterisk), bleibt auch dieser Informationsgehalt eher beschränkt, zumal für EMB keine solche Studie vorliegt. Wenn für JM noch Studien aus anderen Analysekontexten herangezogen werden, liegen insgesamt 22 Studien vor, wobei vor allem MNV untersucht wird; EMB wird dagegen eher vernachlässigt (20 MNV; vier EMB). Damit liegen für MNV potentiell genügend Informationen vor, um sie in Bezug auf die Parameter und Standardfehler bewerten zu können; für EMB dagegen nicht. Nur zwei Studien befassen sich mit H0 und nur eine davon bewertet die Technik im Zusammenhang mit SEM. Von allen MDTs liegen für H0 am wenigsten Informationen vor.

Eine ähnliche Anzahl wie für JM lässt sich auch für die ML-Verfahren, und insbesondere für Direct-ML beobachten. Für Direct-ML gibt es 16 Studien im Rahmen von SEM (weiß), wovon in neun die Fit-Indices untersucht werden. Damit darf Direct-ML als gut untersuchte MDT für diesen Kontext gelten (weniger dagegen für andere Analysekontexte). Anders sieht

[41] Das gilt für alle Varianten der MI. Denn die MI ist von den Analysemodellen losgelöst und sollte deshalb prinzipiell für viele statistische Analysen unverzerrte Parameter und Standardfehler liefern (in gewisser Weise gilt das auch für Direct-ML und EM, wobei bei ersterer eine Bewertung immer im Hinblick auf das jeweilige Analysemodell erfolgt). Liegen demnach in verschiedensten Analysekontexten unverzerrte Schätzungen der Parameter und Standardfehler vor, dann kann davon ausgegangen werden, dass dies für SE-Modelle auch der Fall sein könnte.

es wiederum für EM aus. Zwar liegen dafür insgesamt zehn Studien vor, da EM aber auf zweierlei Weise eingesetzt werden kann (als Imputationsvariante und als Indirect-ML), reduziert sich diese Anzahl auf fünf, respektive sechs Untersuchungen. Werden zudem nur die Studien im Bereich von SEM betrachtet, reduziert sich der Informationsgehalt weiter auf drei respektive vier Studien (weiß). Während die Parameter, unabhängig davon ob EM als Imputationsvariante oder als Indirect-ML verwendet wird, ähnlich sein werden, kann die jeweilige Variante die Schätzung der Standardfehler beeinflussen. Somit liegen für EM zwar genügend Informationen für die Parameterschätzung vor, nicht aber für die Standardfehler und für die Fit-Indices (fünf Studien; dreimal Indirect-ML, zwei für die Imputationsvariante).

Grundsätzlich gilt deshalb an dieser Stelle folgendes:

1. Nur für Direct-ML, EM und für MNV liegen genügend Studien vor, um die Leistung im Hinblick auf die Schätzung der Parameter im Rahmen von SEM (zum Teil auch in anderen Analysekontexten) und unter differenzierten Datensituationen bewerten zu können. Für die Standardfehler liegen dagegen nur für Direct-ML und MNV genügend Informationen vor, da diese bei EM von der gewählten Variante beeinflusst sein können.
2. Es gibt nur sehr wenige Informationen wie EMB, FCS, H0 und PMM im Hinblick auf die Parameter und Standardfehler zu bewerten sind. Das gilt vor allem für den SEM-Kontext, weniger für andere Analysekontexte (dafür stehen mehr Informationen zur Verfügung).
3. Für die Fit-Indices liegen für alle MDTs – mit Ausnahme von Direct-ML – keine (EMB, FCS, H0 und PMM) oder nur sehr wenige Informationen (EM, MNV) vor.

Weiterhin werden auch nur selten mehrere MI-Varianten zueinander untersucht (nur in zehn von 28 Studien, die sich mit der MI befassen; und nur in einer Studie im Zusammenhang mit SEM) oder in Relation zu den ML-Verfahren (13 Studien, davon zehn im SEM-Kontext). Von diesen zehn Studien werden in sieben Studien Vergleiche zu MNV angestellt, und nur jeweils einmal zu EMB, FCS, H0 oder PMM. Das hat zur Folge,

- dass erstens ungeklärt ist, wie die MI-Techniken unter gleichbleibenden Bedingungen zueinander abschneiden, und
- dass zweitens für die MI-Varianten, mit Ausnahme von MNV, keine Informationen vorliegen, wie diese zu den ML-Verfahren zu bewerten sind.

In den nachfolgenden Unterkapiteln werden die aufgelisteten Studien hinsichtlich ihrer Ergebnisse und ihrer Forschungsdesigns[42] vorgestellt. Beachtenswert ist, dass die Generalisierbarkeit von Ergebnissen aus MC-Studien eingeschränkt ist, weil die vorliegenden Ergebnisse vom gewählten Forschungsdesign abhängen. Aus diesem Grund interessiert im Folgenden lediglich, ob aus den verschiedenen Studien allgemeine Tendenzen in Bezug auf die Einflussgrößen aus Kapitel 3.6 und 4.3 ableitbar sind. Das wiederum sollte zeigen, welche der berichteten Einflussgrößen die Performanz der MDTs beeinflussen.

Zur Feststellung, ob die MDTs zuverlässig arbeiten, werden in den Studien verschiedene Prüfkriterien herangezogen. Dazu gehören: der (relative) Bias der geschätzten Parameter und Standardfehler, die Coverage des Konfidenzintervalls, die Effizienz/Genauigkeit der Parameterschätzungen sowie die prozentuale Rate an Modellablehnungen durch die Fit-Indices. Anhand von Abweichungen zwischen dem Populationsmodell oder den Analysen mit komplettem Datensatz und den Analysen nach Ersetzung der fehlenden Werte, kann festgestellt werden, wie gut die MDTs im jeweiligen Anwendungskontext arbeiten (siehe auch Kapitel 6.3). Wenn in den folgenden Unterkapiteln deshalb die Rede davon ist, dass die MDTs zufriedenstellend arbeiten und gute Ergebnisse vorliegen, dann bezieht sich dies darauf, dass die angesprochenen Prüfkriterien für die Parameter und Standardfehler den Schluss zulassen, dass diese unverzerrt geschätzt werden. Liegen gute Ergebnisse für die Fit-Indices vor, dann heißt das, dass die Ablehnungsrate der Modelle gering ist (ca. 5 %).

Das erste Unterkapitel befasst sich mit den FCS-Studien, das zweite mit den Studien zu JM und den ML-Verfahren, bei welchen keine Fit-Indices bewertet werden. Im dritten Unterkapitel werden die Studien zu den Fit-Indices herangezogen (Autoren mit Asterisk in Tabelle 1).

5.1 Bisherige Ergebnisse zu FCS

Das besondere an FCS ist, dass für einzelne Skalenniveaus unterschiedliche Möglichkeiten bestehen, um die Imputationen zu generieren. Damit ist die Performanz von FCS davon abhängig, wie FCS spezifiziert wird. Die Studienergebnisse zu FCS können aus diesem Grund nicht isoliert von der jeweiligen FCS-Spezifikation betrachtet werden, stattdessen hat dies im Hinblick auf die untersuchte Spezifikation zu erfolgen. Die untersuchten FCS-Spezifikationen lassen sich Tabelle 2 entnehmen. Für die Generierung der Imputationen auf binären Variablen werden

[42] Im Anhang finden sich drei Tabellen, welche die Forschungsdesigns der einzelnen Studien ausführlich zusammenfassen (Tabelle A1 bis Tabelle A3).

meist binär logistische Regressionen verwendet (zum Teil auch *random forest*[43] oder multinomial bzw. *renomalized* logistische Regressionen, wobei letztere eine Variation der multinomial logistischen Regression darstellt, die weniger Rechenzeit benötigt; vgl. Kropko u. a. 2013: 9). Missings auf ordinalen Variablen werden mit *proportional odds models*, multinomial logistischen Regressionen oder *machine learning*-Verfahren wie *random forests* und *classification and regressions trees* ersetzt, für metrische Variablen sind es lineare Regressionen oder PMM. Potentiell eignet sich PMM auch für binäre oder ordinale Variablen; das wird in den Studien aber nicht getestet.

Tabelle 2: Spezifikationen von FCS in den betrachteten Studien

Autor(en)	**Evaluierte FCS-Spezifikationen bei folgenden Skalenniveaus:**		
	binär	**ordinal**	**metrisch**
Pritikin u. a. (2018)	--	*proportional odds model*	PMM
Jia/Wu (2019)	binär log. Reg., *random forests*	multinomial log. Reg., *random forests*	--
Raghunathan u. a. (2001)	--	--	lin. Reg.
Zhang u. a. (2017)	--	*proportional odds model*	--
White/Carlin (2010)	--	--	lin. Reg.
Kropko u. a. (2013)	multinomial log. Reg., *renormalized* log. Reg.	multinomial log. Reg., *renormalized* log. Reg.	lin. Reg.
Wu u. a. (2015)	binär log. Reg.	multinomial log. Reg.	--
Lang/Wu (2017)	--	multinomial log. Reg., *classification and regressions trees*	--
McNeish (2017)	--	--	PMM, lin. Reg.
Kleinke (2017)	--	--	PMM
Kleinke (2018)	--	--	PMM
Yu u. a. (2007)	--	--	PMM, lin. Reg.
van Buuren u. a. (2006)	binär log. Reg.	multinomial log. Reg.	lin. Reg.
Lee/Carlin (2010)	binär log. Reg.	*proportional odds model*	--

Anmerkungen: log. Reg.: logistische Regression; lin. Reg.: lineare Regression.

Wie nun FCS (bzw. die FCS-Spezifikationen) beim Umgang mit Missing Values zu bewerten ist, erfolgt zum einen in Relation zu anderen MDTs (Direct-ML, EMB, H0 oder MNV; siehe Tabelle 1) und zum anderen unter verschiedensten Bedingungen. Diese Bedingungen werden auch als exogene Modellparameter oder als Simulationskonfigurationen bezeichnet. Von diesen Konfigurationen wird angenommen, dass sie die Güte der Imputationen mit FCS beeinflussen könnten; sie stellen die Einflussfaktoren dar. Als mögliche Einflussfaktoren werden verschiedene Variablenskalierungen (binär, ordinal und metrisch) und unterschiedliche Samplegrößen (min. 20, max. 10.000) herangezogen. Auch wird untersucht, inwiefern die Missing Values

[43] Sehr vereinfacht: *random forests* sind wie *classification and regressions trees* Klassifikationsverfahren die auf *machine learning*-Algorithmen basieren und durch selbstständige Entscheidungen, die in einer abhängigen Variablen vorhandenen Fälle anhand von Prädiktoren und von *cutpoints*, welche die Entropie am geringsten werden lassen, in Klassen einteilen. Entsprechend der Klassifikationen, welche diese Verfahren vornehmen, werden die fehlenden Werte imputiert (vgl. van Buuren 2018: 88 ff.). Für einen einführenden Überblick über solche Verfahren siehe: Ghani/Schierholz (2017).

selbst einen Einflussfaktor auf die Zuverlässigkeit von FCS darstellen, indem verschieden hohe Anteile an Missing Values (10 % bis 62,5 %) oder die Ausfallmechanismen (MCAR, MAR und NI) berücksichtigt werden. Zusätzlich wird geprüft, wie die Ergebnisse von FCS zu bewerten sind, wenn die Verteilungen der Variablen variiert werden, oder ob der Erfolg von FCS von verschiedenen Modellstrukturen abhängig ist: In zwei Studien werden CFA- bzw. SE-Modelle als Populationsmodelle verwendet (Jia/Wu 2019; Pritikin u. a. 2018), in den meisten Studien sind es aber Regressionsmodelle (logistische und lineare Regressionsmodelle). Vereinzelt werden auch Korrelationen (van Buuren u. a. 2006), Intercepts, Gruppen- und Zeiteffekte (Zhang u. a. 2017) oder deskriptive Statistiken (Yu u. a. 2007) untersucht (für detaillierte Angaben siehe die Tabellen A1 und A2 im Anhang).[44]

Werden die Ergebnisse dieser Studien zu FCS verallgemeinert und systematisiert, lassen sie sich wie folgt zusammenfassen:

- Die Performanz von FCS ist relativ unabhängig von den zugrundeliegenden Populationsmodellen der Simulationen und davon ob MCAR oder MAR gegeben ist (siehe Raghunathan u. a. 2001; Yu u. a. 2007; Zhang u. a. 2017); nicht aber von NI, denn dabei kommen keine zufriedenstellenden Ergebnisse zustande (McNeish 2017).
- Potentiell problematisch könnte sein, wenn ein Modell verschieden skalierte Variablen enthält und FCS mit verschiedenen Spezifikationen zur Ersetzung der Imputationen herangezogen wird (wenn also im Imputationsprozess für jede der Variablen ein jeweiliges Schätzverfahren angelegt wird) (Pritikin u. a. 2018).
- Liegt ein geringer Anteil an Missing Values vor, arbeitet FCS, unabhängig von der Variablenskalierung oder der gewählten Spezifikation, zufriedenstellend (siehe Verweise unter dem nachfolgenden Punkt). Ausgenommen davon sind die Varianten mit *ranfom forest* und *classification and regressions trees*, denn für beide Varianten lassen sich unter vielen Konfigurationen keine zufriedenstellenden Ergebnisse beobachten (Jia/Wu 2019; Lang/Wu 2017).
- Mit zunehmendem Missinganteil lassen sich mit FCS weiterhin gute Ergebnisse erwarten, wenn die Verteilungen der Variablen normal bzw. symmetrisch normal[45]

[44] Die Studien unterscheiden sich weiterhin darin, wie viele Variablen sie verwenden und welche Zusammenhänge sie für die Variablen vorgeben. Anzumerken ist, dass diese beiden Variationsfaktoren eher keinen Einfluss auf die Güte der Schätzung der Parameter und Standardfehler nehmen.

[45] Symmetrisch normal: Die Fälle streuen symmetrisch um den Skalenmittelpunkt und nähern sich einer Normalverteilung an. Der Skalenmittelpunkt weist dabei die meisten Fälle auf, die Kategorien links und rechts davon weisen mit jedem, sich von der Mitte entfernenden, zusätzlichen Skalenpunkt dagegen immer weniger Fälle auf.

sind oder eine größere Fallzahl vorliegt (Kropko u. a. 2013; Lang/Wu 2017; Lee/Carlin 2010; Raghunathan u. a. 2001; van Buuren u. a. 2006; White/Carlin 2010).

- Nimmt der Anteil an Missing Values zu und liegen kleine Fallzahlen und/oder (stark) asymmetrische oder nicht mehr normale Verteilungen der Variablen vor, dann bekundet FCS Probleme. Besonders betroffen erscheinen dabei die Spezifikationen mit den logistischen Regressionsmodellen (Jia/Wu 2019; McNeish 2017; Wu u. a. 2015).
- PMM ist recht robust gegenüber Verletzungen der Normalverteilungsannahme. Sofern die Größe des Spendersets berücksichtigt wird (drei bis fünf Spender) bekundet PMM erst Probleme, wenn zu den Verletzungen der Normalverteilungsannahme sehr hohe Missinganteile und kleine Fallzahlen (< 100) hinzukommen; wobei PMM auch dann bessere Ergebnisse liefert als alternative MDTs (Kleinke 2017; 2018; McNeish 2017; Yu u. a. 2007).
- Dass sich FCS wegen der flexiblen Spezifikation der IMs grundsätzlich besser eignet als alternative MDTs (Direct-ML, EMB oder MNV) lässt sich diesen Studien nicht entnehmen. Oftmals schneiden diese sogar besser ab; auch mit ihnen können in den meisten Fällen zufriedenstellende Ergebnisse erwartet werden (die Erkenntnisse für diese Alternativen werden in Kapitel 5.2 diskutiert).

5.2 Bisherige Ergebnisse zu JM und den ML-Verfahren

Im Gegensatz zum vorherigen Kapitel werden hier die Ergebnisse zu JM und ML vorgestellt.[46] Grundsätzlich ändern sich die Simulationskonfigurationen zu denen aus Kapitel 5.1 kaum. Anders als für FCS können für JM und ML aber noch Studien identifiziert werden (Studien zu JM und ML in Tabelle 1), die quasi-metrische Variablen berücksichtigen (drei bis fünf Kategorien) (Leite/Beretvas 2010; Lin 2010), wenngleich auch bei diesen zusätzlichen Studien nur selten eine Variation der Verteilungen vorgenommen wird (Gold/Bentler 2000; Leite/Beretvas 2010). Zudem werden auch nur vereinzelt andere Populationsmodelle herangezogen (wie ein Latent Growth Model oder die Nähe der Imputationen zu den wahren Werten; Chen u. a. 2018 bzw. Lin 2010).[47]

Direct-ML, EM, EMB und MNV (zum Teil auch H0; beruhend auf wenigen Informationen) zeigen sich unter MCAR oder MAR, unabhängig von den getesteten Modellen, robust gegen-

[46] Das betrifft auch die Ergebnisse für diese MDTs aus den besprochenen Studien in Kapitel 5.1.
[47] Siehe auch Tabelle A2 für eine detaillierte Auflistung der Forschungsdesigns.

über den simulierten Bedingungen und arbeiten unter verschiedensten Konfigurationen zufriedenstellend. Schlussendlich lassen sich den Studien folgende Ergebnisse entnehmen:

- Liegen NI-Missings vor, sollten alle nur bis zu einem gewissen Anteil an Missings eingesetzt werden. Grundsätzlich liegen bei NI immer etwaige Verzerrungen vor; Direct-ML hebt sich dabei etwas ab, da damit zum Teil weniger verzerrte Ergebnisse vorliegen als bei Spezialmethoden[48] (Chen u. a. 2018; Newman 2003).
- Ist die Fallzahl groß genug und liegen metrische Variablen vor (ob normalverteilt oder nicht), unterscheiden sich die Ergebnisse von Direct-ML, EM, EMB und MNV (zum Teil auch H0) kaum voneinander; sie liefern recht ähnliche Ergebnisse. Zwar wird die Performanz dieser MDTs mit zunehmendem Missinganteil etwas beeinträchtigt, in vielen Fällen ist aber weiterhin mit guten Ergebnissen zu rechnen (Asparouhov/Muthén 2010c; Enders 2001c; Gold/Bentler 2000; Graham u. a. 1996; Honaker/King 2010; King u. a. 2001; Newman 2003; Olinsky u. a. 2003).
- Für EM liegen in Bezug auf die Standardfehler etwas widersprüchliche Ergebnisse vor. Einerseits zeigt sich, dass bei höheren Missinganteilen die Standardfehler nicht mehr zufriedenstellend geschätzt werden (das gilt für beide EM-Varianten: Li 2010; Li/Lomax 2017; Newman 2003), andererseits lassen sich aber auch zufriedenstellende Ergebnisse beobachten: Gold u. a. (2003) und Olinsky u. a. (2003).
- Ist die Fallzahl etwas kleiner und liegt ein sehr hoher Missinganteil und/oder nicht normal verteilte Variablen vor, dann liegen größere (aber meistens noch vertretbare) Verzerrungen vor; in einem solchen Fall schneiden EMB und MNV etwas schlechter ab als Direct-ML und EM (Kropko u. a. 2013; Olinsky u. a. 2003).
- Gute Ergebnisse lassen sich mit Direct-ML, EMB, MNV und teilweise mit H0 erwarten, wenn nicht metrisch skalierte Variablen vorliegen (binäre oder ordinale) und/oder deren Verteilung (stark) asymmetrisch ist, sofern die Fallzahl groß genug oder der Missinganteil nicht zu hoch ist (Asparouhov/Muthén 2010c; Finch 2010; Jia/Wu 2019; Leite/Beretvas 2010; Lin 2010; Pritikin u. a. 2018; Wu u. a. 2015).

Letztendlich zeigen die Ausführungen aus diesem und dem vorangegangenen Kapitel, dass Verletzungen der zugrundeliegenden Annahmen der MDTs nur bedingt zu problematischen Ergebnissen führen und in vielen Fällen gute Ergebnisse erwartet werden können. Klare Tendenzen, wonach sich die Performanz der MDTs verschlechtert, wenn die Annahmen verletzt

[48] Wie das zusätzlich getestete Diggle-Kenward Selection Model und ein Pattern Mixture Model.

sind, lassen sich aber dennoch ausmachen. Weiterhin zeigt sich, dass die Performanz der MDTs, neben den identifizierten Einflussgrößen aus der jeweiligen Verfahrenslogik der einzelnen MDTs (siehe Kapitel 3.6 und 4.3), auch durch die Missinganteile und durch Kombinationen aus den Einflussgrößen beeinflusst ist.

5.3 Bisherige Ergebnisse zu den Fit-Indices

In diesem Abschnitt werden Studien vorgestellt, die eine Evaluation der Performanz hinsichtlich der Fit-Indices durchführen. Vor allen Dingen unterscheiden sich diese Studien zu denen aus Kapitel 5.1 und 5.2 darin, dass hierbei nur CFA- und SE-Modelle evaluiert werden (in zwei Studien stehen Latent Growth Models im Fokus; Ferro 2014 und Shin u. a. 2017). Im Großen und Ganzen entsprechen die Forschungsdesigns hinsichtlich der Fallzahlen und der Missinganteile (MCAR und MAR bis max. 30 %) den vorherigen Studien. Mit einer Ausnahme (Teman 2012) werden ausschließlich metrische Variablen untersucht, wobei im Großteil der Studien eine Variation der Variablenverteilung vorgenommen wird. Evaluiert werden in allen Studien die Chi²-Statistik und der dazugehörige p-Wert (im Weiteren nur p-Wert). Andere Fit-Indices werden dagegen kaum untersucht. Li (2010) und Li/Lomax (2017) evaluieren noch den *Root Mean-Square Error of Approximation* (RMSEA); Ferro (2014) sowie Teman (2012) berücksichtigen zusätzlich dazu noch den *Tucker Lewis Index* (TLI), den *Comparative Fit Index* (CFI) und das *Standardized Root Mean-Square Residual* (SRMR; nicht bei Ferro 2014).[49]

Die Bewertung der CFA- und SE-Modelle mittels dem p-Wert, nachdem Direct-ML, EM oder MNV zur Handhabung der fehlenden Werte eingesetzt wurden, erscheint in vielen Fällen unproblematisch. Oftmals liegt die Rate an Modellablehnungen bei ca. 5 % (unabhängig davon ob MCAR oder MAR vorliegt) (Enders/Bandalos 2001; Gold u. a. 2003; Peters/Enders 2002; Wang 2007). Tendenziell ist die Rate an Modellablehnungen mittels dem p-Wert höher, wenn bei konstantem Missinganteil zum einen kleine Fallzahlen vorliegen und zum anderen die Normalverteilungsannahme verletzt wird (Enders 2001b; Savalei/Falk 2014). In einem großen Teil der Studien zeigt sich Direct-ML recht robust, was diese Punkte betrifft. Zwar gehen mit Direct-ML dann auch etwas höhere Ablehnungsraten einher, diese liegen aber, wenn denn ein Vergleich zu anderen MDTs gegeben ist, zumeist unter denen der anderen MDTs. Das wiederum impliziert, dass Direct-ML etwas besser abschneidet, auch wenn nicht immer die gewünschten Ablehnungsraten erreicht werden (Savalei/Bentler 2005; Shin u. a. 2017).

[49] Siehe Tabelle A3 im Anhang. Auch lassen sich den Studien keine anderen Ergebnisse für die Schätzung der Parameter und der Standardfehler entnehmen, als diejenigen aus Kapitel 5.2.

Anders als für den p-Wert, liegen nur wenige Studien für andere Fit-Indices vor. Für RMSEA zeigen Li (2010) und Li/Lomax (2017), dass dieser Index vor allem dann geeignet ist, wenn größere Fallzahlen gegeben sind. Bei konstantem Missinganteil erhöht sich die Ablehnung der Modelle mittels RMSEA erheblich, wenn die Fallzahl kleiner wird. Bei höheren Quoten an fehlenden Werten und kleinen Fallzahlen, können keine zufriedenstellenden Ergebnisse mit RMSEA mehr beobachtet werden. Dagegen zeigt Ferro (2014), dass mit Direct-ML und MNV unter allen Bedingungen zufriedenstellende Ergebnisse erwartet werden können und dass beide Techniken adäquate Ablehnungsraten mit den getesteten Fit-Indices generieren. Teman (2012) weist zudem gute Ergebnisse mit diesen Fit-Indices nach, wenn Direct-ML bei ordinalen Variablen eingesetzt wird (er testet aber keine der hier aufgeführten MI-Varianten[50]). Oftmals zeigen sich gute Werte für die Ablehnungsraten. Es zeigt sich aber auch, dass die Ablehnung der Modelle zunimmt, wenn bei gleichbleibendem Anteil an Missing Values die Fallzahl abnimmt und/oder asymmetrische Verteilungen vorliegen.

5.4 Desiderate in der Forschung

In den vorangegangenen Kapiteln wurden verschiedenste Ergebnisse aus MC-Studien im Hinblick auf die Handhabung von fehlenden Werten zusammengetragen. Gleichzeitig zeigt die Vorstellung des Forschungsstandes für FCS Lücken auf, da diese Technik entweder nur in wenigen Studien evaluiert wurde, oder das Forschungsdesign dieser Studien bestimmten Punkten nur wenig Beachtung schenkte. Einige dieser Desiderate wurden bereits in der Einleitung[51] dieser Arbeit aufgegriffen. An dieser Stelle können sie ausdifferenziert und durch weitere Desiderate im Kontext der Forschung zu Missing Values mittels der MI ergänzt werden. Sie definieren letztlich die Forschungslücke, welche diese Arbeit zu füllen versucht und tragen zur Verdeutlichung der Relevanz der Arbeit bei.

Erstens wird in keiner der berichteten Studien FCS bei Variablen mit einem quasi-metrischen Skalenniveau untersucht (ob in der Spezifikation mit linearen Regressionen oder mit PMM) und *zweitens* wird FCS meist nur bei Daten untersucht, deren Variablen einheitliche Skalierungen aufweisen (entweder metrische, oder ordinale, oder binäre, aber keine gemischten Daten[52]). Da *drittens* keine Studie mit quasi-metrischen Variablen existiert, kann auch nicht

[50] Stattdessen untersucht er eine MI-Variante mit einem multivariaten probit-Modell und weist hinsichtlich der Fit-Indices nach, dass die Ablehnung der Modelle steigt, wenn bei konstantem Missinganteil eine geringe Fallzahl und/oder asymmetrische Verteilungen vorliegen. Weiterhin zeigt er, dass bei 25 % Missing Values die Rate der Modellablehnung zum Teil sehr hoch ist und Direct-ML besser abschneidet als seine getestete MI-Variante.
[51] Das betrifft erstens, drittens und fünftes. Die anderen Punkte sind neu und stellen die ergänzten Desiderate dar.
[52] ‚Gemischte Daten': zu verstehen als Daten, die Variablen mit unterschiedlichen Skalenniveaus enthalten.

bewertet werden, wie FCS (ob in der Spezifikation mit linearen Regressionen oder mit PMM) abschneidet, wenn die Variablenverteilungen nicht mehr symmetrisch normal, sondern (stark) asymmetrisch sind. Diese drei Punkte sind für die empirische Sozialforschung allerdings von größerer Bedeutung, da sowohl quasi-metrisch skalierte Variablen, als auch Datensätze mit gemischt skalierten Variablen dabei häufig als Grundlage zur statistischen Analyse herangezogen werden. Eine Bewertung, ob FCS bei solchen Gegebenheiten plausible Imputationen generiert, ist demnach überfällig. Zusätzlich zu diesen Punkten liegt *viertens* für die FCS-Spezifikation mit PMM keine Evaluation vor, wie sie sich bei Variablen mit wenigen Skalenpunkten bewährt (binäre oder ordinale Variablen); deshalb kann auch nicht bewertet werden, wie PMM bei (stark) asymmetrisch verteilten binär oder ordinal skalierten Variablen abschneidet. Da in der Literatur davon gesprochen wird, dass sich PMM potentiell für alle Variablenskalierungen eignen könnte und somit bei verschiedensten Datensituationen einsetzbar wäre, dafür allerdings keine Studie vorliegt, welche das für empirische Verteilungseigenschaften testet, ist auch hierbei einiges an Forschungsbedarf vorhanden. Es liegen *fünftens* kaum Erkenntnisse vor, wie FCS (unabhängig von den möglichen Spezifikationen) im SEM-Kontext zu bewerten ist und es gibt keine Ergebnisse dazu, wie FCS (unabhängig von den möglichen Spezifikationen) im Hinblick auf die Fit-Indices dieses Analysekontextes abschneidet. Da diese dabei aber einen integralen Bestandteil der statistischen Analyse darstellen, führt die Nicht-Berücksichtigung derselben letztlich dazu, dass nicht klar ist, ob sich FCS überhaupt für diesen Analysekontext eignet.

Außerdem lassen sich *sechstens* nur wenige Studien identifizieren, die sich mit alternativen MI-Varianten wie EMB oder H0 (unabhängig vom Analysekontext) auseinandersetzen. Auch für diese Techniken gibt es damit nur wenige Ergebnisse bei möglichst differenzierten Simulationskonfigurationen (bei quasi-metrischen Variablen, gemischten Daten oder auch variierten Verteilungen) und zudem kaum Ergebnisse für ihre Eignung im Zusammenhang mit SEM und demnach auch keine Ergebnisse zu den Fit-Indices. Damit gilt auch für diese MI-Varianten, dass nicht klar ist, wie sie bei quasi-metrischen Variablen oder Daten mit gemischt skalierten Variablen (ob diese nun symmetrisch normal oder asymmetrisch verteilt sind) zu bewerten sind und ob sie sich, aufgrund der fehlenden Evaluation der Fit-Indices, für die statistische Analyse mittels SE-Modellen eignen. Es ist deshalb nicht nur eine Evaluation von FCS und PMM unter verschiedensten Bedingungen zu leisten, sondern auch eine Evaluation von EMB und H0. Weiterhin ist *siebtens* anzuführen, dass, obwohl es Studien gibt, die sich mit MNV und den Fit-Indices auseinandersetzen, keine Studie existiert, die sich mit MNV und den populären Fit-Indices (wie RMSEA, dem CFI oder dem SRMR) bei vorliegenden quasi-metrischen Variablen

oder Daten mit gemischten Variablenskalierungen befassen. Auch hierzu ist demnach Forschungsbedarf vorhanden.

Wie an diesen Ausführungen zu erkennen ist, werden vor allem die Fit-Indices in der bisherigen Forschung vernachlässigt. Das kann zwei Gründe haben: Zum einen ist anzuführen, dass die MI zuvorderst dafür da ist, trotz fehlender Werte, Aussagen über die zugrundeliegende Population zuzulassen (was bei unverzerrten Schätzungen der Parameter und Standardfehler möglich wird), wohingegen die Fit-Indices nicht von vorrangigem Interesse sind. Das zeigt sich zum anderen auch darin, dass keine Richtlinien vorliegen, wie mit den Fit-Indices von multipel imputierten Daten umgegangen werden soll, wie es sie für die Parameter und Standardfehler gibt (*Rubin´s rules*). Die Anwendenden müssen demnach ohne Richtlinie arbeiten und sich bei der Zusammenführung der *m* Werte für die Fit-Indices auf das Analyseprogramm verlassen (meist wird der Durchschnitt der Fit-Indices gebildet; M*plu*s bringt bei der Chi²-Statistik zusätzlich eine Korrektur an, welche die Anzahl der Datensätze berücksichtigt; vgl. Asparouhov/Muthén 2010b: 2 f.). Ob allerdings die Zusammenführung der Fit-Indices durch die Analyseprogramme zu den gewünschten Ergebnissen führt, ist ungeklärt. Eine Möglichkeit, um dies zu überprüfen, wäre zu vergleichen, wie die Fit-Indices von multipel imputierten Daten im Vergleich zu einfach imputierten Daten zu bewerten sind und ob die Verfahrensweise des Analyseprogramms tendenziell schlechtere oder bessere Ergebnisse liefert. Li (2010) und Li/Lomax (2017) machen im Falle von RMSEA einen solchen Vergleich und zeigen, dass für die Analysen auf den mit MNV imputierten *m* Datensätzen und für den einfach imputierten Datensatz mittels EM, ähnliche Ergebnisse vorliegen. Das wiederum würde bedeuten, dass die Zusammenführung der *m* Werte der Fit-Indices kein vorrangiges Problem darstellt. Eine Evaluation, ob dies auch bei anderen Fit-Indices und/oder bei den oben genannten Datensituationen der Fall ist, oder ob dies auch für die anderen MI-Varianten gegeben ist, ist ungeklärt (*achtens*).

Schlussendlich können einige Studien identifiziert werden, die sich mit verschiedensten MDTs befassen. Allerdings gibt es nur wenige Studien die verschiedene MI-Varianten zueinander vergleichen (vor allem im Kontext von SEM). Gleichzeitig fehlen auch Vergleiche, wie sich verschiedenste MI-Varianten zu Direct-ML oder EM verhalten (mit Ausnahme von MNV). Es bleibt damit unklar (*neuntens*), welche der MI-Varianten sich wann eignet, wann eine Variante einer anderen vorzuziehen ist, oder welche der vorgestellten Varianten sich am ehesten zur Behandlung der fehlenden Werte eignet und *zehntens*, wie diese MI-Varianten gegenüber den ML-Verfahren abschneiden.

Aus der Darstellung der Forschungslücke leiten sich die notwendigen Bedingungen für das Forschungsdesign dieser Arbeit ab. Dieses muss einige der aufgelisteten Punkte inkorporieren, um einen Teil zur Füllung derselben beitragen zu können, um die in der Einleitung formulierten Fragen beantworten zu können und um die Ziele dieser Arbeit erreichen zu können. Zusätzlich zu den Desideraten konnten in den vorangegangenen Kapiteln auch Einflussfaktoren auf die Performanz der MDTs identifiziert werden. Diese Ergebnisse, sowie die abgeleiteten Einflussfaktoren aus den Kapiteln 3.6 und 4.3, dienen als Grundlage für die zu testenden Hypothesen über das Abschneiden der MDTs unter empirischen Verteilungseigenschaften.

5.5 Ableitung der Hypothesen

An dieser Stelle gilt es anzuführen, dass die Hypothesen im Hinblick auf die Tatsache formuliert werden, dass bestimmte Randbedingungen hingenommen werden. Dazu gehören, dass ein MAR-Ausfallmechanismus vorliegt, dass die Konvergenz der MDTs sichergestellt ist, dass für die MI genügend Datensätze (m) imputiert werden und diese nicht miteinander korrelieren, dass das Analysemodell mit dem IM übereinstimmt (im Hinblick auf die Anzahl an Variablen), dass die Modellstruktur des IMs mit dem Analysemodell übereinstimmt (H0), dass die Analysemodelle korrekt spezifiziert sind (Direct-ML) und dass für PMM ein konstantes Spenderset vorliegt. Weiterhin wurde gezeigt, dass es für diese Arbeit notwendig ist eine Einfachimputation als Vergleichsparameter für die Fit-Indices heranzuziehen; EM soll folglich als Imputationsvariante verwendet werden (die Hypothesen beziehen sich auf diese Variante).

Sowohl die ML-Verfahren, als auch alle MI-Varianten (mit Ausnahme von PMM) gehen davon aus, dass entweder eine gemeinsame multivariate Normalverteilung der Daten vorliegt, oder dass die als metrisch definierten Variablen (quasi-metrische Variablen) normalverteilt sind. Wenn demnach diese Annahme eingehalten wird, dann kann davon ausgegangen werden, dass alle MDTs in etwa dieselben Ergebnisse erbringen. Das zeigt sich auch in der bisherigen Forschung: Liegen metrische Variablen vor, die normalverteilt sind, dann ergeben sich für alle berichteten MDTs kaum verzerrte Schätzungen für die Parameter und die Standardfehler. Zwar liegen kaum MC-Studien vor, die sich mit quasi-metrischen Variablen auseinandersetzen, aber aufgrund der Tatsache, dass einerseits Konsens besteht, wonach Methoden, die metrisch skalierte Variablen voraussetzen, auch bei quasi-metrischen Variablen unverzerrte Ergebnisse generieren, und andererseits für MNV dabei gute Ergebnisse vorliegen, dürfen auch für diejenigen MDTs, welche ebenfalls die Annahme einer multivariaten Normalverteilung treffen, bei quasi-metrischen Variablen unverzerrte Ergebnisse erwartet werden. Da zudem FCS, H0 und PMM lineare Regressionen zur Imputation, respektive zur Identifikation der Spender verwenden,

sollte dies auch für diese drei gelten. Demnach gilt: *Weisen die Variablen ein (quasi-)metrisches Messniveau auf und sind (symmetrisch) normalverteilt, dann dürften die Ergebnisse der MDTs annähernd identisch und unverzerrt sein (Hypothese 1).*

Ausgehend davon kann angenommen werden, dass bei Verletzungen der (multivariaten) Normalverteilungsannahme mit verzerrten Ergebnissen zu rechnen ist. In den MC-Studien zeigt sich aber, dass die einzelnen MDTs durchaus robust gegenüber solchen Verletzungen sein können und die Verzerrungen oftmals unproblematisch bleiben (in Einzelfällen kann es aber dennoch zu erheblich verzerrten Ergebnissen kommen). Nichtsdestotrotz ist aber eine Tendenz erkennbar, die zeigt, dass mit zunehmender Abweichung von der (multivariaten) Normalverteilungsannahme auch die Verzerrungen zunehmen. Weil für FCS mittels linearer Regressionen keine Ergebnisse vorliegen, bleibt auch unbekannt, ob deren Performanz bei nicht normalverteilten (quasi-)metrischen Variablen beeinflusst wird. Aufgrund der Tatsache, dass auch die anderen MI-Varianten lineare Regressionen zur Ersetzung der Missings auf (quasi-)metrischen Variablen heranziehen und sich dabei zeigt, dass mit zunehmenden Verletzungen der Normalverteilungsannahme auch die Ergebnisse verzerrter sind (auch wenn die Verzerrungen meist unproblematisch bleiben), sollte dies auch für FCS der Fall sein. Eine Ausnahme stellt PMM dar, da PMM auf tatsächlich beobachtete Daten zurückgreift und die linearen Regressionen lediglich dafür verwendet, möglichst ähnliche, tatsächlich beobachtete Werte zu finden. Damit können keine Werte vorliegen, die nicht auch beobachtet werden. Das führt dazu, dass die Verteilung der Daten recht gut erhalten bleibt und dass PMM eher nicht durch Verletzungen der Normalverteilungsannahme beeinflusst ist. Es lässt sich folgende Hypothese formulieren: *Verletzen die (quasi-)metrisch skalierten Variablen die (multivariate) Normalverteilungsannahme, dann sind die Schätzergebnisse, mit Ausnahme für PMM, stärker verzerrt (Hypothese 2*).

Weil für statistische Analysen meist nicht nur (quasi-)metrische Variablen, sondern auch diskrete Variablen benutzt werden, stellt sich die Frage, wie die einzelnen MDTs abschneiden, wenn Imputationen auf gemischten Daten vorgenommen werden. Ergebnisse aus MC-Studien zeigen, dass Direct-ML, EMB oder MNV auch mit diskreten Variablenskalierungen durchaus zurechtkommen, obwohl diese MDTs eigentlich nicht für solche Skalierungen prädestiniert sind. Einzuschränken ist allerdings, dass bisher nur Untersuchungen vorliegen, die entweder die eine (binäre) oder die andere (ordinale) Skalierung der Variablen berücksichtigen. Wie diese MDTs aber zu bewerten sind, wenn Variablen mit unterschiedlichen Skalenniveaus zugleich behandelt werden, wurde bislang nicht erforscht. Aufgrund der Tatsache, dass für diese MDTs

quasi-metrische Variablen keine Probleme darstellen sollten und aufgrund dessen, dass sie zudem diskrete Variablen handhaben können, ist davon auszugehen, dass sie auch mit gemischten Daten zurande kommen sollten. Da für EM keine Befunde vorliegen was diskrete Variablen betrifft, EM aber wie Direct-ML, ML-Schätzwerte generiert und diese sich für gemischte Daten eignen sollte, kann auch für EM davon ausgegangen werden, dass kaum Verzerrungen vorliegen werden. H0 sollte auch bei gemischten Daten keine Probleme bekunden, da diese die verschiedenen Skalenniveaus explizit berücksichtigen kann. Für FCS kann es problematisch sein, wenn verschiedene Schätzverfahren für die Missings im IM angewendet werden, da damit die Gefahr besteht, dass FCS zu keiner gemeinsamen Verteilung konvergiert. Letztlich zeigt sich in einer der MC-Studien (Pritikin u. a. 2018), dass dann keine zufriedenstellenden Ergebnisse produziert werden. Allerdings führt die Literatur zu FCS aus, dass FCS nicht unbedingt zu einer gemeinsamen Verteilung konvergieren muss, um plausible Imputationen zu generieren; das wiederum widerspricht den Ergebnissen dieser MC-Studie. Da FCS für verschiedenste Variablenskalierungen zufriedenstellende Ergebnisse erbringt, sollte dies auch bei gemischten Daten der Fall sein. In Bezug auf PMM legt die Literatur nahe, dass PMM für alle Variablenskalierungen geeignet sei; es fehlen allerdings Ergebnisse aus MC-Studien. Aufgrund der Verfahrenslogik von PMM ist anzunehmen, dass es zu unplausiblen Imputationen kommt, weil für diskrete Variablen wenige Kategorien und damit auch nur wenige, potentielle Spender vorliegen. Das wiederum führt dazu, dass PMM oftmals ein und denselben Spender in das Set der Spender aufnimmt und somit denselben Spender des Öfteren imputiert; dieses wiederholte Sampeln desselben Spenders sollte letztlich Einfluss auf das Schätzergebnis nehmen. Liegen demnach gemischte Daten vor, könnte dies für PMM ein potentielles Problem darstellen. Damit kann folgendes abgeleitet werden: *Mit Ausnahme von PMM werden alle MDTs bei gemischten Daten nur leicht verzerrte Ergebnisse liefern, sofern die Variablen symmetrisch (normal) verteilt sind (Hypothese 3).*

Sollten gemischte Daten vorliegen, die zudem gewisse Asymmetrien in den Verteilungen der Variablen aufweisen, dann sollten sich auch eher verzerrte Ergebnisse beobachten lassen. Das liegt zum einen daran, dass für die MDTs bereits von einem Einfluss der Verteilungen ausgegangen wird, wenn die (quasi-)metrischen Daten die (multivariate) Normalverteilungsannahme verletzen und diese durch das Hinzukommen von diskreten Variablen immer auf die ein oder andere Weise verletzt wird. Zum anderen zeigt sich für Direct-ML (und damit auch EM), EMB oder MNV, dass bei diskreten Variablen die Ergebnisse verzerrter werden, je stärker die Asymmetrien ausfallen. Für H0 ist diese Annahme zudem sinnvoll, als dass die Handhabung

der fehlenden Werte auf jeglichem Skalenniveau von der Normalverteilungsannahme abhängt.[53] Demzufolge kann tendenziell von einer Verschlechterung der Performanz dieser fünf MDTs ausgegangen werden, wenn die gemischten Daten gewisse Asymmetrien aufweisen (auch wenn die Ergebnisse nicht unbedingt problematische Verzerrungen aufweisen, wie es eben auch in den MC-Studien zum Teil der Fall ist). Zwei Sonderfälle liegen für PMM und für FCS vor. Denn für PMM kann es zwar vorkommen, dass bei gemischten Daten, wegen der diskreten Variablen, keine zufriedenstellenden Ergebnisse mehr produziert werden, diese sollten allerdings unabhängig von der Variablenverteilung sein, da PMM auf tatsächlich beobachtete Daten zurückgreift und eher nicht von den Verteilungen beeinflusst sein sollte. Demnach sollten sich die zu erwartenden Verzerrungen aufgrund der Skalenniveaus bei zunehmend asymmetrisch verteilten, gemischten Daten nicht weiter verschlimmern. Für FCS besteht bei Asymmetrien die Gefahr darin, dass bei solchen Verteilungen die diskreten Variablen nur gering besetzte Zellen aufweisen könnten. Ist das gegeben, können die logistischen Regressionsmodelle Schwierigkeiten bei der Ersetzung der fehlenden Werte bekunden, was dann das Schätzergebnis beeinflusst. In den MC-Studien zeigt sich, dass vor allem diese Modelle bei asymmetrisch verteilten Daten anfällig sind. In einigen Fällen können dabei keine zufriedenstellenden Ergebnisse mehr beobachtet werden. Es ist deshalb anzunehmen, dass mit asymmetrischen Verteilungen die logistischen Regressionsmodelle mittels FCS nicht mehr zufriedenstellend arbeiten und keine zufriedenstellenden Ergebnisse mehr produzieren; weil zudem Verteilungseinflüsse bei (quasi-)metrischen Variablen erwartet werden, könnten für FCS asymmetrisch verteilte, gemischte Daten problematisch sein. Demnach kann angenommen werden, *dass zunehmende Asymmetrien bei gemischten Daten dazu führen, dass FCS nicht mehr zufriedenstellend arbeitet, dass PMM dadurch nicht beeinflusst wird (die Verzerrungen verschlimmern sich nicht) und dass die Verzerrungen durch die anderen MDTs eher zunehmen (Hypothese 4).*

Zusätzlich zur Variablenverteilung als Einflussfaktor ist es denkbar, dass die Samplegröße einen Einfluss auf die MDTs nimmt. Nicht zuletzt sollte dies auf Direct-ML und EM zutreffen, da deren asymptotische Eigenschaft von der Fallzahl beeinflusst wird, was zu größeren Verzerrungen bei geringerer Samplegröße führen kann. Zudem können die MC-Studien nachweisen, dass die Verzerrungen in den Parametern und Standardfehlern bei allen MI-Varianten etwas größer ausfallen, wenn kleinere Fallzahlen gegeben sind. Auch zeigt sich in den MC-Studien,

[53] H0 verwendet lineare Regressionen zur Imputation der metrischen Variablen und die fehlenden Werte auf den diskreten Variablen werden durch die Generierung einer normalverteilten metrischen Hintergrundvariablen mit entsprechenden Schwellenwerten ersetzt.

dass bei kleinen Fallzahlen die ML-Verfahren leicht bessere Ergebnisse liefern als EMB und MNV. In aller Regel bleiben die Verzerrungen aber auch bei kleinen Fallzahlen eher vernachlässigbar. Für alle MDTs gilt deshalb: *Je kleiner das Sample ist, desto eher ist mit verzerrten Ergebnissen zu rechnen (Hypothese 5).*

Zwar kann der Missinganteil den Ausführungen zu den MDTs in Kapitel 3 und Kapitel 4 nicht als Einflussfaktor entnommen werden, aber die MC-Studien zeigen in vielen Fällen, dass mit zunehmendem Missinganteil auch die Verzerrungen zunehmen – wenngleich auch diese, bis zu sehr hohen Anteilen, meist unbedeutend bleiben. Es ist für alle MDTs anzunehmen, *dass bei höheren Anteilen an Missing Values die Verzerrungen in den Ergebnissen zunehmen werden (Hypothese 6).* Eine Besonderheit liegt für EM als Einfachimputation vor. Denn es ist davon auszugehen, dass bei höheren Missinganteilen die Standardfehler verzerrt geschätzt werden, weil die Unsicherheit, die mit den fehlenden Werten einhergeht, nicht adäquat berücksichtigt werden kann (auch wenn in jedem Expectation-Step eine Korrektur des geschätzten Wertes angebracht wird). Zuweilen zeigt sich in den MC-Studien ein widersprüchliches Bild: In einigen Fällen werden die Standardfehler korrekt geschätzt, in anderen nicht. Aufgrund dessen, dass in den MC-Studien keine klare Tendenz auszumachen ist, sei an dieser Stelle an den Ausführungen aus Kapitel 4.2 festgehalten. Damit ergibt sich folgende Hypothese: *Mit EM erhöht sich unter größeren Anteilen an Missing Values die Wahrscheinlichkeit eines falschen inhaltlichen Schlusses (Hypothese 6.1).*

Zuletzt sei darauf verwiesen, dass alle Verzerrungen, wie die Ergebnisdarstellungen der MC-Studien in den Kapiteln 5.1 und 5.2 zeigen, wohl tendenziell größer ausfallen werden, wenn Kombinationen aus den vorgestellten Eigenschaften der Daten vorliegen. Wenn z. B. ein hoher Missinganteil bei einer kleinen Fallzahl und normalverteilten Daten gegeben ist (und damit zwei potentielle Einflussfaktoren, welche die Verzerrungen vergrößern: der Anteil an Missings und eine kleine Fallzahl), dann werden die Verzerrungen größer ausfallen, als wenn ein hoher Missinganteil bei einer großen Fallzahl und normalverteilten Daten vorliegt (und demnach nur ein potentieller Einflussfaktor gegeben ist: der Missinganteil). *Demzufolge sollten sich, unabhängig davon, ob es sich um Daten mir nur (quasi-)metrisch skalierten Variablen oder um gemischte Daten handelt, bei Kombinationen der potentiellen Einflussfaktoren größere Verzerrungen ergeben, als wenn nur einer dieser Einflussfaktoren gegeben ist (Hypothese 7).*

Momentan gibt es für Varianten der MI kaum Studien, die sich mit einzelnen Fit-Indices beschäftigen. Das bedeutet, dass es für die MI auch kaum Hinweise gibt, wie diese im Hinblick auf verschiedene Fit-Indices zu bewerten ist. Aus diesem Grund muss überprüft werden, wie

gut die einzelnen MI-Varianten es schaffen, akzeptable Fit-Indices zu generieren, sodass das resultierende Modell anhand dieser zu akzeptieren ist (wenn es korrekt spezifiziert wird). Eine konkrete Hypothese lässt sich hierbei nicht ableiten, sodass das spätere Ergebnis teilweise einen explorativen Charakter aufweist. Es lässt sich allerdings folgende Frage stellen: *Führt die Behandlung der fehlenden Werte mit den MI-Varianten zu korrekten Schlussfolgerungen mit den Fit-Indices (F1)?* Weiterhin werden die meisten Fit-Indices (CFI, RMSEA, SRMR oder TLI) auch mit den ML-Verfahren kaum untersucht. Zwar gibt es für den p-Wert einige Untersuchungen (mit Direct-ML, EM oder MNV), aus denen sich auch einige Einflussfaktoren ableiten lassen (bei konstantem Missinganteil und abnehmender Fallzahl oder zunehmender Abweichung von der Normalverteilung werden Modelle mittels p-Wert eher zurückgewiesen), allerdings ist unklar, ob sich diese auch auf andere Fit-Indices übertragen lassen; vor allem auch dann, wenn keine metrischen Variablen gegeben sind. Aus diesem Grund sollen anstatt eines Hypothesentests, folgende Fragen beantwortet werden: *Beeinflussen die dargelegten Bedingungen die MDTs, sodass die Performanz der Fit-Indices dadurch beeinträchtigt wird (F2.1)? Wenn ja, welche dieser Bedingungen sind besonders einflussreich (F2.2)?* Eine Zusammenfassung der Hypothesen und der Fragen zu den Fit-Indices lässt sich Tabelle 3 entnehmen.

Tabelle 3: Zusammenfassung der Hypothesen

Empirische Verteilungseigenschaften (Hypothese)	**Parameter und Standardfehler beeinflusst?**							
	FCS	PMM	MNV	EMB	H0	Direct-ML	EM Parameter	EM SE
Quasi-metrische Variablen; symmetrisch normal (H1)	✗	✗	✗	✗	✗	✗	✗	
Quasi-metrische Variablen; (stark) asymmetrisch (H2)	✓	✗	✓	✓	✓	✓	✓	
Gemischte Daten; symmetrisch (normal) (H3)	✗	✓	✗	✗	✗	✗	✗	
Gemischte Daten; (stark) asymmetrisch (H4)	✓	✗	✓	✓	✓	✓	✓	
Kleine Fallzahl (H5)	✓	✓	✓	✓	✓	✓	✓	
Missinganteil (H6 und H6.1)	✓	✓	✓	✓	✓	✓	✓	✓
Kombinationen der Einflussfaktoren (H7)	✓	✓	✓	✓	✓	✓	✓	

Fit-Indices	
Varianten der MI (F1)	Führt die Behandlung der fehlenden Werte mit den MI-Varianten zu korrekten Schlussfolgerungen mit den Fit-Indices?
Simulationskonfigurationen (F2.1 und F2.2)	Beeinflussen die dargelegten Bedingungen die MDTs, sodass die Performanz der Fit-Indices dadurch beeinträchtigt wird? Wenn ja, welche dieser Bedingungen sind besonders einflussreich?

Anmerkungen: Die Hypothesen lassen sich auf eine Ja- und Nein-Antwort mittels der Frage ‚Parameter und Standardfehler beeinflusst?' zusammenfassen: ✗ bedeutet Nein, ✓ bedeutet Ja. Bsp.: Keine der MDTs wird verzerrte Ergebnisse produzieren (H1). Außer PMM werden die Ergebnisse für alle MDTs mit zunehmend asymmetrischen Verteilungen verzerrter werden (H2).

6 Forschungsdesign

In diesem Kapitel wird das Forschungsdesign dieser Arbeit vorgestellt. Dazu gehört neben einer kurzen Einführung in die methodische Herangehensweise auch die Definition der Simulationskonfigurationen und die Festlegung der Kriterien, mit welchen die MDTs bewertet werden sollen. Weiterhin wird in diesem Kapitel die rechentechnische Umsetzung des Designs vorgestellt sowie die Sicherstellung der Randbedingungen diskutiert, auf denen die Hypothesen basieren (unter anderem auch die Sicherstellung der Konvergenz der MDTs).

6.1 Monte-Carlo-Simulationsstudien

Die Performanz statistischer Analysemethoden ist meist nur unter der Bedingung bekannt, dass die Samplegröße unendlich groß wird und dass die, für die Methoden notwendigen, Verteilungseigenschaften der Daten gegeben sind. Empirische Datensätze sind aber nur endlich groß und weisen nur bedingt die notwendigen Verteilungseigenschaften der Daten auf. Um herauszufinden, wie gut einzelne statistische Analysemethoden unter empirischen Bedingungen arbeiten, werden MC-Simulationen[54] benötigt. Mithilfe solcher Simulationen lassen sich Datenstrukturen systematisch variieren und es können empirische Verteilungseigenschaften simuliert werden, in denen verschiedenste Annahmen von statistischen Schätzmethoden verletzt sind. Infolgedessen lassen sich dann Effekte einer oder mehrerer kombinierter Verletzungen auf das Verhalten dieser Schätzmethoden identifizieren.

Bei einer MC-Simulation handelt es sich um einen computergesteuerten Prozess, bei welchem aus einer vorher klar definierten Population wiederholt Zufallsstichproben gezogen werden. Damit ist bei MC-Simulationen die Population – anders als bei empirischen Studien – bekannt. Sie besteht zum einen aus einem statistischen Modell mit entsprechend definierten Zusammenhängen und zum anderen aus bestimmten Dateneigenschaften. Aus dieser Population werden dann wiederholt Zufallsstichproben der Größe *N* gezogen (auch: Replikationen). Auf jeder Stichprobe wird anschließend das statistische Modell geschätzt. Je nach vorliegendem Interesse werden die entsprechenden Ergebnisse dieser Modellschätzungen evaluiert (bspw. die Parameter). Die Sammlung der Ergebnisse aller Stichproben ergibt deren empirische Verteilung (*sampling distribution*) und die Merkmale dieser Verteilung (z. B. die Mittelwerte der Parameter) können dann in Relation zu den im Vorhinein festgelegten Populationswerten

[54] Zum Unterschied zwischen MC-Simulationen und anderen Simulationsstudien: MC-Studien sind statistische Verteilungsuntersuchungen, in denen Stichprobendaten und folglich empirische Stichprobenverteilungen generiert werden. Simulationsstudien generieren ebenfalls Daten, aber nicht zwingenderweise Stichprobendaten (vgl. Bandalos/Leite 2013: 626).

gesetzt werden. Anhand von vorliegenden oder nicht vorliegenden Abweichungen zwischen dem ‚empirischen' Wert und dem Populationswert lassen sich dann die Performanz und damit die Eigenschaften einer Schätzung evaluieren (vgl. Mooney 1997: 1 ff.; Paxton u. a. 2001: 289).

Das Forschungsdesign von Studien mit MC-Simulationen muss demzufolge in drei Teile unterteilt werden: Erstens muss festgelegt werden, welche Schätzmethoden evaluiert werden sollen (etwa verschiedene Modellschätzer, oder auch verschiedene MDTs). Als zweites muss die Population definiert werden und als drittes müssen die Kriterien dargelegt werden, die herangezogen werden, um die ausgewählten Schätzmethoden zu evaluieren.

Die Definition der Population erfolgt durch die Festlegung der Simulationskonfigurationen. Diese werden auch als exogene Modellparameter, unabhängige Variablen, Modellierungskriterien oder als Experimentalbedingungen bezeichnet. Hierunter fallen einerseits die Modelleigenschaften (das ist das Populationsmodell selbst, welche Art von Modell es ist, wie viele Variablen es enthält und wie komplex es ist, inklusive der Werte für die Modellparameter) und andererseits die Eigenschaften der Daten (wie die Samplegröße, die Verteilung der Variablen, deren Messniveau, oder auch der Anteil an Missing Values). Ob und wie sich die Simulationskonfigurationen auf die Ergebnisse der zu evaluierenden Schätzmethoden auswirken und ob sich durch Variationen der Simulationsbedingungen andere Ergebnisse erwarten lassen, wird durch die Bewertungskriterien geprüft (endogene Modellparameter oder abhängige Variablen). Diese beinhalten u. a. die Unverzerrtheit oder die Effizienz der Schätzungen (vgl. Bandalos/Gagné 2012: 100 ff.).

Weil in MC-Studien durch die Festlegung der Population alle möglichen Einflussgrößen auf die zu evaluierenden Schätzmethoden bestimmt und kontrolliert werden können, und weil die Randbedingungen konstant gehalten werden können (der Rahmen der Untersuchung wird durch die Simulationskonfigurationen exakt vorgegeben), können die Ergebnisse, welche die Schätzmethoden liefern, direkt auf diese Einflussgrößen zurückgeführt werden. Schlussendlich werden dadurch kausale Schlussfolgerungen erlaubt, da der kontrollierte Untersuchungsrahmen alle anderen Erklärungsfaktoren bereits im Vorhinein ausschließt (vgl. Krause 2019: 50). Genau dieser Punkt offenbart allerdings auch eine Schwachstelle dieser Art von Studie. Denn die Untersuchungsbedingungen sind immer manipuliert. Damit sind die Ergebnisse von MC-Studien von der Repräsentativität der vorgegebenen Bedingungen abhängig. Das führt einerseits dazu, dass die Brauchbarkeit einer MC-Studie möglicherweise eingeschränkt ist, wenn in der Konzeption der Studie die Komplexität von empirischen Daten nicht abgebildet wird. Andererseits

besitzen die Ergebnisse einer MC-Studie immer nur für die vorgelegten Bedingungen Gültigkeit und können nicht ohne weiteres auf Sachverhalte übertragen werden, die nicht in der Studie überprüft werden (vgl. Bandalos/Leite 2013: 625 ff.; Skrondal 2000: 138). Aus diesem Grund ist es notwendig, die Simulationskonfigurationen möglichst an empirisch vorzufindenden Dateneigenschaften und vorliegenden MC-Studien auszurichten. Ersteres dient dazu, Aussagen für die Empirie zuzulassen, letzteres ermöglicht den Vergleich der Ergebnisse zum aktuellen Forschungsstand, was prinzipiell zur Übertragbarkeit der Ergebnisse beiträgt.

6.2 Simulationskonfigurationen

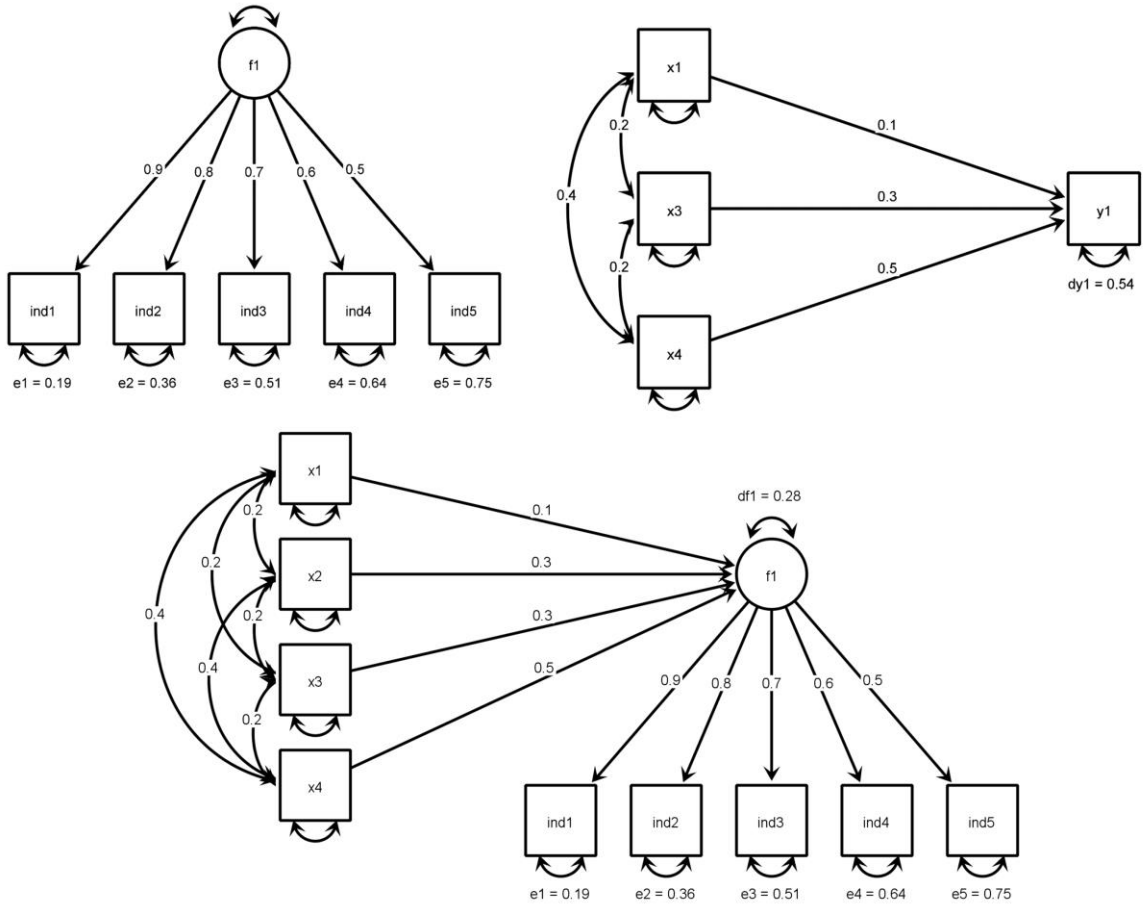
Während die Modell- und Dateneigenschaften die Population einer MC-Studie definieren, stehen die zu evaluierenden Schätzmethoden damit nur bedingt im Zusammenhang. Sie können als weitere Konfigurationsgrößen betrachtet werden, haben aber mit der Population an sich nichts zu tun. Werden verschiedene Schätzmethoden (in diesem Fall sind das die MDTs), in das Design einer Studie aufgenommen, kann verglichen werden, welche der Methoden für bestimmte Modell- oder Dateneigenschaften am geeignetsten ist. Die einzelnen MDTs werden demnach als unabhängige Parameter definiert und dazu verwendet, um die fehlenden Werte in den generierten Daten zu berücksichtigen. Auf den imputierten Daten (bei Direct-ML auf den Daten mit Missings) wird dann das interessierende Modell geschätzt. Diese Modellschätzungen werden im nächsten Schritt mit dem Populationsmodell verglichen. Da die Populationswerte bekannt sind, wird daraus ersichtlich, welche MDT sich unter welchen Dateneigenschaften am ehesten eignet, die Missing Values zu handhaben.

Das zentrale Anliegen dieser Arbeit ist die Evaluation von FCS und die Evaluation von PMM. Zudem soll ein Vergleich dieser Varianten zu JM erfolgen. Hierbei zeigte sich, dass MNV bereits mehrfach untersucht wurde, EMB in der MC-Forschung aber kaum Berücksichtigung fand. Infolgedessen werden beide Varianten untersucht. Gleichzeitig wurde auch H0 bisher kaum untersucht, sodass auch diese Variante als Prüfgröße herangezogen wird. Zudem wurde das Ziel ausgegeben, wonach FCS mit Direct-ML zu vergleichen ist. Aus diesem Grund wird auch Direct-ML in das Design aufgenommen. Die Einfachimputation mittels EM wird berücksichtigt, um eine Evaluation der Zusammenführung der *m* Fit-Indices zu erlauben (siehe Kapitel 5.4). Demzufolge werden insgesamt sieben verschiedene MDTs in das Design aufgenommen. Die Implementation, die Konfiguration und die Sicherstellung der Konvergenz dieser Verfahren wird in Kapitel 6.4.2 diskutiert.

6.2.1 Modelleigenschaften: die Populationsmodelle

Zunächst muss ausgewählt werden, welche(s) Modell(e) geprüft werden soll(en). Werden im Sinne eines Modellierungskriteriums mehrere Modellarten spezifiziert, erhöht sich die Übertragbarkeit der Ergebnisse dementsprechend. Da für verschiedene MDTs im SEM-Kontext kaum differenzierte Untersuchungen existieren, liegen auch nur wenige Befunde hinsichtlich deren Verhaltens in einfachen und komplexeren Modellstrukturen vor. Deshalb sollen drei, im Grunde aufeinander aufbauende, Modelle spezifiziert werden: ein einfaktorielles CFA-Modell (Modell 1), ein rekursives Pfadmodell (Modell 2) sowie ein Gesamtmodell, das die Strukturen der ersten beiden Modelle integriert (Modell 3). Die Modelle inklusive ihrer Parametereigenschaften sind in der folgenden Abbildung 9 dargestellt.

Abbildung 9: Populationsmodelle



Anmerkungen: oben links: Modell 1; oben rechts: Modell 2; unten: Modell 3. Ein doppelseitiger Pfeil, der nur auf eine jeweilige Variable/ein jeweiliges latentes Konstrukt zeigt, stellt die Varianzen/Residualvarianzen dar. Varianzen der unabhängigen Variablen sind auf 1.0 festgesetzt (Wert nicht dargestellt).

Neben der Modellart werden bei der Modellspezifikation auch Werte für die Zusammenhänge zwischen einzelnen Variablen (das können Faktorladungen, Kovarianzen, Strukturpfade u. a. sein) und deren Varianzen festgelegt. Diese dienen als Populationswerte und sollten sich so weit wie möglich an empirisch häufig anzutreffenden Werten sowie an anderen MC-Studien

orientieren. In Anlehnung an die bisherige Forschung werden für die Faktorladungen (λ) Werte von .9, .8, .7, .6 und .5 gewählt, was zu Faktorladungen von durchschnittlich .7 führt. Die Varianz des Faktors wird auf einen Wert von 1.0 festgesetzt.[55] Die Fehlervarianzen (*errors* oder *uniquenesses*) werden berechnet, indem die quadrierte Faktorladung von der Faktorvarianz abgezogen wird ($1-\lambda^2$). Sie weisen dadurch die folgenden Werte auf: .19; .36; .51; .64 und .75. Die Effekte der exogenen Variablen auf die endogene Variable, sowohl im Pfad- als auch im integrierten Strukturmodell, werden mit .1, .3 und .5 angegeben. Für die bestehenden Kovarianzen zwischen den exogenen Variablen werden Werte in Höhe von .2 und .4 spezifiziert. Die Varianzen der unabhängigen Variablen werden in den Modellen wiederum auf 1.0 fixiert. Die Residualvarianzen (oder *disturbances*) für Modell 2 (dy1) und Modell 3 (df1) werden anhand der erklärten Varianz (R^2) mit Hilfe der Regeln von Wright (1934; siehe dazu auch: Duncan 1975; Loehlin/Beaujean 2017) berechnet. Sie belaufen sich für Modell 2 auf .538, für Modell 3 auf .28. Mit den festgelegten Parametern gehen folglich mittelgroße und geringe Anteile an gebundener Varianz einher. Alle diese Werte spiegeln wider, was sich in der angewandten Literatur und in MC-Studien[56] finden lässt.

Diese Modellspezifikationen mit diesen Parametrisierungen sollten zur Vergleichbarkeit zu anderen Studien im gleichen/ähnlichen Themengebiet beitragen, da sich die ausgewählten Parameter auch in anderen Studien finden. Auch sollten sie eine gewisse Übertragbarkeit auf die Empirie zulassen, da alle drei Modelle in der empirischen Forschung zum Einsatz kommen. So ist ein einfaches CFA-Modell vor einer jeden Analyse eines SE-Modells zu rechnen, um die Passung der analytisch abgeleiteten Modellstruktur zu prüfen und Modell 2 stellt ein multivariates lineares Regressionsmodell dar (auch eine häufig eingesetzte Analysemethode). Vor allem ein Modell aus manifesten Variablen und latenten Faktoren, in dem über Strukturpfade eine kausale Verknüpfung unterstellt wird, ist eine Modellart, die viele Eigenschaften von Analysen im Zusammenhang mit SEM miteinander verbindet.

[55] Durch die Fixierung der Varianzen auf 1.0 und die vorgestellten Berechnungsweisen der Fehler- und Residualvarianzen wird eine Standardisierung der Effekte erreicht. Es handelt sich bei den Populationsparametern aller Modelle demnach um standardisierte Effektgrößen. Siehe dazu auch die Diskussionen im M*plus*-Forum: http://www.statmodel.com/discussion/messages/9/980.html?1501616031 und http://www.statmodel.com/discussion/messages/11/24502.html?1505005079.

[56] Es sei auf die Studien in Kapitel 5 verwiesen (das gilt für alle Rechtfertigungen in Kapitel 6.2).

6.2.2 Dateneigenschaften: Variablenskalierungen und Variablenverteilungen

Eine Prämisse der Arbeit ist die Untersuchung der MDTs bei quasi-metrischen und bei gemischten Daten. Aus diesem Grund werden die einzelnen Variablen wie folgt skaliert: Die latenten Konstrukte (Modell 1 und Modell 3) werden mit quasi-metrischen Indikatorvariablen geschätzt (ind1 – ind5). Das rekursive Pfadmodell (Modell 2) besteht aus einer quasi-metrisch definierten (x1, 5er-Skala), einer binären (x3) und einer ordinalen (x4, 4er-Skala[57]) exogenen Variable. Die abhängige Variable ist wiederum eine quasi-metrisch definierte Variable (y1, 5er-Skala). Für Modell 3 dient der latente Faktor f1 als abhängige Variable. Gleichzeitig wird noch eine weitere quasi-metrisch definierte exogene Variable aufgenommen (x2, 5er-Skala). Damit weisen die Indikatoren, die abhängige Variable in Modell 2 sowie die determinierenden exogenen Variablen Skalenniveaus auf, die durchaus in der praktischen Analyse von SE-Modellen weit verbreitet sind. Gleichzeitig werden mithilfe dieser Festlegungen Daten generiert, die einerseits rein quasi-metrisch sind und andererseits unterschiedliche Variablenskalierungen aufweisen.

Als weitere Konfigurationsgröße soll die Verteilung der Variablen miteinbezogen werden. Sowohl die gewünschten Skalierungen als auch die Verteilungen der Variablen hängen in MC-Studien eng miteinander zusammen, denn quasi-metrische und diskrete Variablen werden dadurch erzeugt, dass standardnormalverteilte Variablen in sinnhafte Kategorien ‚geschnitten' werden (im Weiteren Kategorisierung). So können je nach Wunsch verschieden viele Kategorien erzeugt werden. Die gewünschten Verteilungen werden während der Kategorisierung generiert, indem bestimmt wird, welcher prozentuale Anteil an Fällen auf die jeweilige Kategorie entfallen soll (siehe Bandalos 2014; vgl. Muthén/Muthén 2012: 781 ff.). Eine Kategorisierung der Variablen hat allerdings einen Nebeneffekt: In einem solchen Fall sind die Werte für die Skewness und die Kurtosis nur noch bedingt aussagekräftig.[58] Das liegt daran, dass bei kategorisierten Variablen auch extreme Verteilungen, die grafisch eine Verletzung der Normalverteilung anzeigen, angemessene Werte für die Skewness und Kurtosis hervorbringen können.

---

[57] Die in dieser Arbeit für die spätere Analyse nicht erst in Dummies umgewandelt wird, sondern als quasi-metrisch analysiert wird. Das ist zwar generell nicht zu empfehlen, allerdings steht nicht die Performanz eines Modellschätzers zur Diskussion, sondern die Fähigkeit verschiedener MDTs fehlende Werte zu behandeln. Dies wiederum sollte unter Bedingungen geschehen, die möglichst nicht optimal sind. Mit Einbezug von unterschiedlich skalierten Variablen, die zum Teil auch Annahmen des späteren Modellschätzers verletzen, soll dies sichergestellt werden. Weiterhin wird der Einfluss der benutzten Schätzmethode aus den Ergebnissen isoliert, sodass der reine Einfluss der MDTs sichtbar wird, weshalb diese Handhabung unproblematisch erscheint (siehe Kapitel 6.4.3).

[58] Abgesehen von der Tatsache, dass sich die Berechnungen von Kurtosis und Skewness zwischen einzelnen statistischen Analyseprogrammen unterscheiden. Ist das Analyseprogramm nicht angegeben, können auch die Werte nicht angemessen eingeordnet werden, weil diese auf verschiedenen Berechnungsarten basieren. Das kann, bei Anwendung unterschiedlicher Programme, zu unterschiedlichen Ergebnissen führen (siehe Joanes/Gill 1998).

Neben der Problematik, dass keine quantitativen Maße angegeben werden können, welche die Verteilung klassifizieren, resultiert eine Kategorisierung von Variablen immer in einer mehr oder minder starken Abweichung von der Normalverteilung. Das liegt daran, weil kategorisierte Variablen naturgemäß diskret und damit recht grobe Messungen der latenten kontinuierlichen Hintergrundvariablen sind (siehe Bollen 1989b; vgl. Kaplan 2000: 83). Da aber die Performanz der MDTs unter normalverteilten Variablen geprüft werden soll, muss die Kategorisierung der Variablen so erfolgen, dass die Variablen noch annähernd als normalverteilt gelten können: Das wird erreicht, indem die Variablen so zugeschnitten werden, dass sie als symmetrisch normal einzuordnen sind. Verschiedene Studien[59], welche die Performanz statistischer Verfahren unter kategorisierten Variablen untersuchen, definieren eine Verteilung mit fünf Antwortkategorien als symmetrisch normal, wenn die Mittelkategorie zwischen 40 % und 50 % der Fälle aufweist. Zusätzlich sollten die äußeren Kategorien zwischen 5 % und 10 % der Fälle auf sich vereinigen. Verteilungen, die bei fünf Antwortkategorien demnach z-Werte von -1.5, -.5, .5 und 1.5 aufweisen, werden in dieser Arbeit als symmetrisch normal definiert.[60]

Weil auch Hypothesen vorhanden sind, die Aussagen darüber treffen, wie sich die einzelnen MDTs bei nicht mehr normalverteilten Variablen verhalten, werden zusätzlich zur symmetrisch normalen Verteilung, zwei asymmetrische Verteilungen in das Design aufgenommen. Verteilungen, bei denen sich ca. 65 % der Fälle links des Skalenmittelpunkts befinden, werden als asymmetrisch definiert; bei stark asymmetrischen Verteilungen liegen dagegen ca. 80 % der Fälle links vom Skalenmittelpunkt. Diese Werte werden auch für die binäre und die ordinale Variablen herangezogen, was in den Verteilungen resultiert, die in Abbildung 10 dargestellt sind. Die Werte für Skewness und Kurtosis (der Vollständigkeit halber) sowie die Anteils- und z-Werte lassen sich Tabelle 4 entnehmen. Die ausgewählten Verteilungen entsprechen in etwa denjenigen, die in bisherigen MC-Studien implementiert wurden. Mit dem Einbezug von Variablen unterschiedlichen Skalenniveaus und deren Verteilungsvariation dürfte das vorliegende

[59] Hierzu müssen zusätzliche MC-Studien zu denen aus Kapitel 5 herangezogen werden, da in den dortigen Studien meist keine Kategorisierung der Variablen stattfindet. Die Aussagen beziehen sich auf folgende Arbeiten: Finney/DiStefano (2013); Rhemtulla u. a. (2012); Teman (2012); Yang-Wallentin u. a. (2010); Muthén/Kaplan (1992); Lei (2009); Forero u. a. (2009).

[60] Da die Kategorien anhand einer Standardnormalverteilung mit einem Durchschnitt von null und einer Standardabweichung von eins erstellt werden, lassen sich die Anteile der jeweiligen Kategorien in *thresholds* angeben. Diese *thresholds* sind die z-Werte einer Standardnormalverteilung. Je nachdem wie hoch der Anteil an Fällen für die einzelnen Kategorien gewählt wird, desto größer/kleiner sind die z-Werte. Die z-Werte lassen sich berechnen, wenn im Vorhinein die Anteilswerte der einzelnen Kategorien bekannt sind. Die Anteilswerte der Kategorien lassen sich berechnen, wenn die jeweiligen z-Werte bekannt sind. In dieser Arbeit wird sowohl das eine als auch das andere Vorgehen angewendet. Die Berechnungen werden mithilfe eines Tools der Universität Köln durchgeführt (zu finden unter: http://eswf.uni-koeln.de/glossar/surfstat/normal.htm).

Design die gängige empirische Praxis nachempfinden, in welcher Daten benutzt werden, die für die zu schätzenden Modelle eben nicht optimal sind, da sie oftmals deren zugrundeliegenden Annahmen verletzen.

Tabelle 4: Skewness, Kurtosis, Anteils- und z-Werte der gewählten Verteilungen[61]

Werte für Skewness und Kurtosis

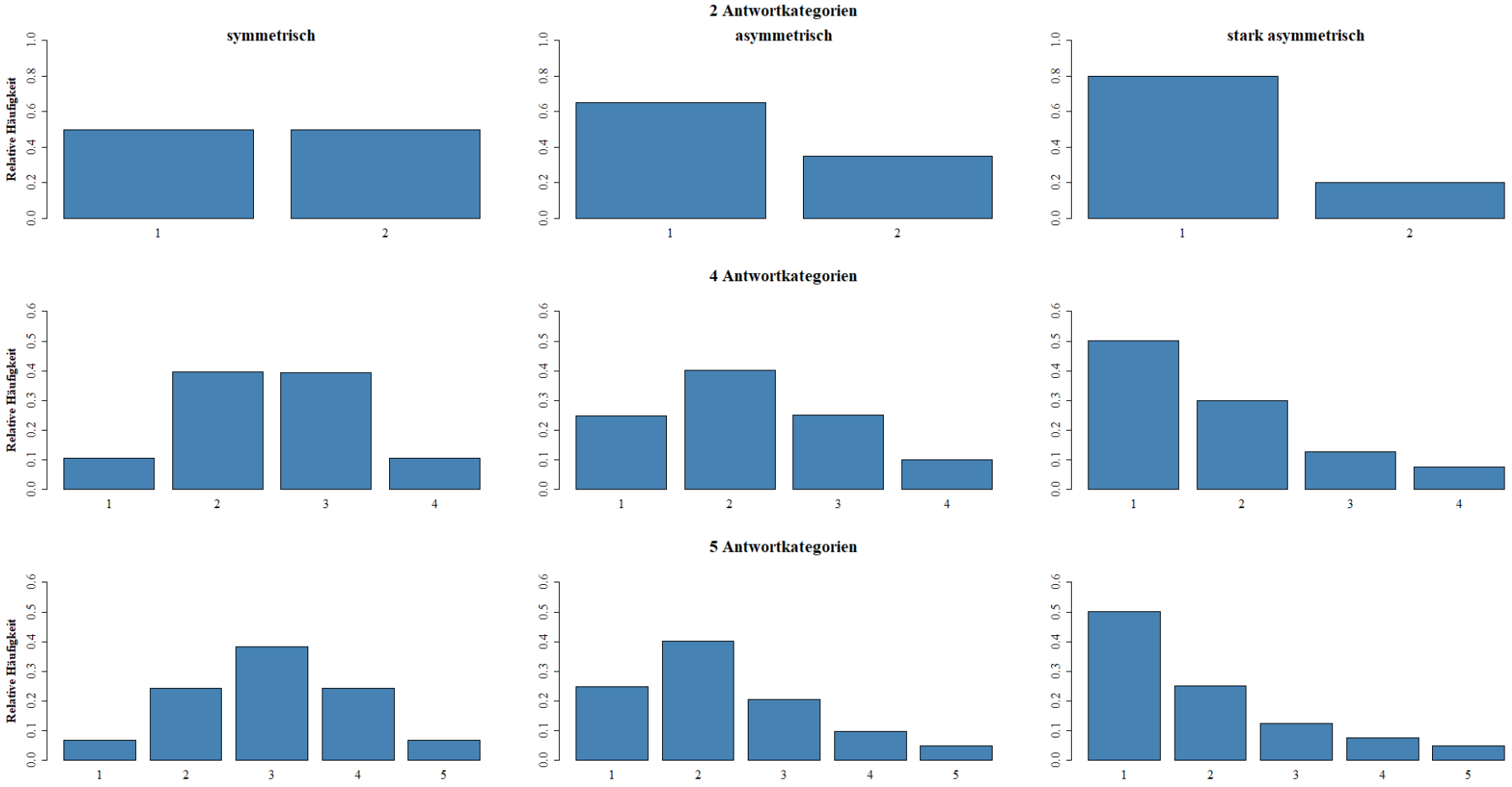
Antwortkategorien	**Verteilung**					
	symmetrisch		*asymmetrisch*		*stark asymmetrisch*	
	Skewness	Kurtosis	Skewness	Kurtosis	Skewness	Kurtosis
2	0.00	-2.00	0.63	-1.60	1.50	0.24
4	0.00	-0.52	0.35	-0.74	1.01	0.00
5	0.00	-0.47	0.74	-0.06	1.17	0.41

Anteils- und z-Werte

	Verteilung	**Fälle pro Kategorie in %**					**z-Werte (*thresholds*)**			
2	*symmetrisch*	50	50				0			
	asymmetrisch	65	35				0.385			
	stark asymmetrisch	80	20				0.84			
4	*symmetrisch*	10.6	39.4	39.4	10.6		-1.25	0	1.25	
	asymmetrisch	25	40	25	10		-0.675	0.385	1.285	
	stark asymmetrisch	50	30	12.5	7.5		0	0.84	1.44	
5	*symmetrisch*	6.7	24.2	38.3	24.2	6.7	-1.5	-0.5	0.5	1.5
	asymmetrisch	25	40	20	10	5	-0.675	0.385	1.05	1.65
	stark asymmetrisch	50	25	12.5	7.5	5	0	0.675	1.15	1.65

[61] Die dargestellten Werte (Tabelle 4) sowie die Verteilungen in Abbildung 10 beruhen auf Simulationsdaten. Aus einer Standardnormalverteilung mit einem Durchschnitt von null und einer Standardabweichung von eins, werden 1.000.000 Fälle gezogen. Diese werden dann anhand der *thresholds* entsprechend zugeschnitten. Die Skewness und Kurtosis werden nach Typ 2 in Joanes/Gill (1998) berechnet (auf dieser Berechnungsform basieren auch die Werte für Kurtosis und Skewness in SPSS). Die Simulation und die Berechnungen werden mit der R-Version 3.4.3 in R-Studio (Version 1.1.419) und den entsprechenden Paketen durchgeführt (*stats* für die Simulation; *psych* in der Version 1.7.8 für die Berechnungen).

Abbildung 10: Verteilungen der Variablen



6.2.3 Dateneigenschaften: Samplegrößen und fehlende Werte

Wenn möglich, sollten für empirische Analysen möglichst Datensätze mit einer großen Fallzahl vorliegen. Schon allein deshalb, weil verschiedene Analysemethoden eine unterschiedlich hohe Anzahl an Fällen benötigen, um möglichst gute Schätzergebnisse zu generieren. Allerdings variieren die Stichprobenumfänge in der Empirie stark. Eine einzige Samplegröße, die das Gros der empirischen Analysen abbildet, wird sich nicht finden lassen. Zudem konnte herausgearbeitet werden, dass die Performanz der MDTs wohl durch die Samplegröße beeinflusst wird, weshalb es angebracht ist, die Samplegröße in dieser MC-Studie zu variieren. Für die vorliegende Arbeit werden mit einer Fallzahl von 250 und von 750 zwei Samplegrößen definiert, die häufig in MC-Studien und empirischen Studien anzutreffen sind.

Schlussendlich liegt das Interesse der Arbeit in der Bewertung verschiedener MDTs und eine Bewertung derselben ist nur möglich, wenn fehlende Werte vorliegen, die gehandhabt werden müssen. In MC-Studien kann demzufolge zum einen der Ausfallmechanismus der Missings als exogener Parameter miteinbezogen werden und zum anderen der Anteil an Missing Values selbst. Da bei empirischen Daten die Annahme, dass den fehlenden Werten ein MCAR-Mechanismus zugrunde liegt, oftmals nicht aufrechterhalten werden kann, werden die fehlende Werte mit einem MAR-Ausfallmechanismus simuliert.[62] Es wird angenommen, dass MDTs, die für diesen Ausfallmechanismus zufriedenstellende Ergebnisse erbringen, auch für MCAR geeignet sind. Der Anteil an Missing Values orientiert sich wiederum an bereits geleisteten MC-Studien. Unterteilt werden diese in eine geringe, eine mittlere und eine hohe Quote (in den meisten Fällen liegt diese zwischen 25 % bis 40 %). Diese Arbeit folgt dieser Einordnung und setzt das Maximum an Missing Values auf 35 % fest. Insgesamt sollen drei Stufen an Anteilsquoten simuliert werden. Das sind: 5 %, 20 % und 35 %. Damit liegen die Missinganteile im gängigen Bereich der aktuellen Forschungsarbeit.

In Tabelle 5 werden die Simulationskonfigurationen zusammengefasst. Insgesamt liegen pro MDT 54 mögliche Konfigurationen vor. Das wiederum entspricht im Gesamten einem Umfang von 378 Simulationskonfigurationen, die mit dieser Arbeit geprüft werden. Bevor im nächsten Unterkapitel die Bewertungskriterien vorgestellt werden, muss zunächst noch die Anzahl der Replikationen festgelegt werden, also wie viele Stichproben aus einer der Konfigurationen ge-

[62] Die Missing Values sollen dabei ein generelles Muster aufweisen, da dieses die höchsten Anforderungen an die MDTs stellt (siehe dazu Kapitel 2).

zogen werden. Diese Anzahl beläuft sich in dieser Arbeit auf 500. Somit liegen pro Konfiguration ausreichend Fälle vor, um später auch komplexere statistische Analysen mit den Daten durchführen zu können.

Tabelle 5: Unabhängige Variablen in dieser Arbeit – Simulationskonfigurationen

Modelleigenschaften:
- Modell 1: CFA Modell
- Modell 2: Pfadmodell
- Modell 3: integriertes Strukturmodell

Missing Data Techniken:
- Direct-ML
- EM
- EMB
- FCS
- H0
- MNV
- PMM

Dateneigenschaften:
- Samplegröße
  - N = 250
  - N = 750
- Variablenverteilung
  - symmetrisch
  - asymmetrisch
  - stark asymmetrisch
- Anteile an fehlenden Werten unter MAR
  - 5 %
  - 20 %
  - 35 %

Simulationskonfigurationen:
- 3 Modelle*2 Samplegrößen*3 Verteilungen*3 Missingquoten = 54 Konfigurationen
- 54 Konfigurationen für jede der 7 MDTs entspricht 378 Simulationskonfigurationen

6.3 Bewertungskriterien

An den Bewertungskriterien kann abgelesen werden, ob die Performanz der MDTs durch die Simulationskonfigurationen beeinflusst ist oder nicht. Sie lassen sich in zwei Ebenen unterteilen: in die Modell- und in die Parameterebene. Auf der Modellebene lassen sich die Fit-Indices der Modellschätzungen analysieren, auf der Parameterebene geht es um die Überprüfung, ob akzeptable Schätzergebnisse im Hinblick auf die Populationsparameter erfolgt sind. Im Grunde wird hierbei die Präzision der Modellschätzung beurteilt: Das ist neben der Unverzerrtheit der Schätzung der Parameter und Standardfehler auch deren Konsistenz und Effizienz.

Die aufgestellten Hypothesen nehmen einen Zusammenhang zwischen den Simulationskonfigurationen und der Performanz der einzelnen MDTs an. Solche Hypothesen lassen sich in MC-Studien auf zwei Arten bewerten: Es können neben deskriptiven Analysen auch inferenzstatistische Modelle gerechnet werden (siehe Boomsma 2013). Die deskriptiven Analysen erfolgen im Hinblick auf die Vergleichbarkeit dieser Arbeit zu anderen Studien, weswegen Bewertungskriterien ausgewählt werden, die auch in anderen Studien Verwendung finden: Das ist für die Modellebene die Ablehnungsrate mittels der Fit-Indices (*rejection rate*) und für die Parameterebene sind das der relative Bias der Parameterschätzung, deren relative Effizienz und der relative Bias des Standardfehlers. Im Gegensatz zu den deskriptiven Analysen lassen die inferenzstatistischen Modelle (auch: Meta-Modelle), durch die Möglichkeit Effekte höherer Ordnung miteinbeziehen zu können, die Identifikation der hauptsächlichen Einflussgrößen auf

einzelne Bewertungskriterien zu. Für die Meta-Modelle werden Regressionsschätzungen herangezogen. Auf der Modellebene handelt es sich um logistische Regressionsmodelle, auf der Parameterebene kommen lineare Regressionsmodelle zum Einsatz. Durch die Kombination von deskriptiven und inferenzstatistischen Analysen, sollte eine umfangreiche Evaluierung der MDTs gewährleistet und die Vergleichbarkeit zu anderen Studien sichergestellt sein.

6.3.1 Modellebene: Fit-Indices

Die Fit-Indices bei der Schätzung von SE-Modellen zeigen an, inwiefern ein spezifiziertes Modell auf die empirischen Daten passt bzw. wie sehr eine geschätzte Kovarianzmatrix mit einer beobachteten Matrix übereinstimmt. In einer MC-Studie ist bereits im Vorhinein sichergestellt, dass eine Passung zwischen den beiden vorliegt, da die Daten anhand eines bekannten Populationsmodells erzeugt werden. Es können also akzeptable Fit-Werte erwartet werden. Weil allerdings Missing Values vorhanden sind, führt dies dazu, dass die SE-Modelle nicht direkt geschätzt werden können (mit Ausnahme von Direct-ML). Erst müssen die fehlenden Werte ersetzt werden, bevor die Modellschätzungen erfolgen können, welchen die Fit-Indices entnommen werden können. Schaffen es die MDTs die fehlenden Werte entsprechend zu handhaben, dann sollten die Werte für die Fit-Indices akzeptabel sein. Es sollen insgesamt vier Fit-Indices untersucht werden, die gemeinhin auch in der Empirie verwendet werden (und teilweise auch bereits in anderen MC-Studien zur Performanz von MDTs untersucht wurden):

- die korrigierte Chi²-Statistik[63],
- das SRMR (*Standardized Root Mean-Square Residual*),
- der RMSEA (*Root Mean-Square Error of Approximation*) und
- der CFI (*Comparative Fit Index*).[64]

Der Chi²-Anpassungstest gibt einen Hinweis auf die Qualität des Modells, indem er prüft, ob es signifikante Unterschiede zwischen den beobachteten und geschätzten Kovarianzen gibt. Weichen die geschätzten Kovarianzen nicht signifikant von den beobachteten Kovarianzen ab, besitzt das geschätzte Modell einen akzeptablen Fit. Allerdings ist der Test (nicht nur[65]) im

[63] Robust gegenüber Verletzungen der Normalverteilungsannahme.
[64] Unterscheiden lassen sich die Fit-Indices in absolute/globale und relative/inkrementelle Fit-Indices. Nur beim CFI handelt es sich um einen relativen Fit-Index. Die relativen Fit-Indices geben an, wie weit das spezifizierte Modell von einem Nullmodell oder perfekten Modell entfernt liegt (unter anderem Bentler 1990; Bollen 1989a). Die anderen drei Indices gehören zu den absoluten Fit-Indices. Sie prüfen, inwiefern die beobachtete Kovarianzmatrix mit der Kovarianzmatrix der Population, die durch das Modell impliziert wird, übereinstimmt.
[65] So hängt das Ergebnis des Tests auch von der Anzahl der verwendeten Indikatoren pro Faktor ab. Werden sehr viele Indikatoren pro Faktor ausgewählt, kann der Chi²-Wert eines Messmodells nach oben verzerrt werden. Das bedeutet dann einen schlechteren Modellfit (vgl. Urban/Mayerl 2014: 119).

Hinblick auf den Stichprobenumfang sehr anfällig. Im Grunde gilt: Je höher die Fallzahl ist, desto höher ist auch die Teststärke dieses Tests und desto einfacher ist es, ein statistisch signifikantes Ergebnis zu erhalten. Das führt häufig fälschlicherweise dazu, dass der Test ein korrekt spezifiziertes Modell zurückweist und die Nullhypothese ablehnt (die annimmt, dass das Modell mit den Daten übereinstimmt), obwohl diese beibehalten werden sollte.

Das SRMR misst die durchschnittliche Abweichung der geschätzten von den beobachteten Korrelationen. Bei einem Wert von .06 bedeutet das, dass das Modell die Korrelationen mit einem durchschnittlichen Fehler von .06 wiedergibt (vgl. Bentler 2006: 352). Hu/Bentler (1999) geben an, dass Werte unter .08 als akzeptabel zu betrachten sind.

RMSEA ist ein Maß für die Differenz zwischen der beobachteten Kovarianzmatrix pro Freiheitsgrad und der Kovarianzmatrix der Population, wie sie durch das Modell impliziert wird. Das Maß bezieht in seine Berechnung die Komplexität des Modells in Form der Freiheitsgrade mit ein, was dazu führt, dass mit sinkender Anzahl an Freiheitsgraden RMSEA anfälliger wird und zu größeren Werten tendiert (er bevorzugt sparsamere Modelle[66]). Auch für RMSEA gilt, dass dessen Werte möglichst klein sein sollten. Werte unter .05 deuten auf eine gute Modellanpassung hin (vgl. Browne/Cudeck 1992: 239). Zusätzlich zur Punktschätzung lässt sich das 90 %ige Konfidenzintervall berechnen. Dieses sollte Werte zwischen .00 und .10 aufweisen. Mit dem Konfidenzintervall lassen sich nach MacCallum u. a. (1996: 133 f.) drei Hypothesen testen: die *exact-fit*, die *close-fit* und die *not-close-fit* Hypothese. Die *exact-fit* Hypothese stellt die Bedingung, dass die untere Intervallgrenze die Null miteinschließt. Tut sie das nicht, ist die Hypothese eines exakten Fits zu verwerfen. Schließt das untere Intervall den Wert .05 mit ein, so ist die *close-fit* Hypothese zu akzeptieren. Wenn das obere Intervall unter einem bestimmten Wert liegt (z. B. .07 nach Steiger 2007 oder .10 bei einer weniger strengen Auffassung nach MacCallum u. a. 1996), dann ist die *not-close-fit* Hypothese zu verwerfen, was wiederum für einen akzeptablen Modellfit spricht.

Im Gegensatz dazu, vergleicht der CFI die Ergebnisse der Schätzung eines Nullmodells (oder Unabhängigkeitsmodells), bei dem angenommen wird, dass alle Variablen im Modell unabhängig voneinander sind (also Kovarianzen aufweisen die null sind), mit den Ergebnissen der Schätzung des spezifizierten Modells. Er gibt dann die anteilsmäßige Verbesserung im Modellfit an. Der CFI weist einen Wertebereich zwischen null und eins auf und je näher der Wert

[66] Kenny u. a. (2014) zeigen in ihrer Studie, dass RMSEA bei einer geringen Anzahl an Freiheitsgraden gar nicht erst zur Evaluation für die Modellpassung benutzt werden sollte.

des CFIs an eins liegt, desto besser ist der Fit des Modells. Für eine gute Modellanpassung sollten Werte über .95 vorliegen.

Die Bewertung der Fit-Indices erfolgt über die Ablehnungsrate (*rejection rate*). Dazu werden die Fit-Indices dichotomisiert. Für jede Replikation innerhalb der Simulationskonfiguration wird eine dichotome Variable ausgegeben, die angibt, ob das Modell anhand des Fit-Indexes verworfen wird (Wert: 1) oder nicht (Wert: 0). Wenn diese dichotome Variable aufsummiert wird und durch die Anzahl der Replikationen innerhalb einer Simulationskonfiguration dividiert wird, dann ergibt sich der Anteil an Modellen, die zurückgewiesen/akzeptiert werden. Als Kriterium für die Dichotomisierung der Fit-Indices werden die berichteten Grenzwerte verwendet. Der Chi²-Wert wird anhand seines p-Wertes dichotomisiert (der Wert wird als signifikant aufgefasst, wenn $p < .05$ ist). Alle signifikanten Werte deuten eine Zurückweisung des Modells an und umgekehrt. Für das SRMR wird ein Grenzwert von .08, für RMSEA von .05 herangezogen. Wenn das/der SRMR/RMSEA des Modells $\leq .08/.05$ ist, wird dies als akzeptable Passung aufgefasst. Weiterhin wird die Ablehnungsrate mittels dem Konfidenzintervall von RMSEA evaluiert. Unterschreitet die untere Intervallgrenze den Wert .05 *und* überschreitet die obere Grenze den Wert .10 *nicht*, dann wird auch dies als akzeptable Passung aufgefasst. Für den CFI wird die Grenze bei .95 festgelegt; Modelle mit einem CFI $\geq .95$ werden akzeptiert.

Für alle Fit-Indices werden Ablehnungsraten von 5 % als akzeptabel aufgefasst. Das kann insofern begründet werden, als dass die einzelnen Replikationen anhand von Populationsmodellen mit perfektem Fit generiert werden. Aufgrund der Vielzahl an Ziehungen aus der Population (je 500) können die Ablehnungsraten des p-Wertes der Chi²-Statistik in Bezug zum Alpha-Fehler interpretiert werden, was bedeutet, dass bei ca. 5 % der Stichproben/Replikationen das Modell fälschlicherweise abgelehnt wird, obwohl es eigentlich aus der zugrundeliegenden Population stammt. Da nun im Idealfall alle Fit-Indices bei der Modellbewertung zum selben Schluss führen sollten, kann dieser Wert auf die anderen Fit-Indices übertragen werden.

6.3.2 Parameterebene: Parameterbias – Unverzerrtheit der Schätzung[67]

Die Verzerrung der Parameterschätzung (Parameterbias oder kurz: Bias) bezeichnet die Abweichung des geschätzten – ob relativ oder absolut – vom wahren, festgelegten Populationswert. Für den absoluten Bias wird der festgelegte Populationswert des Parameters θ_i vom jeweils

[67] Zusätzlich zu den explizit zitierten Studien dienen in den Kapiteln 6.3.2 bis 6.3.4 folgende Studien, auch in Bezug auf die Formeln, als Quellen. Für den Parameterbias: Gold u. a. (2003); Leite/Beretvas (2010); Li (2010); Olinsky u. a. (2003); Teman (2012); Wu u. a. (2015). Für die relative Effizienz: Arbuckle (1996); Enders (2001b); Wang (2007); Zhu (2014). Für den SE-Bias: Newman (2003); Orcan (2013); Yoo u. a. (2007).

geschätzten Wert $\hat{\theta}_{ij}$ für den Parameter *i* innerhalb der Konfiguration *j* subtrahiert und danach dessen Betrag gebildet:

$$Bias_{abs} = |\hat{\theta}_{ij} - \theta_i|. \quad (6.1)$$

Mit dem absoluten Bias liegt damit ein Wert für die Abweichung zwischen geschätztem und festgelegtem Wert pro Replikation innerhalb einer Konfiguration vor. In deskriptiver Hinsicht ist dieses Bewertungskriterium wenig aussagekräftig, allerdings lässt er sich für eine inferenzstatistische Analyse mittels linearer Regression heranziehen. Aus diesem Grund wird auch der Betrag des Wertes gebildet, denn der absolute Bias kann negative Werte annehmen. Dadurch läuft man bei Regressionsanalysen Gefahr, dass sich die negativen und positiven Werte auf der abhängigen Variablen gegenseitig aufheben, was letztlich dazu führen kann, dass kein Zusammenhang und damit keine Varianzaufklärung zwischen den unabhängigen und abhängigen Variablen vorliegt. Durch die Betragsbildung wird erreicht, dass nur ‚positive' Verzerrungen vorliegen. Das führt dazu, dass tatsächliche Zusammenhänge zwischen den unabhängigen Variablen und der abhängigen Variablen sichtbar werden und sich Aussagen darüber machen lassen, welchen Anteil an der Varianzaufklärung einzelne Variablen besitzen. Damit wird sichtbar, welche der modellierten Bedingungen am einflussreichsten im Hinblick auf die Unverzerrtheit der Parameterschätzung ist.[68]

Im Gegensatz zum absoluten Bias handelt es sich bei seinem relativen Maß um ein Kriterium auf Konfigurationsebene. Der relative Bias zeigt an, wie groß die durchschnittliche Abweichung des geschätzten Parameters vom Populationswert ist. Positive Werte bedeuten, dass die geschätzten Parameter durchschnittlich über dem Populationswert liegen, was als Überschätzung der Parameter zu deuten ist, während negative Werte auf eine Unterschätzung hindeuten. An diesem Maß lässt sich direkt ablesen, wie gut die einzelnen MDTs in Relation zueinander sind. Aufgrund seiner einfachen Interpretierbarkeit kommt dieses Maß in nahezu allen MC-Studien zu MDTs zum Einsatz. Das erlaubt einen Vergleich der Ergebnisse aus dieser Arbeit und dem bisherigen Forschungsstand. Beim relativen Bias wird der durchschnittlich geschätzte Parameter *i* innerhalb einer Konfiguration *j* ($\hat{\bar{\theta}}_{ij}$) in Relation zu dessen Populationswert θ_i gesetzt und in einem Prozentwert ausgedrückt:

$$Bias_{rel} = \left(\frac{\hat{\bar{\theta}}_{ij} - \theta_i}{\theta_i}\right) * 100. \quad (6.2)$$

[68] Die Samplegröße dient in diesen Meta-Modellen als unabhängige Variable (siehe die Kapitel 7.2, 8.2, 9.2). Damit wird die Prüfung der Konsistenz der Parameterschätzung möglich (nämlich ob die MDTs bei zunehmender Samplegröße weniger verzerrte Ergebnisse produzieren).

Wie üblich werden Werte über 10 % bzw. 15 % als problematisch und interpretationswürdig angesehen (vgl. Muthén u. a. 1987: 446).

6.3.3 Parameterebene: Effizienz

Ob eine Parameterschätzung effizient ist, lässt sich mithilfe des *mean squared errors* (MSE) erfassen. Das ist die quadratische Abweichung der Schätzfunktion vom zu schätzenden Wert. Der MSE berücksichtigt in seiner Berechnung, neben dem Bias, zusätzlich die Varianz der Schätzungen über die Replikationen hinweg. Berechnet wird der MSE mit der Varianz des betreffenden Parameters *i* innerhalb einer Konfiguration *j* und dessen quadrierten Bias[69]:

$$MSE = Var(\hat{\theta}_{ij}) + (\hat{\bar{\theta}}_{ij} - \theta_i)^2. \tag{6.3}$$

Sind die Parameterschätzungen unverzerrt, dann gibt der MSE die Effizienz der Schätzung wieder. In einem solchen Fall entspricht der MSE der Stichprobenvarianz dieser Schätzungen und je kleiner diese Varianz ist, desto effizienter ist die Schätzung. Liegen dagegen verzerrte Parameterschätzungen vor, dann ist der MSE als quantitatives Maß für die Genauigkeit der Schätzung zu verstehen. In beiden Fällen gilt jedoch, dass die Schätzung besser ist, je kleiner der MSE ausfällt. Unterscheidet sich dieser nur sehr gering von null, bedeutet dies, dass sich der geschätzte Wert $\hat{\bar{\theta}}_{ij}$ nur unwesentlich vom Populationswert θ_i unterscheidet. Für sich allein betrachtet ist der MSE allerdings wenig sinnvoll. Erst wenn dieser in einem relativen Sinne zum Einsatz kommt, lassen sich gehaltvolle Aussagen treffen. Ein solches Maß liegt mit der relativen Effizienz (RE) vor. Hierbei werden die MSE-Werte zweier Schätzmethoden (hier: zweier MDTs) in Relation gesetzt, sodass sich Aussagen tätigen lassen, welche dieser Schätzmethoden unter welchen Bedingungen effizienter ist bzw. genauere Schätzungen erzielt (vgl. Carsey/Harden 2014: 89).

Im vorliegenden Fall bedeutet das, dass die MSEs zwischen einzelnen MDTs für den jeweiligen Parameter innerhalb einer Konfiguration zu vergleichen sind. Die Berechnung der relativen Effizienz erfolgt dann folgendermaßen (als Beispiel sollen FCS und Direct-ML dienen):

$$RE = \frac{MSE_{FCS}}{MSE_{FIML}}. \tag{6.4}$$

Ist dieses Verhältnis größer als eins, dann ist Direct-ML effizienter/genauer als FCS. Wenn beide Methoden gleich gut sind, dann liegt das Verhältnis bei eins. Werte, die kleiner als eins sind, zeigen eine effizientere/genauere Schätzung mit FCS an. Als interpretationswürdig wird eine Relation von kleiner .9 und größer 1.1 angesehen. Wenn kein Bias vorliegt, bedeutet eine

[69] Nicht zu verwechseln mit dem bereits vorgestellten absoluten Bias und dem relativen Bias. Wie der relative Bias auch, ist dieser Bias auf der Konfigurationsebene zu verorten. Er wird oft als Roh-Bias bezeichnet, soll hier aber keine weitere Rolle spielen, da er lediglich zur Berechnung des MSE benötigt wird.

solche Relation, dass das Sample ca. 10 % kleiner sein darf (RE = .9), um mit FCS eine gleich effiziente Schätzung zu erzielen wie mit Direct-ML, oder 10 % größer sein muss (RE = 1.1), damit mit FCS die gleiche Effizienz erzielt wird, wie mit Direct-ML. Schlussendlich gibt die relative Effizienz an, welche der beiden Schätzungen bei gleichbleibender Fallzahl weniger Varianz aufweist und damit effizienter ist (siehe Enders 2001c; Enders/Bandalos 2001).

6.3.4 Parameterebene: Standardfehlerbias

Auch die Verzerrung des Standardfehlers (SE-Bias) kann untersucht werden. Weil für die Standardfehler aber keine Populationsparameter festgelegt werden können, wird zur Berechnung des SE-Bias die durchschnittlich berechnete Standardabweichung (SD) benutzt (vgl. Muthén/Muthén 2017: 472). Dieses Maß wird auch als empirischer Standardfehler bezeichnet. Für den absoluten SE-Bias wird die Standardabweichung eines Parameters *i* innerhalb einer Konfiguration (SD_i) vom jeweils geschätzten Standardfehler $\widehat{SE}_{ij}$ für den Parameter *i* innerhalb der Konfiguration *j* abgezogen:

$$SE\text{-}Bias_{abs} = |\widehat{SE}_{ij} - SD_i|. \tag{6.5}$$

Der absolute Wert für den SE-Bias kann, wie der absolute Parameterbias, als abhängige Größe in ein lineares Regressionsmodell aufgenommen werden. Damit kann neben dem Einfluss einzelner Modellvariablen auch die Konsistenz der Standardfehler überprüft werden.

Neben der absoluten Verzerrung für den Standardfehler wird noch deren relative Ausprägung untersucht. Das erlaubt Aussagen darüber, welche MDT den Standardfehler in den einzelnen Simulationskonfigurationen eher unter- oder überschätzt. Wird der Standardfehler eher unterschätzt, dann ist die Möglichkeit eines falschen inferenzstatistischen Schlusses erhöht. Eine Überschätzung des Standardfehlers deutet eine eher konservativere inferenzstatistische Schlussfolgerung an. Der relative SE-Bias wird berechnet, indem die Differenz zwischen dem durchschnittlich geschätzten Standardfehler innerhalb einer Konfiguration ($\widehat{\overline{SE}}_{ij}$) von dessen Populationswert innerhalb der Konfiguration (SD_i) abgezogen und durch den Populationswert dividiert wird. Das Ergebnis wird wieder in Prozent ausgedrückt:

$$SE\text{-}Bias_{rel} = \left(\frac{\widehat{\overline{SE}}_{ij} - SD_i}{SD_i}\right) * 100. \tag{6.6}$$

Wie für den Parameterbias werden Werte über 10 % bzw. 15 % als problematisch und interpretationswürdig angesehen. Tabelle 6 fasst die Bewertungskriterien zusammen.

Tabelle 6: Abhängige Variablen in dieser Arbeit

Bewertungskriterium		Grenzwerte	Analysen
Modellebene			
Ablehnungsrate/ *rejection rate*	Fit-Indices pro Replikation: 1. p-Wert Chi² 2. RMSEA 3. RMSEA KI (90 %) 4. SRMR 5. CFI	 1. $\geq .05$ 2. $\leq .05$ 3. $\leq .05$ u. $\leq .10$ 4. $\leq .08$ 5. $\geq .95$	deskriptiv + log. Regressionsmodelle
Parameterebene			
Parameterbias	relativer Bias pro Konfiguration	10 % bzw. 15 %	deskriptiv
	absoluter Bias pro Replikation	--	lin. Regressionsmodelle
Relative Effizienz	Vergleich der MDTs	--	deskriptiv
Standardfehlerbias	relativer SE-Bias pro Konfiguration	10 % bzw. 15 %	deskriptiv
	absoluter SE-Bias pro Replikation	--	lin. Regressionsmodelle

6.4 Umsetzung des Forschungsdesigns

Zur Umsetzung des vorgestellten Forschungsdesigns werden zwei Softwarepakete verwendet: M*plus* in der Version 7.31 (Muthén/Muthén 2012) und R-Studio in der Version 1.1.419 (RStudio Team 2016) auf Basis der Programmiersprache R in den Versionen 3.2.2[70] und 3.4.3 (R Core Team 2015; 2017). Die notwendigen Daten mit den Charakteristiken der exogenen Modellparameter werden mithilfe von M*plus* und dem Einsatz des *MplusAutomation* Pakets (Hallquist/Wiley 2018; Hallquist 2018) generiert.[71] Mit diesem ist es möglich, den Prozess der Datengenerierung sowie die Analysen mit M*plus* zu automatisieren. Neben der Datengenerierung

[70] Die R-Version 3.4.3 dient allen notwendigen Berechnungen als Grundlage. Eine Ausnahme stellt die Implementation von MNV dar. Hierbei wird die Version 3.2.2 verwendet, da das benutzte Paket (*norm2*) zum Zeitpunkt der Simulationsdurchführungen nicht für eine neuere Version verfügbar war.

[71] Für die Generierung der Daten ist ein computergesteuerter Zufallsprozess verantwortlich, wodurch jede Replikation jeweils eine Zufallskomponente enthält. Damit weist jede Replikation Eigenheiten auf, die sie von anderen Replikationen zufällig unterscheiden. Diese Eigenheiten gleichen sich im Mittel zwar aus, wenn viele Replikationen gezogen werden, allerdings wird durch die Zufallskomponente in den Ziehungen die Wiederholbarkeit einer MC-Studie eingeschränkt: Wird bei gleichem Design eine erneute Simulation mit einer gleichen Anzahl an Replikationen durchgeführt, kann es sein, dass sich die Ergebnisse zwischen den Simulationen unterscheiden. Um die Wiederholbarkeit sicherzustellen, muss dem Zufallsprozess deshalb ein numerischer Startwert (Seed) zugeschrieben werden, was dazu führt, dass bei einer wiederholten Simulation die Reihenfolge der Zufallszahlen identisch bleibt. Das stellt nicht nur die Wiederholbarkeit sicher, sondern auch die Nachvollziehbarkeit der Ergebnisse. Aus diesem Grund werden den drei Populationsmodellen willkürlich gewählte Seeds vorangestellt. Da auch die MI-Varianten mit Zufallsziehungen arbeiten und sich die Imputationen bei einer Wiederholung des Imputationsprozesses unterscheiden, wenn kein Seed vorangestellt wird, werden auch diesen willkürliche Seeds zugewiesen (siehe Anhang O1). Durch diese Maßnahmen kann die Wiederholbarkeit der vorliegenden Arbeit sichergestellt werden (vgl. Krause 2019: 83). Ein weiteres Problem des Zufallsprozesses bei einer MC-Studie ist der MC-Fehler. Denn trotz eines gleichbleibenden Seeds wird sich bei der Variation einer Konfigurationsgröße auch der Zufallsprozess unterscheiden. Würde demnach eine Replikation einer Konfiguration (bspw. große Fallzahl) mit einer Replikation einer anderen Konfiguration verglichen werden (bspw. kleine Fallzahl), dann könnte nicht ausgeschlossen werden, ob die unterschiedlichen Ergebnisse in den Replikationen aufgrund der variierten Fallzahl zustande gekommen sind, oder ob dafür die Zufallskomponente verantwortlich ist. Aus diesem Grund werden auch viele Replikationen notwendig, denn darin wird die Zufallskomponente im Mittel ausgeglichen, was zu einer Reduktion des MC-Fehlers führt, ihn allerdings nicht gänzlich reduziert.

werden zudem die Behandlung der fehlenden Werte mit Direct-ML, respektive die Ersetzung der Missings mittels H0 in M*plus* durchgeführt. Für die anderen MDTs (EM, EMB, FCS, MNV und PMM) werden verschiedene R-Pakete verwendet (siehe Kapitel 6.4.2). Alle Modellschätzungen werden wiederum in M*plus* getätigt.[72] Die Aufbereitung der Analyseergebnisse, deren Auswertung und Visualisierung erfolgt dann ausschließlich in R-Studio mit ausgewählten Paketen.[73] Weil die Performanz einzelner MDTs im Fokus steht, werden für jede der MDTs identische Datensätze verwendet. Es werden 500 verschiedene Datensätze für die 54 Konfigurationen erstellt, die dann versiebenfacht werden und damit die berechneten 378 Simulationskonfigurationen abbilden.

6.4.1 Implementation von MAR

MAR fußt auf der Annahme, dass die Missing Values in einer Variable *Y* von anderen Modellvariablen *X* abhängig sind, nicht aber von der Ausprägung in *Y* selbst. In M*plus* wird die Generierung von MAR mithilfe logistischer Regressionsmodelle erreicht. Dazu wird für jede Variable, die Missing Values aufweisen soll, ein eigenes logistisches Regressionsmodell spezifiziert, bei der die abhängige Variable eine Indikatorvariable ist, die anzeigt, ob ein Wert in der eigentlichen Variablen fehlend sein soll oder nicht. Mithilfe der Ausprägungen in der Indikatorvariablen wird dann der Wert in der eigentlichen Variablen gelöscht (vgl. Muthén/Muthén 2012: 422).[74] Bei diesen Modellen handelt es sich um bivariate logistische Regressionen, bei welchen eine unabhängige Variable für die Missings verantwortlich ist. Für Modell 1 werden z. B. auf dem dritten Indikator (ind3) Missing Values eingefügt. Die entsprechende logistische Regression sieht in diesem Falle folgendermaßen aus:

$$ind3_I = intercept + slope * ind1. \quad (6.7)$$

Die Variable $ind3_I$ ist die Indikatorvariable der eigentlichen Variablen (ind3) und ind1 ist diejenige Variable, auf welche die Missing Values zurückzuführen sind. Da die Missing Values in $ind3_I$ (und damit auch in ind3) durch ind1 verursacht werden, nicht aber von der Ausprägung in ind3 abhängig sind, unterliegen diese dem MAR-Ausfallmechanismus. Durch die Variation

[72] Die *templates* zur Erstellung der M*plus*-Files (zur Datengenerierung), die M*plus*-Files für die Anwendung von Direct-ML und H0 sowie die Files zur Modellanalyse finden sich im Anhang O1.

[73] Zu den wichtigsten gehören unter anderem: *ggplot2* (Version 3.1.0; Erstellung Grafiken), *stargazer* und *flextable* (Version 5.2.2 respektive Version 0.4.4; Erstellung Tabellen), *brglm* (Version 0.6.1; Schätzung logistischer Regressionen mit PML) und *mfx* (Version 1.1; Berechnung der AMEs).

[74] Das kann an folgendem Beispiel verdeutlicht werden: Die erste Spalte eines hypothetischen Datensatzes enthält die Ausprägungen der eigentlichen Variablen (bspw. eine 5er Skala), die zweite Spalte enthält die Ausprägung der Indikatorvariablen (0 = nicht Missing, 1 = Missing). Im Folgenden werden aus den Zeilen in der ersten Spalte diejenigen Fälle gelöscht (Missing gesetzt), bei denen die Indikatorvariable den Wert für Missing enthält.

des *intercepts* und des *slopes* lässt sich mittels dieser Regressionen der Anteil an Missing Values in der jeweiligen Indikatorvariablen steuern. Je nachdem, wie hoch dieser gemäß den festgelegten Anteilen sein soll, wird die Ausprägung beider Komponenten angepasst (*templates* im Anhang O1). Die Missing Values werden durch die entsprechenden Regressionen wie folgt generiert:

- Modell 1 hat auf den Indikatoren drei bis fünf Missing Values (ind3 bis ind5).
- Modell 2 hat auf der abhängigen Variablen y1 und auf den unabhängigen Variablen x3 und x4 Missing Values.
- Modell 3 hat auf allen Indikatoren (ind1 – ind5) und auf den unabhängigen Variablen x3 und x4 Missing Values.
- Als verursachende Variablen dienen: ind1 und ind2 in Modell 1; x1 in Modell 2; x1 und x2 in Modell 3.

6.4.2 Konfiguration der Missing Data Techniken

Der Analysestandard für M*plus* bei fehlenden Werten ist Direct-ML mit dem robusten ML-Schätzer (MLR). Da bei Direct-ML die Behandlung der fehlenden Werte und die Modellanalyse in einem Schritt durchgeführt werden, entsprechen die Modelle zur Behandlung der fehlenden Werte gleichzeitig den Analysemodellen. Das gilt auch für H0. Hierbei entspricht das IM dem Analysemodell und beide Schätzungen (sowohl die Imputations- als auch die Analysephase) werden in M*plus* durchgeführt. Im Gegensatz zu Direct-ML werden allerdings bei H0 die diskreten Variablen als solche klassifiziert (die binär und ordinal skalierten Variablen werden als solche definiert). Die 5er skalierten Variablen werden wie bei Direct-ML als quasi-metrische behandelt. Da bei der Spezifikation der Modelle für H0 Annahmen über die Exogenität und Endogenität einzelner Variablen getroffen werden, sind sowohl die binäre (x3), als auch die ordinale (x4) Variable im IM exogene Variablen.

Die Imputation der fehlenden Werte mit den anderen MDTs erfolgt mit verschiedenen R-Paketen. MNV wird mithilfe des *norm2*-Pakets umgesetzt (Version 2.0.1) (Schafer 2016) und EM sowie EMB mit *Amelia* (Version 1.7.4) (Honaker u. a. 2011). Für FCS wird das Paket *mice* in der Version 2.46.0 eingesetzt (van Buuren/Groothuis-Oudshoorn 2011). Wie bei H0 werden die Variablen x3 und x4 als diskrete Variablen behandelt. Die Imputation der fehlenden Werte in x3 (die binäre Variable) erfolgt anhand binär logistischer Regressionen, in x4 (die ordinale Variable) mit multinomial logistischen Regressionen und in den quasi-metrischen Variablen

mit linearen Regressionen. Auch für PMM wird *mice* eingesetzt; alle fehlenden Werte auf allen Variablen werden mit PMM imputiert (vgl. ebd.: 16 ff.).[75]

Potentiell hätte es für FCS weitere Spezifikationsmöglichkeiten für die diskreten Variablen gegeben (wie *random forests* oder *classification and regressions trees*), auch wären andere Kombinationen möglich gewesen (z. B. eine Kombination aus den logistischen Regressionen und PMM). Allerdings gibt die herausgearbeitete Forschungslücke einerseits klare Vorgaben, wonach PMM und die linearen Regressionen von FCS bei quasi-metrischen Variablen und wonach PMM bei diskreten Variablen noch nicht untersucht wurde, die mit diesen Spezifikationen abgedeckt werden können. Andererseits zeigt sie, dass kaum Untersuchungen von FCS oder PMM bei gemischten Daten und im SEM-Kontext erfolgt sind, weshalb es angebracht erscheint, möglichst Spezifikationen zu wählen, die für die potentiellen Anwendenden leicht zugänglich sind und im Vorhinein nachvollziehbar erscheinen. Das sollte mit dem Einbezug der logistischen Regressionen anstatt der *machine learning*-Verfahren geschehen sein. Mit den gewählten Spezifikationen kann demnach nicht nur eine Reihe an Punkten aus der Forschungslücke aufgegriffen werden, es wird damit auch die Anschlussfähigkeit dieser Arbeit sichergestellt.

6.4.2.1 Anzahl an zu imputierenden Datensätzen (*m*)

Kapitel 3.1 befasste sich mit den Voraussetzungen, die allen Varianten der MI zugrunde liegen. Darin wurde gezeigt, dass der Anteil an Missing Values in etwa die Anzahl an Datensätzen (*m*) bestimmen sollte, die notwendig werden, damit aus Analysen auf Datensätzen, die mit der MI imputiert werden, verlässliche statistische Inferenz hervorgeht. Das Problem mit dieser Empfehlung ist allerdings, dass der Anteil an Missing Values sehr stark davon abhängt, wie hoch die Anzahl an Variablen im IM ist. Werden neben den zu imputierenden Variablen ausnahmslos Variablen in das IM aufgenommen, die keine fehlenden Werte aufweisen, so tendiert deren Gesamtanteil gegen null (vgl. Lall 2016: 426). Dementsprechend unterscheiden sich auch die Gesamtanteile in dieser Arbeit, da in den einzelnen Modellen unterschiedlich viele Variablen vorhanden sind und die Relation zwischen vollständigen und unvollständigen Variablen jeweils eine andere ist. Für alle Modelle wird dadurch eine andere Anzahl an zu imputierenden Datensätzen notwendig. Lässt man allerdings die Variablen ohne Missings außen vor, können alle Modelle designtechnisch nur eine maximale Quote von 35 % erreichen (die Konfiguration mit dem höchsten Missinganteil).

[75] Dem Anhang O1 lassen sich die genauen Konfigurationen entnehmen.

Um dennoch auszuschließen, dass die Ergebnisse aufgrund einer zu geringen Anzahl an Datensätzen beeinträchtigt werden und weil Graham u. a. (2007) zeigen, dass ab einem *m* von 40 die Ergebnisse im Hinblick auf Effizienz und Teststärke relativ stabil sind, wird ein *m* von 50 für jede Konfiguration festgelegt. Diese Anzahl sollte auch aus dem Grund ausreichend sein, als dass bekannt ist, welche Variablen für die Missing Values verantwortlich sind und sich diese in den IMs befinden. Demnach sind Variablen vorhanden, die mit den Variablen, auf denen Missings vorhanden sind, in Beziehung stehen. Das dürfte in einer Reduktion der FMI resultieren, was ein *m* von 50 angemessen erscheinen lässt.

6.4.2.2 Konvergenz der Verfahren

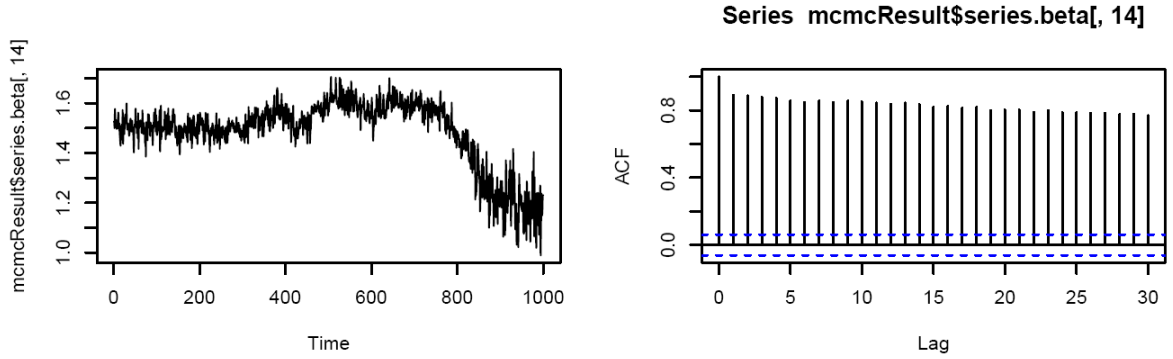
Da es sich bei allen vorgestellten und verwendeten Imputationsmethoden um iterative Verfahren handelt, muss deren Konvergenz sichergestellt werden. Zur Erfassung derselben gibt es zwei Möglichkeiten: Zum einen kann die Konvergenz mithilfe von Grafiken erfasst werden und zum anderen mit dem *potential scale reduction factor* (PSR).[76]

Im Folgenden soll an MNV veranschaulicht werden, was mit der Konvergenz dieser Verfahren gemeint ist. Für die anderen MDTs stehen ähnliche Prüfmöglichkeiten zur Verfügung, es sei an dieser Stelle aber an die Literatur verwiesen (FCS/PMM: van Buuren/Groothuis-Oudshoorn 2011; van Buuren/Vink 2018; EM/EMB: Honaker u. a. 2012).

Für MNV wird die Konvergenz zur gewünschten posterior-Verteilung durch Autokorrelations- und Zeitreihen-Plots erfasst (vgl. Schafer 2016: 29 ff.). Eine Bedingung für die Plausibilität der Imputationen ist, dass diese aus Iterationen hervorgehen, die voneinander unabhängig sind. Ob dies der Fall ist, erfolgt mit den angesprochenen Autokorrelationsplots. Der rechte Plot in Abbildung 11 zeigt, dass die Korrelation zwischen der ersten Iteration und Iteration 30 sehr hoch ist. Werden demnach der erste und der 30ste Datensatz ausgewählt, dann korrelieren die imputieren Werte sehr stark miteinander und die Unabhängigkeit der Imputationen ist nicht gegeben. In diesem Fall müssen zwischen den auszuwählenden Datensätzen mehr Iterationen liegen. Für die vorliegende Arbeit ist die Abhängigkeit der Imputationen verschiedener Iterationen allerdings unproblematisch, da für die MI-Verfahren jeweils 50 parallele MCMC-Ketten generiert werden (außer für H0). Aus diesen wird dann jeweils der Datensatz der letzten Iteration ausgespeichert. Somit ist nach Vorliegen der Konvergenz auch automatisch die Unabhängigkeit der Iterationen sichergestellt (dafür wird aber mehr Rechenleistung notwendig).

[76] Siehe dazu auch: Brooks/Roberts (1997); Brooks/Gelman (1998); Gelman/Rubin (1992); Sinharay (2003).

Abbildung 11: Erfassung Konvergenz für MNV (Beispiel)



Quelle: Schafer (2016: 43).

Demzufolge ist für die MI-Verfahren lediglich sicherzustellen, dass die Kette stabil ist; bei MNV geschieht dies mit den Zeitreihen-Plots (Abbildung 11, links). Hierbei sollte über den Iterationsverlauf keine Systematik zu verzeichnen sein. Das heißt, der Zeitreihen-Plot sollte sich nicht in eine positive oder negative Richtung entwickeln. Abbildung 11 zeigt beispielhaft eine Verletzung an: Hier hat sich die Kette (noch) nicht stabilisiert und die Iterationsanzahl ist zu erhöhen. Liegt eine solche Systematik bei späteren Iterationen nicht mehr vor, ist die MCMC-Kette konvergiert.

Während die Prüfung auf Konvergenz bei EM, EMB FCS und PMM ähnlich verläuft wie bei MNV, wird die Konvergenz bei H0 durch den PSR erfasst. Der PSR misst das Verhältnis zwischen between-chain und within-chain Variation. Die Konvergenz der Ketten ist gegeben, wenn die between-chain Variation im Vergleich zur within-chain Variation nicht größer ist als eins (vgl. Muthén/Muthén 2012: 335). Ist das der Fall, werden nach jeder 100sten Iteration der ersten Kette (um die Unabhängigkeit der Iterationen zu gewährleisten), die imputierten Datensätze ausgespeichert, sodass am Ende die gewünschte Anzahl an Datensätzen vorliegt. Damit gehen bei H0 die *m* Datensätze aus einer einzigen Kette hervor.

Die Prüfung der Konvergenz der MDTs kann in der vorliegenden Arbeit allerdings nicht durchgeführt werden, da die Anzahl an Überprüfungen sehr hoch ist.[77] Um auf eine Überprüfung verzichten zu können, werden deshalb zunächst Testdatensätze der jeweils komplexesten Simulationskonfiguration eines Modells generiert (Modell 1 bis Modell 3 bei einer Fallzahl

[77] Zwar variieren die durchzuführenden Prüfungen einzelner MI-Varianten, um allerdings eine grobe Einschätzung zu ermöglichen, wie viele Überprüfungen durchzuführen wären, wird dies an MNV veranschaulicht. Für MNV müsste *jede* Kette mit mehreren Plots geprüft werden. Für eine Replikation wären das 50 Files, die mehrere Plots enthalten. Insgesamt müssten für *eine* Konfiguration dann 25.000 solcher Files geprüft werden. Wird dies auf alle Konfigurationen erweitert, dann wären insgesamt 1.350.000 Files zu prüfen. Da diese Zahl mit den weiteren MDTs anwächst, ist eine Prüfung aller Plots ausgeschlossen.

von 750, einem Anteil an Missing Values von 35 % und einer stark asymmetrischen Verteilung). Aus diesen insgesamt 1500 Datensätzen (für jedes Modell werden 500 Datensätze bzw. Replikationen generiert), werden sechs Datensätze (für jedes der drei Modelle zwei Datensätze) zufällig ausgewählt, auf welchen dann die MDTs angewendet werden.[78]

Zunächst werden für die MDTs die Standardwerte für die Iterationsanzahlen verwendet. Diese zeigen sich für EM und EMB zur Konvergenzerzielung als ausreichend. Deutet aber bereits die Literatur an, dass die bisherigen Vorgaben überholt sind, werden die Iterationszahlen an den aktuell geltenden Standard angepasst. Das ist der Fall bei FCS und PMM, wobei anstatt der standardmäßigen fünf Iterationen, 20 Iterationen verwendet werden. Mit dieser Anzahl kann in den Testdatensätzen Konvergenz erzielt werden (vgl. van Buuren 2012: 113 f.; 2015: 279 f.). Für MNV und teilweise für H0 kann weder mit der standardmäßigen Iterationszahl Konvergenz erzielt werden, noch liegen aktuellere Richtlinien vor, wie viele Iterationen verwendet werden sollten. Für diese beiden Verfahren wird in den Testdatensätzen deshalb die Iterationszahl so lange erhöht bis sich die Ketten stabilisieren. Beim Einsatz von MNV können in den ersten 3000 Iterationen einige Systematiken beobachtet werden (per Default werden 1000 Iterationen vorgeschlagen; vgl. Schafer 2016: 27), die darauf hindeuten, dass die Ketten instabil sind. Ab 5000 Iterationen zeigt sich dann aber eine Stabilisation der Ketten. Bei H0 werden insgesamt 50.000 Iterationen als Default angelegt (vgl. Muthén/Muthén 2017: 656). Für das erste Modell zeigen sich 25.000 Iterationen als ausreichend, für das zweite Modell kann mittels der Default-Variante Konvergenz erzielt werden, und für das dritte Modell sind 75.000 Iterationen notwendig, da auch hierbei einige Schwankungen im PSR bis hin zu hohen Iterationszahlen auszumachen sind.

Zuletzt ist noch die Konvergenz von Direct-ML erwähnt, weil es sich dabei auch um ein iteratives Verfahren handelt. Für Direct-ML erweisen sich die Standardwerte als ausreichend[79] (maximale Anzahl an Iterationen: 1000; Toleranzkriterium: .00005; vgl. ebd.: 697).

Da diese Kriterien für die komplexeste Konfiguration in den Testdatensätzen ausreichend sind, werden diese in den eigentlichen Simulationen für alle Konfigurationen angelegt (somit auch für die weniger komplexen Konfigurationen). Damit kann die Konvergenz der Verfahren für alle Konfigurationen sichergestellt werden, da einerseits durch die Testdatensätze Nachweise vorliegen, dass diese ausreichen und andererseits davon ausgegangen werden kann, dass

[78] Die Testläufe werden auf einem Standard-Desktoprechner durchgeführt (Windows 7 64-Bit, Intel Core i5-750, 2.76 GHz, 16 GB RAM).
[79] Auch die späteren Modellanalysen werden mit diesen Werten berechnet.

die Konvergenz bei weniger komplexen Konfigurationen früher, aber zumindest nicht später eintritt. Tabelle 7 fasst die Konvergenzkriterien zusammen.

Tabelle 7: Iterationszahlen der MDTs

MDT	**Komplexeste Simulationskonfiguration** N = 750; Missing = 35 %; stark asymmetrische Verteilung		
	Modell 1	Modell 2	Modell 3
Direct-ML	Iterationen: 1000; Toleranzkriterium: 0.00005[80]		
EM / EMB	Toleranzkriterium: 0.0001		
FCS / PMM	20		
H0	25.000	50.000	75.000
MNV	5000		

6.4.2.3 bwUniCluster

Für die verschiedenen MI-Techniken müssen für jede Replikation innerhalb einer Konfiguration, fehlende Werte in 50 Datensätzen (*m*) ersetzt werden. Teilweise ist auch die Anzahl an Iterationen, im Gegensatz zu den Standardwerten, erhöht. Beides beeinflusst die Dauer der Simulationsläufe für die einzelnen Konfigurationen. Bei der Durchführung der Imputationen auf den Testdatensätzen benötigte MNV durchschnittlich ca. 80 Sekunden, bis die Missing Values in einem der sechs zugrundeliegenden Datensätze ersetzt waren und die 50 imputierten Datensätze vorlagen (FCS benötigte ca. 35 Sekunden; H0 ca. 65 Sekunden; PMM ca. 35 Sekunden; die Dauer für Direct-ML, EM und EMB war zu vernachlässigen). Mit diesen vier Techniken dauert es also im Durchschnitt ca. 55 Sekunden bis die fehlenden Werte imputiert sind. Rein rechnerisch ergibt sich dadurch eine Gesamtdauer von knapp 70 Tagen für die Simulationsläufe aller Konfigurationen dieser vier Techniken, auf dem, für die Testläufe verwendeten, Standard-Desktoprechner.[81] Werden bei dieser Berechnung noch die anderen drei Techniken berücksichtigt, sowie die Tatsache, dass dann lediglich imputierte Daten aber noch keine Modellschätzungen und auch keine Analyseergebnisse vorliegen, erhöht sich die rechnerisch benötigte Anzahl der Tage nochmals.

Aus diesem Grund werden die Simulationsläufe der MI-Techniken auf dem bwUniCluster durchgeführt. Hierbei handelt es sich um einen Parallelrechner, der vom *Steinbuch Centre for*

[80] Das Toleranzkriterium ist ein Abbruchkriterium für die Iterationsketten: Wenn die maximale relative Änderung aller Parameter von einer Iteration zur nächsten unter diesen Wert fällt, dann ist ML zur maximalen Likelihood konvergiert (vgl. Schafer/Novo 2015: 5).

[81] 55 Sekunden*500 Replikationen*54 Konfigurationen*4 MDTs.

Comuting (SCC) im Rahmen des baden-württembergischen Umsetzungskonzepts für Hochleistungsrechnen in Betrieb genommenen wurde. Damit können gleichzeitig bis zu 50 Konfigurationen gerechnet werden, was eine erhebliche Reduktion der Rechenzeit zur Folge hat.[82]

6.4.3 Modellschätzer der Analysemodelle und Referenzmodell

Nachdem die fehlenden Werte mit den MDTs behandelt wurden, können die eigentlichen Modellschätzungen getätigt werden. Aus Abweichungen dieser Modellschätzungen zu den Populationsmodellen, kann dann mithilfe der Bewertungskriterien erfasst werden, wie gut die fehlenden Werte durch die MDTs gehandhabt werden. Wichtig ist dabei, dass diese Modellschätzungen mit ein und demselben Modellschätzer erfolgen. Damit kann sichergestellt werden, dass Unterschiede in den Analyseergebnissen, nicht durch unterschiedliche Modellschätzer verursacht werden, sondern auf die jeweiligen MDTs zurückgehen. Im vorliegenden Fall wird dafür der MLR-Schätzer verwendet. Mit diesem lässt sich einerseits auf nicht normalverteilte Daten reagieren, indem die Verteilungen der Variablen berücksichtigt werden, um neben den Standardfehlern auch die Chi²-Statistik zu korrigieren (vgl. Muthén/Muthén 2012: 603). Andererseits handelt es sich dabei um den standardmäßigen Modellschätzer von M*plus*, sofern Missing Values vorliegen, sodass dieser in der Empirie wohl sehr häufig zum Einsatz kommt. Auch führt dies dazu, dass für Direct-ML keine alternativen Syntaxfiles notwendig werden.

Die Auswahl des MLR-Schätzverfahrens für die Modellschätzungen auf den imputierten Daten und auf den Daten mit fehlenden Werten (Direct-ML), kann allerdings zu dem Problem führen, dass damit verzerrte Schätzergebnisse einhergehen. Wenn also die Modellschätzungen auf den imputierten Daten, oder den Daten mit Missings, Ergebnisverzerrungen aufweisen, kann nicht unterschieden werden, ob diese Verzerrungen durch den Einsatz der MDTs zustande kommen, oder ob diese durch den MLR-Schätzer verursacht werden. Um diesem Problem zu entgehen, werden zusätzlich zu den oben aufgeführten Simulationskonfigurationen, zusätzliche Konfigurationen gerechnet, in welchen keine Missing Values vorliegen.[83] Diese Konfigurationen dienen für die Direct-ML-Analysen und für die Analysen auf den imputierten Datensätzen als Referenzen. Die Annahme dahinter ist Folgende: Sollten sich für die Konfigurationen, bei denen keine Missing Values vorliegen, Verzerrungen in den Ergebnissen zeigen, so stammen diese von dem gewählten MLR-Schätzer; damit wird bekannt, ob und wie der MLR-Schätzer

[82] Für das bwUniCluster siehe https://www.scc.kit.edu/dienste/bwUniCluster.php. Für die interessierte Leserschaft findet sich im Anhang A2 eine ausführliche Beschreibung des Simulationsvorgangs auf dem Cluster.
[83] Zu den bereits vorliegenden 378 Simulationskonfigurationen kommen in diesem Fall noch einmal zusätzliche 18 Konfigurationen hinzu (3 Modelle*2 Samplegrößen*3 Verteilungen*1 Missingquote). Damit erhöht sich die Gesamtzahl auf 396.

das Schätzergebnis verzerrt. Werden im Anschluss daran die eigentlich interessierenden Analysen zu den einzelnen MDTs gerechnet, dann lässt sich auch hier für jede Konfiguration berechnen, ob es zu Verzerrungen im Schätzergebnis gekommen ist oder nicht. Wie bereits angeführt, weisen diese Ergebnisse aber nicht nur den Einfluss der MDTs auf, sondern eben auch den Einfluss des Modellschätzers. Da der Einfluss des Letzteren aber aufgrund der Analysen auf Daten ohne Missing Values bekannt ist, kann das Ergebnis zu den einzelnen MDTs dazu in Bezug gesetzt und um den Einfluss des MLR-Schätzers bereinigt werden. Letztlich wird so der Einfluss der MDTs auf das Schätzergebnis unter den modellierten Bedingungen sichtbar. Sollten die MDTs jedoch von den modellierten Bedingungen unabhängig sein, dann sollten sich die Ergebnisse zwischen den MDT-Analysen zu denjenigen auf vollständigen Daten nur geringfügig unterscheiden. In diesem Fall wird die ursprüngliche Datenstruktur wiederhergestellt und die beobachteten Verzerrungen entstehen durch den gewählten Modellschätzer.

7 Ergebnisse Modellebene

Durch die Sicherstellung der Konvergenz der imputationsbasierten MDTs wird gewährleistet, dass immer Datensätze vorliegen, die analysiert werden können. Allerdings bedeutet dies nicht, dass die Modellschätzungen mittels MLR-Schätzer auf den imputierten Daten (bzw. mit Direct-ML auf den Daten mit fehlenden Werten) erfolgreich konvergieren. Demzufolge ist es möglich, dass die Modellschätzungen keine oder unplausible Lösungen aufweisen. Da beides mit inakzeptablen Schätzergebnissen gleichzusetzen ist, werden sie aus den Analysen ausgeschlossen. Bevor demnach die Ergebnisse analysiert werden können, müssen zunächst gültige und nichtgültige Modellschätzungen unterschieden werden.

Eine Modellschätzung wird als gültig aufgefasst, wenn diese erfolgreich konvergiert ist *und* dabei keine unplausible Lösung (*improper solution*, oftmals auch als Heywood-Fall bezeichnet) aufweist. Als unplausible Lösung ist eine konvergierte Schätzung zu verstehen, die negative Varianzen schätzt und/oder deren Werte für einzelne Zusammenhänge größer als ± 1.0 sind (vgl. Chen u. a. 2001: 469 f.; Gerbing/Anderson 2016: 99 f.). Die Summe der gültigen Lösungen innerhalb einer Simulationskonfiguration stellt die Fallzahl der jeweiligen Konfiguration für die nachfolgenden Analysen dar, während die Relation dieser Summe zur vorgesehenen Anzahl an Replikationen, als deren Konvergenzrate bezeichnet wird (der Anteil an gültigen Modellschätzungen einer jeweiligen Konfiguration in Prozent).

Bis auf eine einzige Replikation, liegen gültige Modellschätzungen vor. Damit weist, mit Ausnahme einer Konfiguration für EMB (siehe Anhang A3), jede Simulationskonfiguration 500 Fälle auf. Diese Fallzahl gilt für alle folgenden Analysen und Ergebnisdarstellungen, es sei

denn, eine Fallzahl wird explizit genannt. Anzumerken ist, dass das Vorliegen von gültigen Modellschätzungen keine inhaltliche Bedeutsamkeit hat. Es kann nur sichergestellt werden, dass in der Analyse keine unplausiblen Lösungen vorhanden sind. Wie diese Modellschätzungen zu bewerten sind, und ob die Handhabung der Missing Values mit den MDTs zufriedenstellend ist, lässt sich daran nicht ablesen.

7.1 Deskriptive Analyse der Fit-Indices: Ablehnungsraten

In diesem Kapitel werden die Ablehnungsraten der Fit-Indices deskriptiv aufgearbeitet. Da das zweite Modell ein voll saturiertes ist, stehen dafür auch keine Fit-Indices zur Analyse bereit. Aus diesem Grund wird dieses Modell für die Analyse der Fit-Indices nicht betrachtet. Weiterhin werden auch lediglich die Ergebnisse zum dritten Modell berichtet, da dessen Ergebnisse repräsentativ für das erste Modell sind. Das gilt zwar nicht für die absolute Höhe der Ablehnungsraten, aber in Bezug auf die Einflussgrößen.[84] Weil die Ablehnungsraten der Fit-Indices die Bewertungskriterien der MDTs darstellen, attestieren zufriedenstellende Ablehnungsraten den MDTs eine gute Performanz, wohingegen nicht zufriedenstellende Raten eine schlechte Performanz implizieren.

Bevor die Analyse der einzelnen Fit-Indices unter Vorliegen von Missing Values durchgeführt wird, sei zunächst auf das Referenzmodell verwiesen (Tabelle 8). Die Ablehnungsraten des p-Wertes liegen bei ca. 5 %. Im Gegensatz dazu wird anhand des SRMR, des 90 %igen Konfidenzintervalls von RMSEA (im Weiteren auch lediglich Konfidenzintervall) und mit dem CFI kein bzw. kaum ein Modell verworfen. RMSEA liefert unter kleinen Fallzahlen etwaige Modellablehnungen (ca. 5 %), für größere Fallzahlen werden keine Modelle verworfen. Im Hinblick darauf, dass mit 5 % an Modellablehnungen zufriedenstellende Ergebnisse vorliegen, wird mittels der Fit-Indices (mit Ausnahme des p-Wertes) das Modell zu oft akzeptiert. Eine Tendenz, wonach die Modellablehnungen mit abnehmender Samplegröße (außer bei RMSEA), oder zunehmender Asymmetrie größer werden, lässt sich nicht beobachten. Alle Ablehnungsraten der Fit-Indices nach Einsatz der MDTs, die über die Werte in Tabelle 8 hinausgehen, können auf die MDTs zurückgeführt werden, wohingegen gleichbleibende Raten den Einfluss der MLR-Schätzung darstellen. Aus diesem Grund gilt auch, dass die MDTs zuverlässig arbeiten, wenn die Raten des Referenzmodells erreicht werden oder wenn nicht nochmals zusätzliche 5 % an Modellablehnungen hinzukommen. Denn das bedeutet, dass die MDTs die ursprüngli-

[84] Eine unkommentierte Ergebnisdokumentation zum ersten Modell lässt sich dem Anhang O2.1 entnehmen.

che Datenstruktur so wiederherstellen können, dass damit korrekte Modellbewertungen einhergehen. Weil der MLR-Schätzer zufriedenstellende Ergebnisse liefert, werden die Ablehnungsraten der MDTs nicht um dessen Einfluss bereinigt (wie es beim Parameterbias der Fall ist; siehe Kapitel 8). Das heißt, dass alle Ergebnisse zu den Ablehnungsraten den vernachlässigbaren Einfluss der MLR-Schätzung und den Einfluss der MDTs wiedergeben.

Tabelle 8: Ablehnungsraten des Referenzmodells in %

Modellkonfiguration			p-Wert	RMSEA	RMSEA KI (90 %)	SRMR	CFI
750	skew1	mar1	4.2	0.0	0.0	0.0	0.0
	skew2		3.6	0.0	0.0	0.0	0.0
	skew3		4.4	0.0	0.0	0.0	0.0
250	skew1		6.0	4.0	0.0	0.0	0.0
	skew2		6.8	5.0	0.0	0.0	0.0
	skew3		8.2	6.2	0.0	0.0	0.4

Anmerkungen: Modellkonfiguration: Samplegröße, Verteilungen (skew1 = symmetrisch; skew2 = asymmetrisch; skew3 = stark asymmetrisch), Missinganteile (mar1 = 0 %; mar2 = 5 %; mar3 = 20 %; mar4 = 35 %).

Folgende zwei Aspekte werden an dieser Stelle relevant. Zum einen handelt es sich bei der Analyse der Ablehnungsraten um eine explorative Analyse, weil nur wenig Literatur vorliegt und auch keine Hypothesen generiert wurden. Es werden deshalb zunächst die Übersichtstabellen aller Konfigurationsergebnisse analysiert (Tabellen im Anhang A4.1). Daraus kann folgendes entnommen werden: Erstens produzieren die Fit-Indices in der Konfiguration mit wenigen Missing Values, einer großen Fallzahl und einer symmetrischen Variablenverteilung zufriedenstellende Ablehnungsraten. Diese Konfiguration ist damit für die Performanz der MDTs wenig bedeutsam. Für den p-Wert liegen in dieser Konfiguration aber bereits deutlich höhere, nicht mehr zufriedenstellende Ablehnungsraten vor. Zweitens zeigt sich die Konfiguration mit einer kleinen Fallzahl, stark asymmetrischen Verteilungen und sehr hohen Missinganteilen als äußerst problematisch. In vielen Fällen liegen die Ablehnungsraten bei über 90 %.

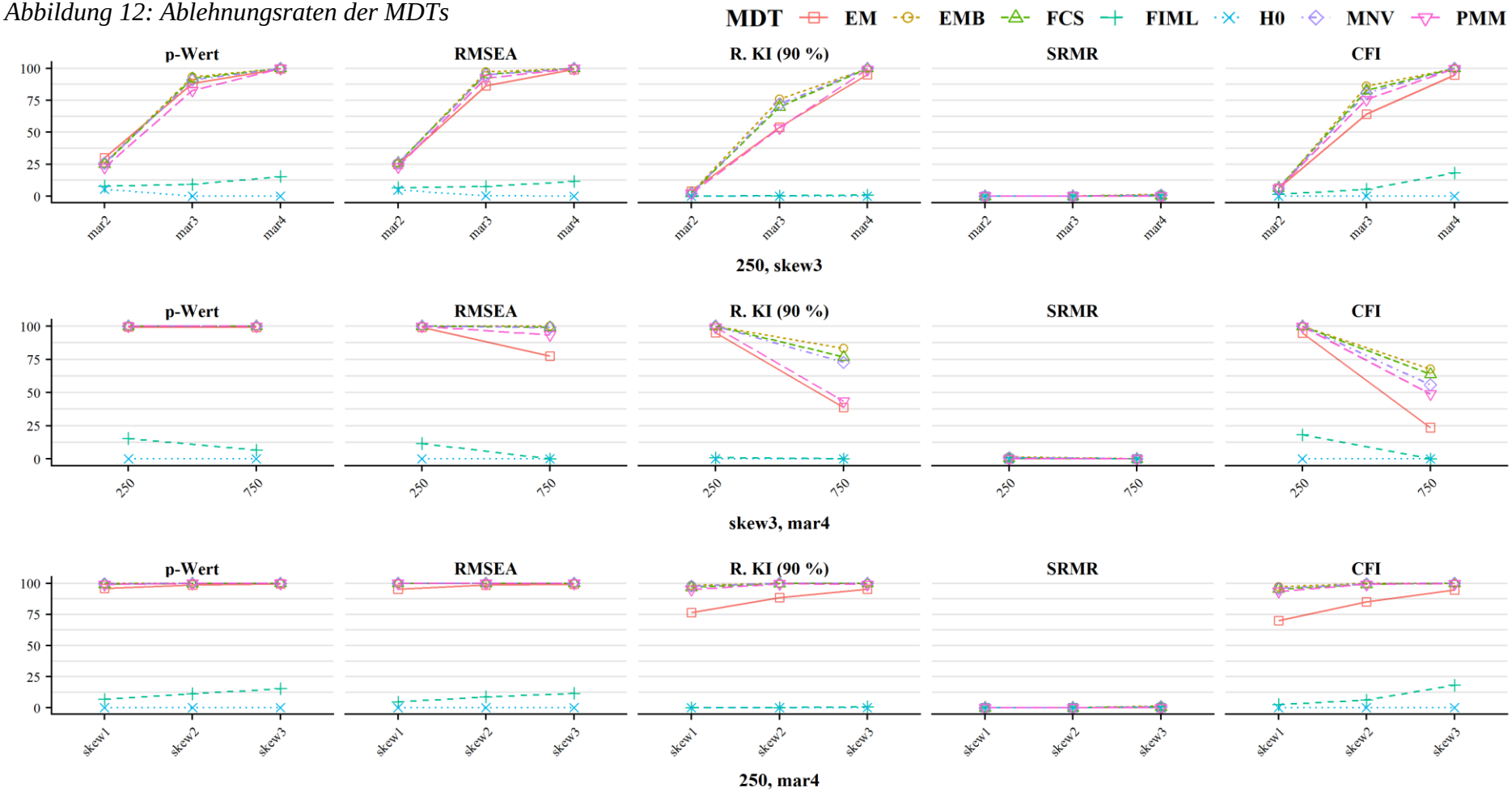
Zum anderen erfolgt die Ergebnisauswertung auf drei verschiedenen Ebenen, um die aufgestellten Fragen zu den Fit-Indices beantworten zu können (siehe Kapitel 5.5). Zunächst steht dabei die Performanz der MDTs zueinander im Fokus. Fraglich ist, ob sich für die MDTs ähnliche Ergebnisse ergeben, oder ob sich MDTs identifizieren lassen, die bessere/schlechtere Ablehnungsraten generieren. Danach ist von Interesse, ob es Unterschiede in den Ergebnissen der untersuchten Fit-Indices gibt. Es steht die Frage im Vordergrund, welche Fit-Indices sich besonders gut eignen, SE-Modelle zu bewerten, wenn Missing Values vorliegen. Zuletzt sollen auch diejenigen Simulationsgrößen identifiziert werden, die für den jeweiligen Fit-Index besonders einflussreich sind. Damit eine solche Aussage möglich wird, müssen die Effekte der einzelnen Simulationsgrößen isoliert betrachtet werden. Aus diesem Grund werden die nicht

interessierenden Simulationsgrößen konstant gehalten (bspw. die Fallzahl und die Verteilungen), wohingegen diejenige Größe variiert wird, deren Effekt im Interesse steht (bspw. der Missinganteil). Durch diese Maßnahme lassen sich die interessierenden Effekte isolieren, was letztlich die gewünschten Aussagen hinsichtlich des Einflusses der Simulationsgrößen erlaubt. In der nachfolgenden Abbildung 12 sind die isolierten Effekte der Missinganteile, der Fallzahl und der Verteilungen dargestellt (respektive die erste, zweite und dritte Reihe in der Abbildung). Konstant gehalten werden jeweils die nicht relevanten Größen. Hierfür werden die angesprochenen problematischen Konfigurationen verwendet, um einen Eindruck zu gewähren, welche Simulationsgröße unter sonst suboptimalen Bedingungen am einflussreichsten ist (Fallzahl von 250, stark asymmetrische Verteilungen und 35 % Missinganteil).

7.1.1 Überblick: Unterschiede zwischen den MDTs und den Fit-Indices

Alle MDTs, welche im Imputationsprozess die Modellstruktur des Analysenmodells nicht berücksichtigen, schneiden bei allen Fit-Indices schlechter ab, als diejenigen, bei denen das Analysemodell bei der Behandlung der fehlenden Werte berücksichtigt wird. Direct-ML und H0 liefern also bessere Ergebnisse als alle anderen MDTs. Grundsätzlich zeigen sich beide auch eher robust gegenüber den simulierten Größen, wenngleich sich leichte Unterschiede erkennen lassen. Generell liegen mit H0 bessere Ergebnisse vor als mit Direct-ML; es lassen sich keine problematischen Datenkonstellationen identifizieren. Für Direct-ML lässt sich dagegen beobachten, dass die Ablehnungsraten höher sind, wenn der Missinganteil hoch und die Fallzahl gering ist. Allerdings können auch dabei in den meisten Fällen noch gute Ergebnisse erwartet werden. Insgesamt liegen mit beiden MDTs eher vernachlässigbare Ablehnungsraten vor, sodass bei der Bewertung der SE-Modelle auch korrekte Schlussfolgerungen erwartet werden können. Da für beide MDTs gute Ergebnisse vorliegen und die Ablehnungsraten des Referenzmodells nicht nochmals erhöht werden (mit dem Einsatz von H0 gehen sogar geringere Ablehnungsraten einher als im Referenzmodell), wird an dieser Stelle auf eine weitergehende Analyse dieser MDTs im Hinblick auf die Fit-Indices verzichtet (siehe Kapitel 7.2).

Abbildung 12: Ablehnungsraten der MDTs



Anmerkungen: R. KI (90 %): 90 %iges Konfidenzintervall von RMSEA. Erste Zeile: Konstanthaltung der Fallzahl auf 250 und der Verteilung auf ‚stark asymmetrisch' (skew3); Variation der Missinganteile. Zweite Zeile: Konstanthaltung der Verteilung auf ‚stark asymmetrisch' (skew3) und des Missinganteils auf 35 % (mar4); Variation der Fallzahl. Dritte Zeile: Konstanthaltung der Fallzahl auf 250 und des Missinganteils auf 35 % (mar4); Variation der Verteilungen. Die Fallzahl für EMB in der Konfiguration ‚250, skew3, mar4' beträgt 499.

Ein etwas anderes Bild zeigt sich für die anderen MDTs. Sie liefern in vielen Fällen keine zufriedenstellenden Ergebnisse mehr. Das ist problematisch, sollten SE-Modelle mittels der untersuchten Fit-Indices bewertet werden. Weiterhin lässt sich feststellen, dass alle MDTs in etwa gleich zu bewerten sind. Größere Unterschiede zwischen den MI-Varianten EMB, FCS, MNV, oder PMM lassen sich nicht erkennen. Die Einfachimputation mittels EM schneidet tendenziell etwas besser ab; wobei die Ablehnungsraten auch dabei in vielen Fällen nicht zufriedenstellend sind. Während eine korrekte Modellbewertung nach Einsatz von Direct-ML und H0 mit allen Fit-Indices einhergeht, lässt sich für die anderen Techniken nur ein zuverlässiger Index identifizieren, der in allen Fällen zu korrekten Bewertungen des Modells führt: das SRMR. Alle MDTs liefern unter allen dargestellten Konstellationen zuverlässige Werte für diesen Index und es lassen sich keine Änderungen in den Ablehnungsraten beobachten, wenn einzelne Simulationsgrößen variiert werden. Damit liegen auch keine Effekte vor, welche diesen Index beeinflussen: Damit ist das SRMR unabhängig von den gewählten Simulationskonfigurationen. Aus diesem Grund wird auch auf eine weiterführende Analyse des SRMR verzichtet (siehe auch Kapitel 7.2).

7.1.2 p-Wert der Chi²-Statistik

Für den p-Wert ist zu erkennen, dass selbst bei wenigen Missing Values alle MDTs[85] zu hohe Ablehnungsraten generieren: Diese belaufen sich bei einer kleinen Fallzahl und stark asymmetrischen Verteilungen selbst bei wenigen Missings auf ca. 25 % (erste Reihe[86]). Weiterhin zeigt sich, dass mit zunehmenden Missinganteil die Ablehnungsraten erheblich größer werden. Grundsätzlich kann ab einem Missinganteil von 20 % nicht mehr davon ausgegangen werden, dass die Modelle korrekt bewertet werden können. Zudem ist offensichtlich, dass weder ein Einfluss der Fallzahl, noch der Verteilungen vorliegt. Damit können die Ablehnungsraten mittels p-Wert weder durch die Erhöhung der Fallzahl, noch durch weniger asymmetrische Verteilungen reduziert werden (zweite, respektive dritte Reihe). Dies bedeutet, dass die Ablehnungsraten des p-Wertes vor allen Dingen durch die Missinganteile bestimmt werden, wohingegen die Fallzahlen und die Verteilungen keine Rollen spielen. Grundsätzlich sollte aufgrund der bereits erhöhten Ablehnungsraten bei geringen Anteilen an fehlenden Werten auf eine Modellbewertung mit dem p-Wert eher verzichtet werden.

[85] Nochmals sei erwähnt, dass Direct-ML und H0 von diesen Ausführungen ausgenommen sind.
[86] Wird in den Kapiteln 7.1.2 bis 7.1.4 auf eine jeweilige Reihe verwiesen, bezieht sich das auf Abbildung 12.

7.1.3 Root Mean-Square Error of Approximation – RMSEA

Zu hohe Ablehnungsraten lassen sich auch mit RMSEA bereits bei geringen Missinganteilen beobachten. Werden diese Anteile größer, kann nicht mehr davon ausgegangen werden, dass das SE-Modell korrekt bewertet wird (erste Reihe). Zudem verbessert sich die Modellablehnungsrate kaum, wenn die Fallzahl erhöht wird. Demnach ist auch dieser Index fallzahlunabhängig (zweite Reihe). Grundsätzlich sollte bei 35 % Missinganteil auf eine Modellbewertung mittels RMSEA verzichtet werden. Zudem haben auch die Variablenverteilungen keinen Einfluss (dritte Reihe). Demzufolge ist die Modellevaluation durch RMSEA davon abhängig, wie hoch der Missinganteil ist. RMSEA sollte also nur eingesetzt werden, wenn ein geringer Missinganteil vorliegt; ansonsten ist mit falschen Modellbewertungen zu rechnen.

Etwas bessere Ergebnisse als mit der Punktschätzung von RMSEA lassen sich mit dessen 90 %igen Konfidenzintervall beobachten. Ein Missinganteil von 5 % ist unabhängig von den Bedingungen unproblematisch. Zudem bewegen sich die Ablehnungsraten bei erhöhten Anteilen an fehlenden Werten unter dem Niveau der Punktschätzungen. Bei 35 % an Missing Values werden aber auch mit dem Konfidenzintervall nahezu alle Modelle zurückgewiesen (erste Reihe). Allerdings lässt sich, im Gegensatz zur Punktschätzung, die Modellbewertung mit dem Konfidenzintervall verbessern, wenn eine große Fallzahl gegeben ist. Ein dementsprechender Effekt kann beobachtet werden (zweite Reihe). Wiederum zeigt sich kein Einfluss der Verteilungen. Einzig EM liefert etwas bessere Ergebnisse, wenn symmetrische Verteilungen vorliegen: die Modellablehnungen werden geringer. Jedoch bleibt diese Reduktion im Hinblick auf die absolute Höhe des Anteils der Modellablehnungen unbedeutend (dritte Reihe). Somit werden die Modellablehnungen durch das Konfidenzintervall von den Missinganteilen und der Fallzahl bestimmt, nicht aber durch die Verteilungen. Das Konfidenzintervall kann den Ergebnissen zufolge zur Modellbewertung herangezogen werden, wenn erstens nur wenige Missing Values vorliegen, oder wenn zweitens bei höheren Missinganteilen die Fallzahl entsprechend hoch ist.

7.1.4 Comparative Fit Index – CFI

Bei geringen Anteilen an Missing Values liefern die MDTs zufriedenstellende Ergebnisse für den CFI. Wie bei den anderen Fit-Indices gehen aber auch mit dem CFI ab 20 % an Missing Values zu viele Modellablehnungen einher (erste Reihe). Die Erhöhung der Fallzahl führt zu einer wesentlichen Verbesserung der CFI-Ablehnungsrate, wobei auch dabei gilt, dass diese weiterhin zu hoch ist (zweite Reihe). Dagegen haben die Verteilungen wiederum kaum einen

Effekt. Einzig für EM zeigt sich eine leichte Tendenz wonach die Modellablehnungen zurückgehen, wenn symmetrische Verteilungen vorliegen; wiederum bleibt die Ablehnungsrate aber auf einem hohen Niveau (dritte Reihe). Schlussendlich gilt für die MDTs im Hinblick auf den CFI ähnliches wie für das Konfidenzintervall von RMSEA: Liegen geringe Quoten an Missings vor, ist die Modellbewertung mit dem CFI unproblematisch, höhere Quoten verlangen aber nach einer größeren Fallzahl. Wie bei den anderen Fit-Indices auch, geht von den Verteilungen eher kein Einfluss aus.

7.2 Modellbasierte Analyse der Ablehnungsraten

Mithilfe der modellbasierten Analysen, den Meta-Modellen, sollen die gewonnenen Erkenntnisse der deskriptiven Auswertung verdichtet werden. Während in der deskriptiven Analyse nur einzelne, problematische Simulationskonfigurationen separat auf der Konfigurationsebene betrachtet wurden, werden in diesem Abschnitt die Analysen für alle Konfigurationen auf der Replikationsebene durchgeführt. Das bedeutet, dass jede Replikation (jeder Fall) in die Modellanalyse aufgenommen wird. Hierfür werden die dichotomisierten Fit-Indices aller Replikationen eines Modells, die für die Berechnung der Ablehnungsraten notwendig waren, in einen eigenen, gemeinsamen Datensatz geschrieben. Weiterhin werden diejenigen MDTs, welche die Modellstruktur bei der Imputation der fehlenden Werte nicht berücksichtigen, zur Komplexitätsreduktion und aufgrund der nicht allzu unterschiedlichen Performanz zu einer Kategorie zusammengefasst. Durch dieses Vorgehen können die durchschnittlichen Effekte der Simulationsgrößen, über alle MDTs und Konfigurationen hinweg, auf die jeweiligen Fit-Indices identifiziert werden. Mit diesem Vorgehen wird ein höherer Abstraktionsgrad erreicht, was allgemeinere Aussagen bezogen auf die Performanz der MDTs erlaubt und gleichzeitig die Ergebnisse der deskriptiven Analyse weiter ausdifferenziert. Insgesamt sollte mit dem Abschluss der modellbasierten Analysen ein umfassendes Bild hinsichtlich der Einflussgrößen auf die Fit-Indices vorliegen.

Auf eine modellbasierte Analyse von Direct-ML und H0 wird in nahezu allen Fällen verzichtet, da in den meisten Fällen der Gesamtanteil an Modellablehnungen unter 5 % liegt und eine Modellschätzung deshalb nicht sinnvoll erscheint.[87] Weil auch kaum Varianz in den Ergebnissen des SRMR vorhanden ist (der Gesamtanteil an Modellablehnungen über alle MDTs

[87] Für den p-Wert von Direct-ML wird eine logistische Regressionsanalyse durchgeführt, da der Gesamtanteil an Modellzurückweisungen bei über 5 % liegt. Zwar können dabei mehrere Effekte identifiziert werden, allerdings ist deren inhaltliche Aussagekraft eher unbedeutend (AMEs < 5 %) und auch die jeweilige Modellanpassung ist unzureichend (Nagelkerke-R^2: 2.6 %). Damit ist die Performanz von Direct-ML in Bezug auf die Fit-Indices robust und nicht von den modellierten Bedingungen abhängig (siehe Anhang A4.2).

hinweg beträgt .03 %) wird auf eine modellbasierte Analyse dieses Indexes verzichtet. Bereits bei den deskriptiven Analysen konnten für keine der MDTs Effekte identifiziert werden, welche die Ablehnungsraten mit diesem Index beeinflussen. Demnach liegt mit dem SRMR ein robustes Maß zur Modellbewertung vor. Weiterhin werden nur die Ergebnisse zum dritten Modell berichtet, da diese, in Bezug auf die Einflussstärken und -richtungen, repräsentativ für das erste Modell sind.[88] Durch dieses Vorgehen reduziert sich die Zahl der zu schätzenden logistischen Regressionsmodelle auf vier: jeweils eine Schätzung pro Fit-Index.

Bevor nun die Ergebnisse der logistischen Regressionsanalysen präsentiert werden, sind einige Punkte anzusprechen: Sie betreffen die Kodierung der unabhängigen Variablen, die Fallzahl, auf welcher die Meta-Modell-Analysen basieren und die Interpretation der Effekte.

Die einzelnen Konfigurationsgrößen der Simulationen dienen in den Meta-Modellen als Prädiktoren. Weil es sich dabei um keine metrischen Variablen handelt, müssen diese in Dummies transformiert werden. Bei allen unabhängigen Variablen in den logistischen Regressionsmodellen handelt es sich damit um Dummyvariablen. Als Referenzkategorien dienen (wenn nicht anders angegeben) jeweils diejenigen Konfigurationen, die sich in den deskriptiven Analysen als weniger problematisch herausstellten:

a. Samplegröße: N = 750;
b. Verteilungsform: symmetrisch (skew1);
c. Missinganteil: 5 % (mar2).

Da die Analysen auf Ebene der Replikationen stattfinden, weisen die zugrundeliegenden Datensätze Fallzahlen von etwa 45.000 Fällen auf (max. 500 Replikationen mal 90 Konfigurationen[89]). Damit geht die Frage einher, wie verlässlich das Signifikanzkriterium ist. Denn bei einer Fallzahl von 45.000 kann davon ausgegangen werden, dass kleinere Unterschiede signifikant werden, die eigentlich vernachlässigbar sind. Aus diesem Grund schlägt die Literatur zur Auswertung von MC-Studien vor, neben dem Signifikanzniveau auch die Effektstärken zu betrachten (vgl. Bandalos/Leite 2013: 660). Hierfür werden die geschätzten Odds-Ratios mithilfe der Transformationsregeln von Borenstein u. a. (2009: 44 ff.) in Pearsonsche Korrelationskoeffizienten überführt. Als ein interpretationswürdiger Effekt wird ein schwacher Effekt mit einem r von .1 angesehen (vgl. Cohen 1988: 79). Dies wiederum entspricht einem Odds-Ratio von

[88] Modell 1: Anhang O2.2.
[89] Die Ergebnisse des Referenzmodells, von Direct-ML und H0 werden nicht aufgenommen und für Modell 2 liegen keine Fit-Indices vor. Damit reduzieren sich die 396 simulierten Konfigurationen auf insgesamt 180 (Modell 1 und Modell 3). Es liegen dann für Modell 3, 90 Konfigurationen und 44.999 Fälle vor. Zur Abweichung von 45.000 kommt es, da für EMB in einer Konfiguration nur 499 Replikationen gegeben sind.

ungefähr .7 bzw. 1.45. Signifikante Effekte mit einer Korrelation von < .1 werden nicht interpretiert. Gleichzeitig wird ein nicht signifikanter Effekt als nicht interpretationswürdig angesehen.

Zur Interpretation der Effekte werden die durchschnittlichen marginalen Effekte (*average marginal effects*: AMEs) herangezogen. Während die Odds-Ratios dazu dienen, die Effektstärken einzuordnen, werden die AMEs für inhaltliche Aussagen benutzt. Weil es sich bei den unabhängigen Variablen der Regressionsmodelle um Dummies handelt, können diese wie folgt interpretiert werden: Die Wahrscheinlichkeit ein Modell abzulehnen liegt durchschnittlich um AME Prozent niedriger/höher als in der jeweiligen Referenzkategorie, wenn alle anderen unabhängigen Variablen konstant gehalten werden (vgl. Urban/Mayerl 2018: 405 f.). Aufgrund dessen, dass aus den Logits oder den Odds-Ratios nicht unmittelbar auf die Ausprägung der AME-Werte geschlossen werden kann, müssen diese in Bezug auf ihre inhaltliche Aussagekraft eingeordnet werden. Diese Einordnung erfolgt anhand der Ausprägung der einzelnen Effekte im jeweiligen Meta-Modell. Sollten größere Unterschiede zwischen den Effekten vorliegen, so sind diejenigen vernachlässigbar, die im Vergleich zu den stärksten Effekten abfallen.[90]

7.2.1 Haupteffekte

In Abbildung 13 sind die geschätzten AMEs (inkl. dem 95 %igen Konfidenzintervall) der vier angesprochenen logistischen Regressionen dargestellt.[91] Die zugehörigen Tabellen finden sich

[90] Meist handelt es sich bei den Effekten, die als vernachlässigbar angesehen werden können, weil sie gegenüber den anderen abfallen, um Effekte, die AMEs von ca. 10 % aufweisen. Aus diesem Grund ist in Abbildung 13 eine Linie zur Orientierung bei y ± 10 % sichtbar.

[91] Während die logistischen Regressionsmodelle, in denen nur die Haupteffekte berücksichtigt werden, keine Schätzprobleme haben, können für die ML-Schätzung bei den Interaktionsmodellen aufgrund (quasi-)vollständiger Separation teilweise verzerrte Ergebnisse beobachtet werden. Aus diesem Grund werden die betroffenen Modelle zusätzlich mit einer korrigierten ML-Schätzung, der *Penalized Maximum Likelihood*-Schätzung (PML) nach Firth (1993), geschätzt. Bei PML handelt es sich um ein Korrekturverfahren, das die geschätzten Koeffizienten der ML-Schätzung mit einer Korrektur (*penalty*) versieht, wenn diese unrealistische Werte annehmen; sich also weit von null entfernen, wie es bei diesem Modellverstoß der Fall sein kann. Dadurch lassen sich auch bei (quasi-)vollständiger Separation zuverlässige Schätzungen der b-Koeffizienten und den zugehörigen Konfidenzintervallen tätigen (vgl. Cole u. a. 2014: 255; Fiebig 2012: 151 ff.). Heinze/Schemper (2002) zeigen allerdings, sollte PML sehr große Schätzwerte für die b-Koeffizienten liefern, dass die inferenzstatistischen Tests dann nicht mit der Wald-Statistik durchgeführt werden sollten. Hierbei kommt es häufiger zu Überschätzungen, was fälschlicherweise in der Annahme der Nullhypothese resultiert. Sie raten deshalb dazu, zusätzlich die Konfidenzintervalle mittels *Profile Penalized Likelihood* zu berechnen (vgl. ebd.: 2418). Da für die vorliegenden Ergebnisse aber nicht unbedingt die Signifikanz, sondern vor allem die Effektstärke als Indikator eines interpretationswürdigen Effektes dient, wird diese nicht berechnet. Es lassen sich auch keine Schätzungen beobachten, bei denen PML Schätzwerte liefert, deren (Nicht-)Signifikanz im Widerspruch zur Effektstärke stehen. Problematisch an der korrigierten Schätzung nach Firth ist allerdings, dass keine automatisierte Möglichkeit besteht, die AMEs daraus zu generieren (zumindest zum Zeitpunkt der Durchführung der Analysen für diese Arbeit). Deshalb, und weil die verzerrten Lösungen mit der ML-Schätzung nur in den Interaktionsmodellen auftreten, wird PML nur dazu benutzt, interpretationswürdige Interaktionen zu identifizieren. Infolgedessen werden diese Interaktionen in separaten Analysen mit der normalen ML-Schätzung geprüft, womit auch die AMEs berechnet werden können. PML wird also lediglich zu Modellbildungszwecken eingesetzt. Aufgrund dieser Problematik sind in der Abbildung 13 nicht alle AMEs aller

im Anhang A4.2. Erst werden die Haupteffekte betrachtet, bevor im kommenden Kapitel die Interaktionsterme[92] berücksichtigt werden. Für alle Fit-Indices lassen sich substantielle Haupteffekte für die Samplegröße, die (stark) asymmetrische Datenkonfiguration sowie die Missinganteile ausmachen. Es gibt allerdings Unterschiede zwischen den einzelnen Fit-Indices im Hinblick darauf, wie stark die einzelnen Effekte im jeweiligen Modell zueinander zu bewerten sind.

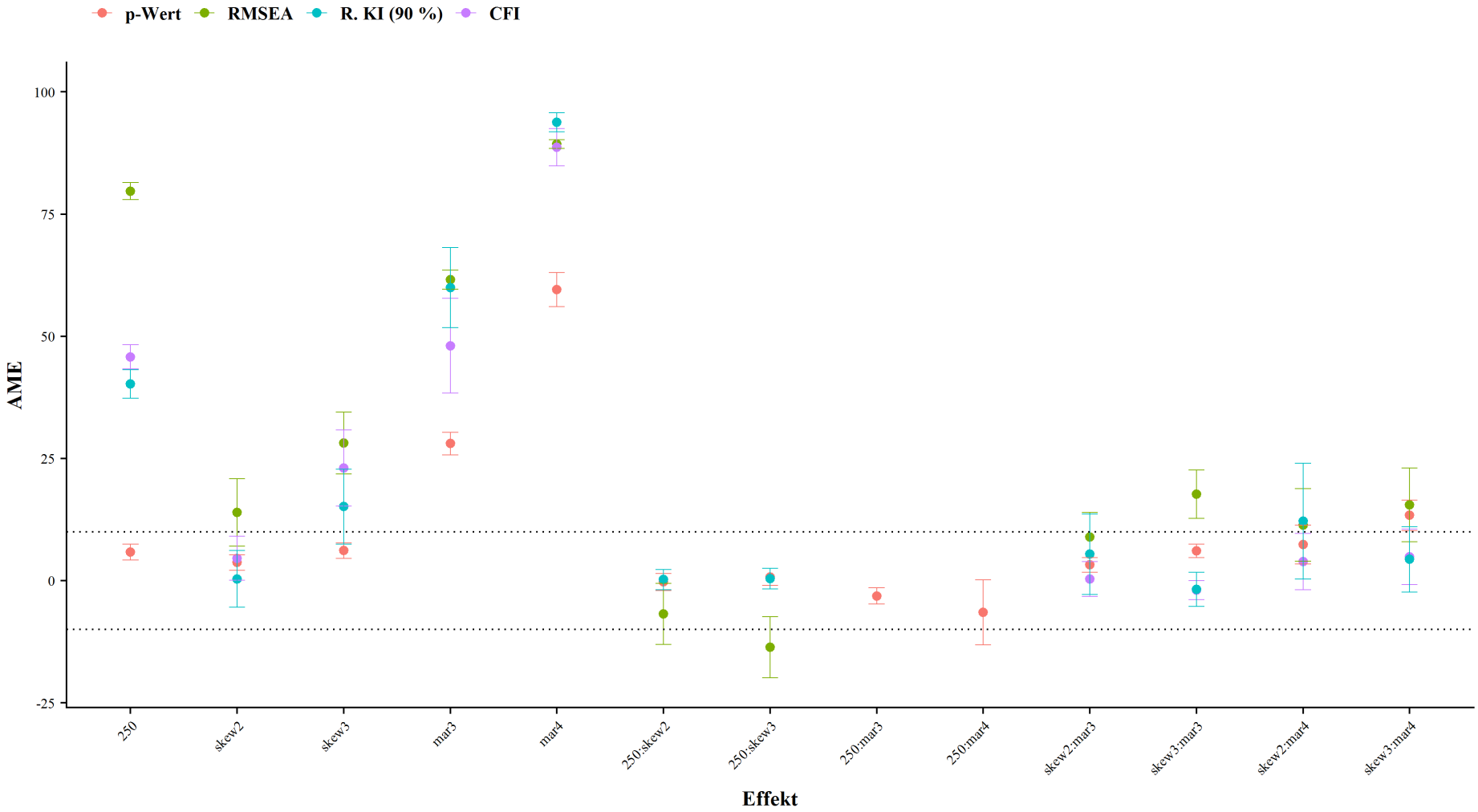
Für den p-Wert zeigt sich, dass neben der Fallzahl auch ein Einfluss der stark asymmetrischen Verteilungskategorie vorliegt. Beide Effekte sind allerdings mit ca. 6 % eher vernachlässigbar. Denn im Gegensatz dazu, haben die beiden Missingkategorien einen stärkeren Einfluss auf die Modellablehnung. Die durchschnittliche Wahrscheinlichkeit mittels p-Wert ein Modell abzulehnen ist bei mittleren oder hohen Quoten an Missings, im Vergleich zur Referenzkategorie und unter Konstanthaltung aller anderen Variablen, um 28 % bzw. um 59 % höher. Damit liegen für die Missingkategorien auch die stärksten Effekte für den p-Wert vor.

Ähnliches gilt auch für RMSEA und dessen Konfidenzintervall. Mit abnehmender Samplegröße steigen für beide Fit-Indices die Wahrscheinlichkeiten das Modell abzulehnen. Der Einfluss der Fallzahl ist besonders für die Punktschätzung mittels RMSEA mit knapp 80 % besonders deutlich. Beide Indices weisen zudem substantielle Effekte der Missingkategorien auf, wobei die 35 % Kategorie in beiden Fällen den jeweils stärksten Effekt im Modell darstellt (89 % respektive 94 %). Erhöhen sich demnach die Missinganteile von 5 % auf 20 % oder von 5 % auf 35 % dann liegt durchschnittlich eine höhere Wahrscheinlichkeit vor, mit beiden Indices, unter Konstanthaltung aller anderen Variablen, das Modell zurückzuweisen. Zwar gibt es einen solchen Einfluss auch für die Verteilungen, wonach mit zunehmender Asymmetrie die Modellablehnungen wahrscheinlicher werden (für das Konfidenzintervall nur in der stark asymmetrischen Kategorie), allerdings können diese Effekte als eher vernachlässigbar angesehen werden, da sie im Vergleich zu den anderen Effektstärken im Modell bedeutend weniger ausgeprägt sind.

Interaktionen dargestellt. Die Ergebnisse der separaten Analysen finden sich im Anhang A4.2. Neben der Problematik von (quasi-)vollständiger Separation offenbart die Modelldiagnostik für die Interaktionsmodelle in den meisten Fällen erhöhte Multikollinearität. Da allerdings weder die Effektrichtung noch die Einordnung als interpretationswürdiger Effekt dadurch beeinflusst wird, wird dieser Verstoß ignoriert (auch weil in der Forschungspraxis das Vorliegen von Multikollinearität, die bspw. durch Interaktionen ‚künstlich' entstanden ist, als unproblematisch angesehen wird; vgl. Urban/Mayerl 2018: 245 ff.).

[92] Die Hinzunahme der Interaktion zweiter Ordnung resultiert nicht in besseren Modellanpassungen. Aus diesem Grund werden nur Interaktionen der ersten Ordnung berücksichtigt.

Abbildung 13: Ergebnisse der modellbasierten Analyse zu den Ablehnungsraten



Im Grunde gelten die Ausführungen für das Konfidenzintervall auch für den CFI. Denn auch mit diesem steigt die durchschnittliche Wahrscheinlichkeit einer Modellablehnung an, wenn eine geringe Fallzahl oder größere Anteile an Missing Values vorliegen. So ist die Wahrscheinlichkeit einer Modellablehnung, wenn jeweils alle anderen unabhängigen Variablen konstant gehalten werden, bei geringer Fallzahl um durchschnittlich 46 %, bei mittleren Missinganteilen um 48 % und bei hohen Missingquoten um 89 % höher, als bei der jeweiligen Referenzkategorie. Zusätzlich zeigen sich für den CFI auch Effekte für die Verteilungen. Während derjenige für die asymmetrische Kategorie mit ca. 5 % vernachlässigbar bleibt, zeigt sich ein stärkerer Effekt, wenn stark asymmetrische Verteilungen vorliegen (23 %). Auch hier gilt allerdings, dass der Effekt, vor allem im Vergleich zu den sehr ausgeprägten Effekten der Samplegröße und den Missinganteilen eher wieder vernachlässigbar ist.

7.2.2 Interaktionseffekte

7.2.2.1 Fallzahl und Verteilungen

Die Interaktionen aus der Fallzahl und den Verteilungen überschreiten die Schwelle für interpretierbare Effekte im p-Wert- und Konfidenzintervall-Modell nicht. Die Effekte der Verteilungen bleiben demnach für den p-Wert und das Konfidenzintervall über die Fallzahl hinweg stabil. Das bedeutet, dass die Variablenverteilung eine vernachlässigbare Größe im Hinblick auf diese beiden Fit-Indices darstellt.

Im RMSEA-Modell überschreitet dagegen der Interaktionseffekt mit den stark asymmetrischen Verteilungen den festgesetzten Grenzwert; derjenige mit den asymmetrischen Verteilungen bleibt unter der interpretationswürdigen Schwelle. Beide Interaktionseffekte zeigen aber in die gleiche Richtung: In Kombination mit einem kleinen Sample reduzieren sich die Verteilungseffekte etwas, werden aber nicht komplett kompensiert. Hierbei ist allerdings anzumerken, dass dieser Interaktionseffekt auch anders interpretiert werden kann: Wenn (stark) asymmetrische Verteilungen vorliegen, dann ist der Effekt der Samplegröße weniger stark ausgeprägt. Da beide Interaktionseffekte allerdings nur unwesentlich den Fallzahleffekt reduzieren (dieser bleibt mit weiterhin über 65 % recht deutlich), trägt diese Interaktion auch nur bedingt zu einer Verminderung der durchschnittlichen Wahrscheinlichkeit einer Modellablehnung unter kleinen Fallzahlen bei (siehe auch die Analyse getrennt nach der Fallzahl in Tabelle A11).

Für den CFI liegen dagegen stärkere Effekte der Verteilungen für das kleine Sample vor. Ist deren Effekt für das große Sample eher vernachlässigbar (ca. 2 % und 7 %), können für das

kleine Sample Effektstärken von 20 % und 37 % beobachtet werden.[93] Damit kann der Einfluss der Variablenverteilung reduziert werden, wenn ein großes Sample vorliegt. Demnach ist ein größeres Sample, sollten asymmetrische Verteilungen vorliegen, für den CFI begünstigend. Der Effekt kann zwar nicht kompensiert werden kann, dennoch lässt er sich soweit reduzieren, dass er zu vernachlässigen ist.

7.2.2.2 Fallzahl und Missinganteile

Die Interaktionseffekte mit der Fallzahl und den Missinganteilen für den p-Wert liegen unterhalb (Fallzahl und mittlerer Anteil) bzw. knapp über der interpretationswürdigen Grenze (Fallzahl und höchster Anteil). Aufgrund der Tatsache, dass der Haupteffekt mit ca. 60 % aber bereits sehr ausgeprägt ist, erscheint der Interaktionseffekt mit ungefähr -6 % vernachlässigbar. Die Haupteffekte können nicht durch eine Veränderung der Fallzahl kompensiert werden: Die Einflüsse der Missinganteile bleiben auch bei verschieden großen Fallzahlen konstant.

Dies gilt auch für RMSEA. Zwar gibt es hierbei einen moderierenden Effekt der Samplegröße (in einem großen Sample ist der Einfluss der Missingkategorien stärker als in einem kleinen), allerdings kann die Veränderung der Fallzahl den Effekt nicht ausgleichen. Sie moderiert den Einfluss zwar, bedeutend bleibt dieser aber weiterhin: Damit liegen auch für den RMSEA konstant hohe Einflüsse der Missingkategorien vor.[94]

Die Ergebnisse für das Konfidenzintervall und für den CFI sind wiederum sehr ähnlich. Für beide Indices dient die Fallzahl als Moderator, denn sie kann den Effekt der Missingkategorien erheblich reduzieren. Beträgt dieser im kleinen Sample für das Konfidenzintervall etwa 73 % und für den CFI 72 %, reduziert sich der Effekt in einem großen Sample auf 37 %, respektive 10 % (der Unterschied ist jeweils signifikant[95]): Liegen demnach mehr als 20 % Missing Values vor, ist es für beide Indices begünstigend, ein großes Sample zu verwenden.

7.2.2.3 Verteilungen und Missinganteile

In drei der vier Modelle sind diese Interaktionseffekte substantiell: Bei zunehmend asymmetrischen Verteilungen werden die Effekte der Missingkategorien stärker. Dabei können moderate bis größere Zunahmen in den Effekten ausgemacht werden (p-Wert: ca. 6 % bis 13 %; RMSEA: 9 % bis 18 %; Konfidenzintervall: 4 % bis 12 %). Damit ist eine Kombination von Missinganteilen und asymmetrischen Verteilungen für diese Indices problematisch. Lediglich die Höhe

[93] Siehe hierzu die Tabelle A13.
[94] Siehe dazu die separate Analyse getrennt nach der Fallzahl (Tabelle A11).
[95] Die Konfidenzintervalle der betreffenden Logit-Koeffizienten überschneiden sich nicht (Tabelle A12 und A13).

in der Zunahme ist zwischen den Fit-Indices unterschiedlich. Da bereits sehr starke Haupteffekte der Missinganteile vorliegen, werden diese durch asymmetrische Verteilungen nur noch weiter verschärft; unabhängig wie groß die Zunahme durch diese ausfällt. Für den CFI können diese Interaktionen zwar auch als substantiell identifiziert werden. Anders als für die anderen Fit-Indices verschärft sich der Effekt der Missingkategorien aber nur geringfügig, wenn asymmetrische Verteilungen vorliegen.

7.2.3 Zusammenfassung der modellbasierten Analyseergebnisse

Aus den vorangegangenen Analysen lassen sich sechs Ergebnisse ableiten:

1) Die Performanz der MDTs, welche die Modellstruktur bei der Behandlung der fehlenden Werte nicht berücksichtigen (EM, EMB, FCS, MNV und PMM), ist von den modellierten Bedingungen abhängig. Das zeigen nicht nur die interpretationswürdigen Effekte in den einzelnen Modellen, sondern auch deren sehr hohen Anpassungsmaße: Das Nagelkerke R^2 für die einzelnen Interaktionsmodelle reicht von ca. 65 % (p-Wert-Modell) bis 77 % (CFI-Modell).

2) Für die MDTs kann ein Einfluss der Samplegröße auf deren Performanz identifiziert werden. Gleichzeitig zeigt sich auch der Missinganteil als entscheidender Einflussfaktor: a) Die Effekte der Fallzahl unterscheiden sich zwischen den Fit-Indices etwas. So ist RMSEA stark durch die Samplegröße beeinflusst, wohingegen das Konfidenzintervall und der CFI etwas weniger anfällig sind. Kein Einfluss der Fallzahl kann für den p-Wert nachgewiesen werden. b) Die Missingkategorien weisen über alle Fit-Indices die stärksten Effekte auf.

3) Weiterhin kann gezeigt werden, dass die Variablenverteilung einen Effekt auf die Performanz der MDTs hat. Auch hierbei gibt es wiederum Unterschiede zwischen den Fit-Indices. Vor allen Dingen liegen Effekte im RMSEA-Modell vor. Die anderen Modelle weisen dagegen nur schwache (Konfidenzintervall und CFI) oder vernachlässigbare Effekte auf (p-Wert). Zudem gilt, dass die Verteilungseffekte in jedem Modell gegenüber den Effektstärken der Fallzahl und der Missinganteile abfallen.

4) Der Einfluss der Fallzahl kann für RMSEA über die Variablenverteilungen moderiert werden. Liegen (stark) asymmetrische Verteilungen vor, ist der Fallzahleffekt geringer, als bei symmetrischen Verteilungen. Allerdings kann unter keinen Umständen der Fallzahleffekt kompensiert werden: Damit gilt für RMSEA, dass dieser vor allen Dingen bei geringen Fallzahlen eher nicht angewendet werden sollte – unabhängig davon welche Verteilungen vorliegen.

5) Ein moderierender Effekt der Fallzahl liegt, ausgenommen für den p-Wert, für alle Fit-Indices hinsichtlich des Missinganteils vor. Für das Konfidenzintervall und für den CFI schwächt sich der Effekt der Missingkategorien bei zunehmender Fallzahl ab. In allen Fällen

gelingt dies allerdings nicht in dem Maße, als dass der Haupteffekt der Missingkategorien ausgeglichen werden kann. Damit ist ein großes Sample zwar wünschenswert, bei großen Anteilen an Missings kommt es aber dennoch häufig zu Modellablehnungen. Für RMSEA liegt dagegen ein anders gelagerter Effekt vor. Hier zeigt sich, dass bei kleineren Fallzahlen geringer ausgeprägte Effekte vorliegen, als bei großen Fallzahlen. Abermals können die Effekte der Missingkategorien aber nicht kompensiert werden: Schlussendlich sind diese bei kleinen Fallzahlen zwar geringer, aber weiterhin bedeutend. Im Gegensatz dazu, ist der Einfluss der Missinganteile auf den p-Wert über die Fallzahl hinweg konstant; dementsprechend ist es für den p-Wert unerheblich, ob kleine oder große Fallzahlen vorliegen.

6) Eine Zunahme des Einflusses der Missinganteile lässt sich für alle Fit-Indices beobachten, wenn (stark) asymmetrische Verteilungen vorliegen. Diese erhöhen die durchschnittlichen Wahrscheinlichkeiten der Modellablehnung. Liegen dementsprechend (stark) asymmetrische Verteilungen vor, dann wird bei erhöhten Quoten an Missings das Modell eher verworfen, als wenn symmetrische Kategorien gegeben sind. In vielen Fällen ist diese Zunahme allerdings vernachlässigbar, da bereits die Haupteffekte der Missings entsprechend ausgeprägt sind.

7.3 Ablehnungsraten: Ergebnisdiskussion und Einordnung

7.3.1 Exkurs: Anmerkungen zu den Ergebniseinordnungen

Die Ergebnisdiskussionen dieser Arbeit umfassen unter anderem Einordnungen in den aktuellen Forschungsstand. Hierzu müssen zunächst aber einige Anmerkungen gemacht werden. Das liegt daran, dass die Übertragbarkeit der Ergebnisse einer MC-Studie eingeschränkt ist (siehe dazu die Ausführungen in Kapitel 6.1). Das Forschungsdesign dieser Arbeit orientiert sich zwar an anderen Arbeiten mit ähnlichem Themenschwerpunkt, allerdings gibt es zwischen diesen und der Vorliegenden einige Unterschiede. Ein direkter Vergleich ist deshalb aus mehreren Gründen problematisch. Erstens gibt es Unterschiede was die getesteten Modelle angeht. So werden die Modelle dieser Arbeit in einen Vergleich zu bi- und multivariaten linearen Regressionen, Wachstumsmodellen, noch komplexeren SE-Modellen u. a. gesetzt. Zusätzlich können auch Unterschiede in den gewählten Populationsparametern vorliegen. Zweitens gibt es Unterschiede was die weiteren Konfigurationsgrößen betrifft (wie unterschiedliche Fallzahlen, unterschiedliche Anteile an Missing Values, andere Verteilungen oder andere Variablenskalierungen). Drittens kann auch die Implementation der MDTs unterschiedlich sein. Andere Studien legen für FCS andere Schätzungen für die Imputation der fehlenden Werte in einzelnen Variablen an, benutzen für die MI andere Iterationszahlen oder wählen ein anderes *m* etc. Viertens

können auch unterschiedliche Rahmenbedingungen der Simulationen vorliegen (wie eine empirische oder synthetische Datengrundlage). Zuletzt können auch unterschiedliche Bewertungskriterien gegeben sein (z. B. andere Grenzwerte für die Fit-Indices oder gänzlich andere endogene Modellparameter). All dies schränkt die Aussagekraft eines direkten Vergleichs ein. Nichtsdestotrotz verbleibt die Möglichkeit die Ergebnisse in ihrer Tendenz einzuordnen und zu prüfen, ob sich diese Arbeit von den anderen unterscheidet, oder ob sie zu ähnlichen Ergebnissen kommt.[96]

7.3.2 Diskussion zum SRMR, zu Direct-ML und H0

Als erstes soll das SRMR im Fokus stehen. Mit diesem Index können unter allen Bedingungen akzeptable Ablehnungsraten erzielt werden. Mit dem SRMR liegt damit ein Fit-Index vor, der anhand der vorliegenden Ergebnisse mit dem gesetzten Grenzwert von .08 für alle MDTs problemlos empfohlen werden kann. Im Kontext von fehlenden Werten ist für die Praxis dazu zu raten, diesen Wert zu verwenden, anstatt das striktere Kriterium von .05, das bisweilen auch zum Einsatz kommt. Zumal Teman (2012) zeigt, dass bei einem Wert von .05 die Ablehnungsraten für das SRMR beträchtlich höher sind.

Aus den Analysen wird weiterhin deutlich, dass von allen getesteten MDTs H0 die Missing Values in Bezug auf die Fit-Indices am besten handhabt. Damit können unabhängig des Fit-Indexes in allen Fällen zufriedenstellende Ablehnungsraten beobachtet werden. Ähnliches gilt auch für Direct-ML. Auch damit gehen zufriedenstellende Ergebnisse hinsichtlich aller Fit-Indices einher. Anzumerken ist, dass sie im Gegensatz zu H0 leicht höhere Ablehnungsraten produziert. Grundsätzlich führt der Einsatz von Direct-ML aber zu korrekten Modellbewertungen. Somit können beide für den Praxiseinsatz empfohlen werden. Sie erweisen sich gegenüber den modellierten Bedingungen als robust und die Gefahr, Modelle fälschlicherweise zurückzuweisen, ist im Vergleich zu den anderen getesteten MDTs reduziert. Weil Direct-ML einfacher umzusetzen ist, da sie in einigen Statistikpaketen die Standardoption darstellt und die fehlenden Werte direkt berücksichtigt, ohne diese vorher imputieren zu müssen, ist sie leicht zu favorisieren.

Dass beide MDTs gute Ergebnisse liefern, ist der Tatsache geschuldet, dass sowohl Direct-ML als auch H0 die Annahme treffen, dass das spezifizierte Modell zur Behandlung der fehlenden Werte korrekt ist. Demnach wird bereits bei der Behandlung der fehlenden Werte festgelegt, wie das betreffende Modell in der Population vorzufinden ist. In beiden Fällen werden

[96] Alle diese Ausführungen treffen auch auf die Einordnungen in den Kapiteln 8.4.1 und 9.3.1 zu.

also nur Beziehungen berücksichtigt, die spezifiziert sind. Aus diesem Grund werden die Missing Values entsprechend behandelt, dass die daraus resultierenden Kovarianzmatrizen diesem Modell entsprechen. Für H0 kommt gleichzeitig hinzu, dass im Endeffekt für die Missing Values erneut Daten aus dem Populationsmodell generiert werden, da sowohl das Populationsmodell als auch das IM identisch sind. Wie aber beide Techniken zu bewerten sind, wenn die Modelle, die dazu benutzt werden, um die Missing Values zu handhaben, missspezifiziert sind, kann diese Arbeit nicht beantworten. Asparouhov/Muthén (2010c) liefern für H0 einen Hinweis, wonach die Technik gegenüber leichten Missspezifikationen robust ist (vgl. ebd.: 22). Für Direct-ML kann diesbezüglich auf Davey u. a. (2005) verwiesen werden. Die Autoren zeigen, dass die Fit-Indices unter Einsatz von Direct-ML zu oft falsch spezifizierte Modelle nicht zurückweisen: Mit zunehmenden Anteilen an Missing Values wird es wahrscheinlicher, das Modell nicht zu verwerfen.

Eine solche Tendenz, wonach mit zunehmendem Missinganteil die Modellbewertung besser wird, lässt sich in dieser Arbeit, im Vergleich zum Referenzmodell, für H0 ausmachen. Ob allerdings H0 bei falsch spezifizieren IMs auch die Anpassungswerte verbessert, wird in dieser Arbeit nicht geprüft. Sollte dies aber der Fall sein, dann werden Daten generiert, die es wahrscheinlich machen, dass ein falsch spezifiziertes Modell bei hohen Missinganteilen gute Anpassungswerte aufweist. Das wiederum kann für die empirische Praxis problematisch werden, denn dort ist meist nicht bekannt, ob das spezifizierte Modell tatsächlich dem Populationsmodell entspricht. Die Praxis würde wohl bei guten Anpassungswerten ein falsch spezifiziertes Modell akzeptieren und dessen Schätzergebnisse interpretieren. Das könnte aber zu falschen Schlussfolgerungen führen.

Aufgrund der geringen Forschungslage bei missspezifizierten Imputations- und/oder Analysemodellen und aufgrund des besseren Abschneidens von H0 (im Vergleich zum Referenzmodell) bei hohen Missinganteilen sowie wegen der Ergebnisse von Davey u. a. (2005), sollte in der empirischen Praxis darauf geachtet werden, dass das zu schätzende Modell eine fundierte analytische Basis aufweist. Ansonsten könnte es der Fall sein, dass bei hohen Missinganteilen durch den Einsatz dieser beiden MDTs falsch spezifizierte Modelle akzeptiert werden. In Tabelle 9 werden alle Ausführungen dieses Unterkapitels zusammengefasst.

Tabelle 9: Zusammenfassung der Performanz bzgl. der Fit-Indices I

Direct-ML, H0	SRMR	p-Wert	RMSEA	R. KI (90 %)	CFI
Grundniveau	zufrieden-stellend	zufrieden-stellend	zufrieden-stellend	zufrieden-stellend	zufrieden-stellend
Kleine Fallzahl	x	x	x	x	x
Zunehmende Asymmetrien	x	x	x	x	x
Steigender Missinganteil	x	x	x	x	x
Empfehlung	unter allen Bedingungen zur Modellbewertung geeignet				
Einschränkung	falsch spezifizierte Modelle könnten bei hohen Missinganteilen akzeptiert werden				

Anmerkungen: x: kein Effekt; +/- positiver/negativer Effekt; ++/-- stark positiver/negativer Effekt.

7.3.3 Diskussion zu den anderen MDTs und deren Performanz

Im Gegensatz zu Direct-ML und H0 gestaltet sich das Ergebnis für EM, EMB, FCS, MNV und PMM problematischer. Für diese MDTs ist entscheidend, wie viele Missing Values vorliegen; die beiden Haupteffekte der Missingkategorien weisen in allen Modellen die stärksten Effekte auf. Zudem zeigt sich bereits in den deskriptiven Analysen, dass die Missinganteile am meisten zur Modellablehnung beitragen. Auch kann das Ergebnis der deskriptiven Analyse hinsichtlich eines Einflusses der Fallzahl für den CFI und das Konfidenzintervall, mittels der modellbasierten Analyse aufgegriffen werden, denn auch dieser Effekt zeigt sich darin. Für RMSEA liegt ein solcher Effekt in der deskriptiven Auswertung jedoch nicht vor. Im Meta-Modell stellt sich aber heraus, dass der durchschnittliche Effekt der Samplegröße für RMSEA derart stark ausgeprägt ist, dass davon abzuraten ist, diesen Index zur Bewertung der Modelle heranzuziehen, wenn bei kleinen Fallzahlen, Missing Values vorliegen. Denn ohne eine Variation der anderen Simulationsgrößen vorzunehmen, werden bei kleinen Fallzahlen mit wenigen Missing Values (5 %) bereits sehr viele Modelle zurückgewiesen. Dies ist typisch für RMSEA, da dieser vor allem in kleinen Stichproben dazu tendiert, größer und ungenauer zu sein (vgl. Iacobucci 2010: 96). Letztlich bedeutet das verschiedenartige Ergebnis zwischen der deskriptiven und der modellbasierten Analyse, dass die Performanz von RMSEA bei geringen Missingquoten durch die Samplegröße beeinflusst werden kann, während bei sehr hohen Missinganteilen die Performanz von RMSEA eher unabhängig von der Fallzahl ist. Für den p-Wert lässt sich ein Fallzahleffekt eher nicht beobachten. Das ist bemerkenswert, denn der gängigen Erwartung nach, ist der p-Wert größer, je kleiner das Sample ist. Das wiederum würde bedeuten, dass bei kleinen Fallzahlen weniger Modellablehnungen zu beobachten sind. Im Grunde heißt das, dass die Fallzahl für den p-Wert unerheblich ist, wenn Missing Values vorliegen. Alle vier Fit-Indices sind weiterhin durch die Variablenverteilungen beeinflusst. Diese Effekte sind aber in allen Modellen den Effekten der Missinganteile und der Samplegröße nachgelagert. Aus diesem Grund ist die

Variablenverteilung auch eher eine vernachlässigbare Größe. Dass die Verteilungen keinen Effekt aufweisen, ist ein weiteres Ergebnis, das in der deskriptiven Analyse beobachtet werden kann.

Zwar können für alle Fit-Indices bedeutsame Moderatorvariablen ausgemacht werden, in allen Fällen können aber die Haupteffekte damit nicht ausgeglichen werden. So reduzieren sich für den CFI und das Konfidenzintervall mit der Erhöhung der Fallzahl die Effekte der Missingkategorien, aber nicht in dem Maße, als dass diese bedeutungslos werden. Ähnliches gilt auch für RMSEA: Sowohl die Effekte der Missingkategorien (das gilt auch für den p-Wert), als auch der Variablenverteilungen sind bei kleinen Fallzahlen weniger stark ausgeprägt als bei großen Fallzahlen. Weil für RMSEA bei kleinen Fallzahlen bereits ein ausgeprägtes Niveau an Modellzurückweisungen vorhanden ist, kann der Einfluss der Interaktionen vernachlässigt werden. Dementsprechend ist es für RMSEA sinnvoller, wenn eine große Fallzahl vorliegt. Weiterhin kann beobachtet werden, dass sich die Effekte der Missings bei asymmetrischen Verteilungen verschärfen – dieser Zuwachs ist in Anbetracht der bereits sehr ausgeprägten Effekte der Missingkategorien aber unbedeutend.

Im Gegensatz zu Direct-ML und zu H0 ist es demnach äußerst problematisch, Modelle korrekt zu bewerten, wenn eine der MDTs angewendet wird, die bei der Behandlung der fehlenden Werte die Modellstruktur nicht berücksichtigen. Wird in der empirischen Praxis auf eine dieser MDTs zurückgegriffen, so ist entscheidend, wie hoch der Missinganteil ist. Bei geringen Anteilen ist die Modellbewertung anhand des CFIs oder des Konfidenzintervalls unproblematisch; anders dagegen mittels dem p-Wert und RMSEA. Denn beim p-Wert liegt bereits ein ausgeprägtes Grundniveau an Modellablehnungen vor: Bereits in der am wenigsten problematischen Konfiguration zeigen sich für den p-Wert höhere, nicht zufriedenstellende Ablehnungsraten. Dies bedeutet, dass bereits bei wenigen Missing Values, ohne dass andere Simulationsgrößen variiert werden, höhere Ablehnungsraten vorliegen als bei den anderen Fit-Indices. Grundsätzlich kann deshalb der p-Wert bereits bei Missinganteilen von 5 % eher nicht empfohlen werden; ab 20 % an Missings sollte dann auf diesen verzichtet werden. RMSEA kann dagegen bei 5 % an Missings eingesetzt werden, sofern ein großes Sample gegeben ist. Ist das nicht der Fall, sollte auf diesen verzichtet werden. Grundsätzlich gilt, dass RMSEA ab einem Anteil von 20 % nicht zur Modellbewertung eingesetzt werden sollte. Bei solchen Bedingungen empfiehlt es sich, dessen Konfidenzintervall heranzuziehen oder den CFI. Beide liefern bei größeren Fallzahlen und 20 % Missings noch zufriedenstellende Modellbewertungen. Sollten bei diesem Anteil aber gleichzeitig kleine Fallzahlen vorliegen, können auch sie nicht empfohlen werden. Ab

einer Quote von 35 % sollte eine Modellbewertung nur mit dem CFI erfolgen, und auch nur dann, wenn eine größere Fallzahl gegeben ist. Alle anderen Indices werden unter solchen Bedingungen das Modell fälschlicherweise ablehnen.

Insgesamt gilt für diese MDTs ähnliches wie für Direct-ML und H0: In der empirischen Praxis ist es notwendig eine fundierte analytische Basis aufzuweisen; denn rein aufgrund der statistischen Kennwerte der Fit-Indices (mit Ausnahme des SRMR) ist die Wahrscheinlichkeit, ein korrekt spezifiziertes Modell nach Behandlung der fehlenden Werte fälschlicherweise zurückzuweisen, wesentlich erhöht. Während es also mit Direct-ML und H0 zum Problem werden könnte, dass ein falsch spezifiziertes Modell bei hohen Missingquoten angenommen wird, kann es sein, dass mit EM, EMB, FCS, MNV und PMM ein Modell zurückgewiesen wird, das eigentlich korrekt ist. Tabelle 10 fasst diese Ausführungen zusammen.

Tabelle 10: Zusammenfassung der Performanz bzgl. der Fit-Indices II

EM, EMB, FCS, MNV, PMM	**SRMR**	**p-Wert**	**RMSEA**	**R. KI (90 %)**	**CFI**
Grundniveau	zufrieden-stellend	zu hohe Ablehnungsraten	zufrieden-stellend	zufrieden-stellend	zufrieden-stellend
Kleine Fallzahl (FZ)	x	x	++	+	+
Zunehmende Asymmetrie	x	x	x	x	x
Steigender Missinganteil (MA)	x	++	++	++	++
Empfehlung	uneingeschränkt einsetzbar	einsetzbar: bei max. 5 % MA	einsetzbar: bei max. 5 % MA + gr. FZ	uneingeschränkt einsetzbar: bei max. 5 % MA; einsetzbar: bei max. 20 % MA + gr. FZ	uneingeschränkt einsetzbar: bei max. 5 % MA; einsetzbar: bei 20 % u. 35 % MA + gr. FZ
Einschränkung	keine	Wahrscheinlichkeit ein korrekt spezifiziertes Modell dennoch zurückzuweisen ist auch mit diesen Empfehlungen gegeben			

Anmerkungen: x: kein Effekt; +/- positiver/negativer Effekt; ++/-- stark positiver/negativer Effekt.

7.3.4 Einordung der Ergebnisse

Die Replikation der Ergebnisse aus vorangegangenen Studien ist mit dieser Arbeit nur teilweise geglückt:

1. Direct-ML ist in vielen Fällen besser als die anderen MDTs (Enders/Bandalos 2001; Peters/Enders 2002; Savalei/Bentler 2005; Shin u. a. 2017).
2. Bei hohen Missinganteilen erhöht sich die Modellzurückweisung; teilweise wird diese extrem (Li 2010; Li/Lomax 2017; Teman 2012).

3. Die Modellzurückweisung ist bei großen Fallzahlen weniger problematisch als bei kleinen und es lassen sich Einflüsse der Variablenverteilungen nachweisen (auch wenn diese hier nicht von Bedeutung sind).

Im Gegensatz zu anderen Studien zeigt sich allerdings nicht, dass mit MNV unter allen Bedingungen zufriedenstellende Ergebnisse vorliegen (Ferro 2014; Wang 2007), dass Direct-ML von den simulierten Bedingungen beeinflusst ist (Enders 2001b; Savalei/Falk 2014), oder dass auch mit EM zufriedenstellende Ablehnungsraten einhergehen (Gold u. a. 2003). Dies könnte aber vor allem daran liegen, dass in all diesen Studien (außer in der Studie von Teman 2012) metrische Variablen geprüft werden. Demnach kann es sein, dass Direct-ML bei metrisch skalierten Variablen durch die Simulationsbedingungen beeinflusst ist, oder dass MNV bei rein metrischen Daten im Hinblick auf die Modellbewertung weniger Probleme bekundet. Zudem kann das gute Ergebnis von EM wohl nicht repliziert werden, weil EM in dieser Arbeit als Einfachimputation und nicht als Indirect-ML eingesetzt wird.

Mithilfe der vorliegenden Arbeit gibt es nun zum ersten Mal eine umfangreiche Evaluation verschiedener MI-Varianten bei quasi-metrischen und gemischten Daten unter unterschiedlichen Bedingungen. So zeigt sich, dass jede MI-Variante, sofern sie keine Modellstruktur berücksichtigt, bei erhöhten Missinganteilen keine akzeptablen Fit-Werte mehr generiert. Das könnte in der Praxis zu Problemen führen. Ob dies mit diesen Varianten auch bei metrischen Daten der Fall ist, wurde nicht geprüft. Hier muss weitere Forschung ansetzen. Zudem ist nur H0 ein geeignetes MI-Verfahren zur Bewertung der Modellgüte, wenn fehlende Werte imputiert werden sollen. Ob H0 unter rein metrischen Daten auch gute Ergebnisse liefert, kann nicht geprüft werden. Auch hierin liegt weiteres Forschungspotential. Letztlich gilt für alle hier untersuchten MDTs: Der Forschungsstand ist noch wenig ausgeprägt; vor allen Dingen gilt dies für die populären Fit-Indices, die in vielen Fällen zur Modellbewertung herangezogen werden.

Abschließend sei auf zwei weitere Anknüpfungspunkte verwiesen. Erstens zeigt sich, dass die Durchschnittswertbildung der Fit-Indices nicht unbedingt problematisch ist, da EM ähnliche Ergebnisse erbringt. Eine Anwendung spezieller Regeln ist also nicht notwendig. Da eine umfangreiche Prüfung von Fit-Indices unter Missing Values vorerst ausgeblieben ist, sollte dort angeknüpft werden. Zweitens werden in der vorliegenden Arbeit mit den Fit-Indices und deren vorgeschlagenen Grenzwerten oftmals inakzeptable Ablehnungsraten generiert, sodass sich die Frage stellt, inwieweit diese Grenzwerte für Modellbewertungen mit sehr hohen Anteilen an Missing Values geeignet sind. Nun garantieren die Fit-Indices nicht, dass auch die Schätzung

der Parameter und Standardfehler einer Modellschätzung gut sind. Sollte sich im weiteren Verlauf der Untersuchungen herausstellen, dass diese, für diejenigen MDTs, welche die Modellstruktur nicht berücksichtigen, inakzeptabel sind, dann ist der Grenzwert für das SRMR zu weich und die anderen Grenzwerte sind angebracht. Sollten allerdings gute Schätzwerte vorliegen, dann sind die 5 %ige Irrtumswahrscheinlichkeit für den p-Wert, die Grenzwerte für RMSEA und dessen Konfidenzintervall und den CFI zu strikt und derjenige für das SRMR geeignet. Ist das der Fall, dann ist es angebracht hier anzusetzen und in Zukunft Grenzwerte zu ermitteln, mit denen die Anwendenden korrekt spezifizierte Modelle aufgrund von Missing Values nicht zurückweisen.

8 Ergebnisse Parameterebene: Parameterbias und Effizienz

In diesem Kapitel findet eine Bewertung der MDTs anhand der geschätzten Parameter statt. Fraglich ist, ob die gute Passung, die mit Direct-ML und mit H0 erzielt wird, sich auch in unverzerrten Parametern widerspiegelt. Gleichzeitig ist zu erörtern, ob die schlechtere Performanz der Fit-Indices nach Anwendung der anderen MDTs auch zu nicht akzeptablen Schätzwerten der Parameter führt. Weiterhin ist bedeutend, ob sich die Unterschiede, die sich in den Ergebnissen zu den Fit-Indices zwischen den einzelnen MDTs zeigen, auch auf der Parameterebene finden. Zunächst wird der Parameterbias deskriptiv aufgearbeitet, bevor er mit Meta-Modellen analysiert wird. Danach werden die MDTs nach deren Effizienz beurteilt.

8.1 Deskriptive Analyse: Relativer Parameterbias

Da sich die Ergebnisse zwischen den drei getesteten Modellen kaum unterscheiden (wenn, dann in der absoluten Höhe der Bias-Werte), werden nur die Ergebnisse für das dritte Modell dargestellt. Die Ergebnisse zu den Faktorladungen des dritten Modells gelten damit auch für das erste Modell, die Ergebnisse zu den Kovarianzen und Strukturpfaden können entsprechend auf das zweite Modell übertragen werden.[97] Bevor nun die Ergebnisse der einzelnen MDTs vorgestellt werden, wird zunächst der relative Bias des Referenzmodells diskutiert (Tabelle 11).

Bis auf die Strukturpfade der beiden 5er skalierten unabhängigen Variablen (x1 und x2) mit dem Faktor, sind alle relativen Bias-Werte des Referenzmodells negativ. Damit werden die Zusammenhänge in Relation zu den Populationswerten in den meisten Fällen unterschätzt. Für die Faktorladungen zeigt sich, dass der Bias größer wird, wenn die Ladungen abnehmen. Zudem lässt sich erkennen, dass bei zunehmender Asymmetrie auch der Bias größer wird. Ein

[97] Analysen zu Modell 1 und Modell 2 finden sich im Anhang O2.3.

Einfluss der Samplegröße zeigt sich hingegen nicht. Weil nur in zwei Fällen Bias-Werte von über 10 %[98] vorliegen, die 15 %-Grenze dabei aber nicht überschritten wird, kann der MLR-Schätzung eine zufriedenstellende Performanz im Hinblick auf die Reproduktion der Populationswerte der Faktorladungen attestiert werden.

Im Gegensatz zu den Faktorladungen sieht das Bild für die Strukturpfade und die Kovarianzen etwas anders aus. Für die Strukturpfade zeigt sich vor allem für den schwachen Pfad (x1>f1) eine erhebliche Überschätzung des Parameters. Damit liegen mit der MLR-Schätzung größere Parameterwerte vor, als im Populationsmodell vorgesehen. Der Dummypfad (x3>f1) wird hingegen stark unterschätzt. Die beiden anderen Pfade werden zufriedenstellend wiedergegeben, auch wenn einer leicht über-, der andere leicht unterschätzt wird. Eine klare Tendenz wonach die Verteilungen einen Einfluss auf die Güte der Strukturpfade haben, lässt sich hier nicht ablesen, auch wenn der Dummypfad mit zunehmend asymmetrischen Verteilungen eher größere Bias-Werte aufweist. Ein Einfluss der Samplegröße zeigt sich wiederum nicht, da die Bias-Werte über die Fallzahlen konstant sind. Neben dem Bias für die Strukturpfade, lassen sich auch für die Kovarianzen Verzerrungen erkennen: Die Kovarianz zwischen den metrischen Variablen (x1-x2) und die Kovarianzen der metrischen Variablen mit der 4er skalierten Variablen (x1-x4 und x2-x4) weisen dabei weniger problematische Bias-Werte auf als die Kovarianzen mit der Dummyvariablen (x3). Hier können Bias-Werte von über 20 % bis hin zu knapp 40 % beobachtet werden. Weiterhin gilt, dass die Bias-Werte nicht durch die Fallzahl bedingt sind, mit zunehmender Asymmetrie aber zunehmen.

Tabelle 11: Relativer Parameterbias des Referenzmodells

Modellkonfiguration			Faktorladungen				Strukturpfade				Kovarianzen					
			ind2 .8	ind3 .7	ind4 .6	ind5 .5	x1>f1 .1	x2>f1 .3	x3>f1 .3	x4>f1 .5	x1-x2 .2	x1-x3 .2	x2-x3 .2	x1-x4 .4	x2-x4 .4	x3-x4 .2
750	skew1	mar1	-4.6	-4.4	-4.3	-4.8	**35.2**	7.1	**-18.0**	-5.4	-9.7	**-23.7**	**-23.2**	**-10.5**	**-10.8**	**-25.5**
	skew2		-5.0	-5.3	-5.8	-6.7	**29.5**	5.7	**-19.4**	-3.9	**-12.9**	**-26.4**	**-26.5**	**-11.7**	**-11.8**	**-28.0**
	skew3		-6.5	-7.8	-9.8	**-11.5**	**29.7**	4.2	**-23.1**	-3.7	**-21.3**	**-36.2**	**-34.3**	**-17.7**	**-17.5**	**-35.3**
250	skew1		-4.6	-4.7	-3.8	-4.9	**32.5**	7.1	**-17.7**	-4.7	-9.5	**-23.8**	**-24.6**	**-10.2**	**-10.6**	**-27.8**
	skew2		-4.9	-5.5	-5.8	-6.5	**26.7**	6.4	**-19.4**	-3.8	**-12.6**	**-26.6**	**-28.0**	**-11.4**	**-11.2**	**-28.1**
	skew3		-6.5	-8.3	-9.9	**-11.4**	**27.8**	4.4	**-23.0**	-3.4	**-20.3**	**-38.2**	**-34.4**	**-17.7**	**-17.0**	**-37.1**

Anmerkungen: Fettdruck: Überschreitung des gesetzten Grenzwertes (± 10 %).

Das Referenzmodell liefert über alle Konfigurationen und Fit-Indices hinweg sehr gute Anpassungswerte. Das führt in einem Anwendungsfall dazu, dass viele Effekte verzerrt geschätzt werden. Tatsächliche Folgen für die Empirie dürften aber, abgesehen von der absoluten Höhe

[98] Der Einfachheit halber werden auch negative Bias-Werte immer in positiver Richtung interpretiert. Die Richtung der Verzerrung wird vermittelt, in dem die Werte als Unter- oder Überschätzungen eingeordnet werden.

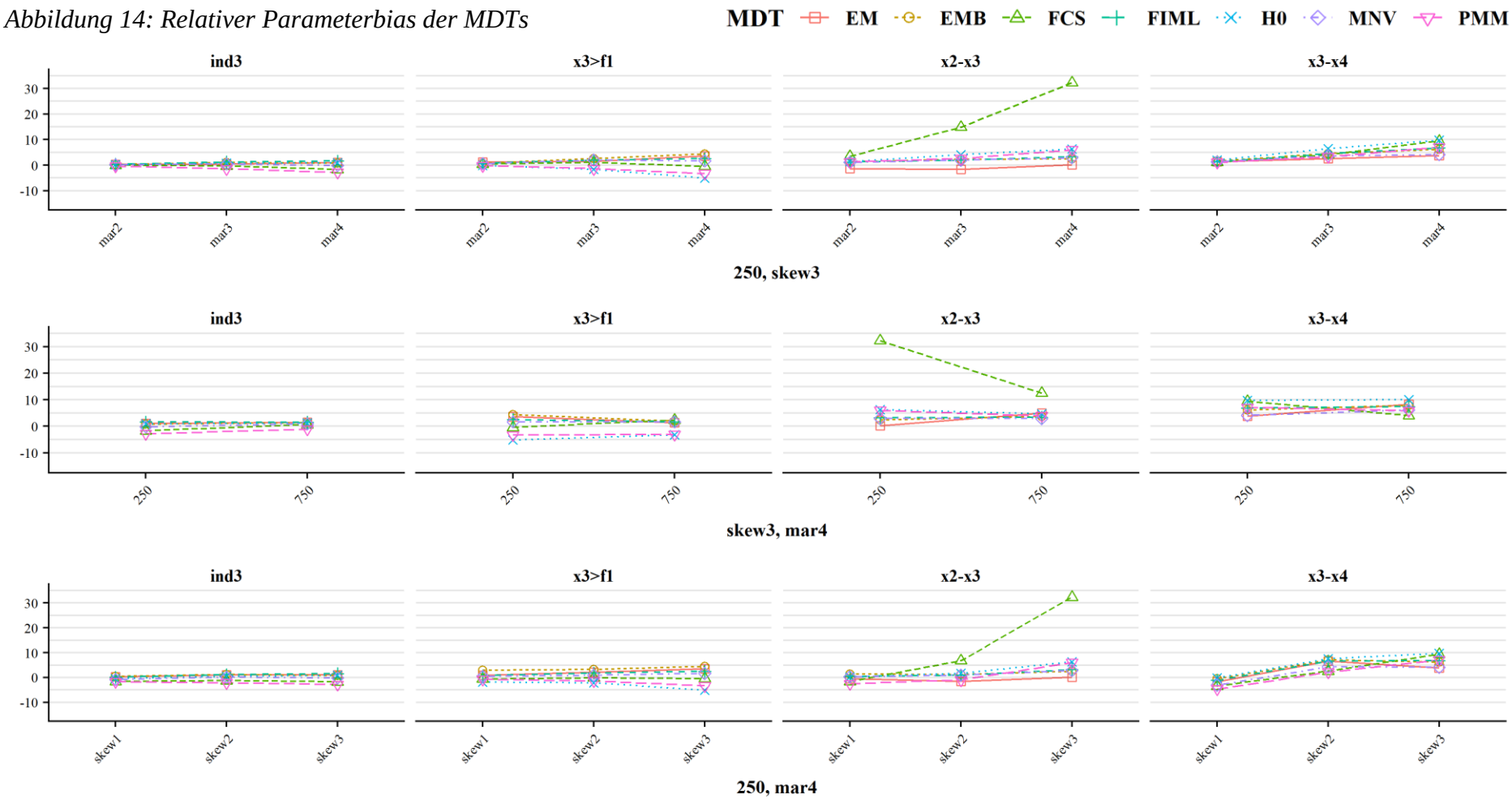
der Koeffizienten, nicht vorliegen. Denn in aller Regel sind die geschätzten Kovarianzen vernachlässigbar und nur die Strukturpfade für die Hypothesentests interessant. Letztlich dürfen dabei erhöhte Bias-Werte vorliegen, sofern der Hypothesentest keinen direkten Vergleich zwischen den Effektstärken vorsieht. Zwar wird der schwache Strukturpfad über-, und der Dummypfad unterschätzt, allerdings dreht sich in beiden Fällen das Vorzeichen nicht. Dadurch wird die Effektrichtung beibehalten. Steht demzufolge nur die Einflussrichtung im Erkenntnisinteresse, dann sind die vorliegenden Befunde eher unproblematisch. Allerdings zeigt sich eben auch, dass die MLR-Schätzung im vorliegenden Fall Probleme hat, die Populationsparameter zufriedenstellend zu reproduzieren. Ob sich andere Schätzmethoden besser eignen, wird nicht geprüft. Dies ist allerdings auch nicht notwendig, denn es steht nicht die Performanz der MLR-Schätzung im Mittelpunkt des Interesses, sondern die Performanz der MDTs. Und um diese beurteilen zu können, ist es lediglich notwendig zu wissen, wie stark das Schätzergebnis durch den ausgewählten Modellschätzer verzerrt ist. Diese Kenntnis erlaubt es, den Einfluss desselben aus den Ergebnissen zu isolieren, sodass der Einfluss der MDTs auf die Verzerrung der Modellschätzung sichtbar wird.

Bevor nun im Folgenden die Bias-Werte der MDTs besprochen werden, sollte auf drei Punkte verwiesen werden. 1) Die deskriptiven Analysen zum Parameterbias orientieren sich an den Analysen zu den Ablehnungsraten. Im Fokus steht demnach das Abschneiden der MDTs zueinander, das Abschneiden der MDTs im Hinblick auf die verschiedenen Teile des Modells (der Faktorladungen, der Strukturpfade und der Kovarianzen) sowie die Identifikation der Einflussgrößen, welche die Performanz der MDTs in Bezug auf die Parameterschätzung am ehesten beeinflussen. Um die Einflussgrößen analysieren zu können, werden die jeweils nicht interessierenden Simulationsgrößen konstant gehalten, wohingegen diejenige Simulationsgröße variiert wird, deren Effekt von Interesse ist. Hierzu werden diejenigen Konfigurationen herangezogen, die bereits bei den Ablehnungsraten analysiert wurden. Im Gegensatz zu den Ablehnungsraten lässt sich dies aber aufgrund der aufgestellten Hypothesen begründen. Denn Hypothese 7 besagt, dass Kombinationen aus mehreren Einflussgrößen, größere Ergebnisverzerrungen ergeben, als Kombinationen, in denen weniger Einflussgrößen vorliegen. Demzufolge sollten bei Kombinationen aus potentiell problematischen Konfigurationsgrößen (Fallzahl bei 250, stark asymmetrische Verteilungen, Missinganteil bei 35 %) größere Bias-Werte auftreten, als bei Kombinationen, in denen eine dieser Größen auf einer weniger problematischen Konfiguration verbleibt (Fallzahl bei 250, stark asymmetrische Verteilungen, Missinganteil bei 5 %). Letztlich lässt sich durch dieses Vorgehen prüfen, ob auch unter möglichst nicht optimalen

Bedingungen gute Parameterschätzwerte mittels der MDTs möglich sind. Sollten sich allerdings keine zufriedenstellenden Bias-Werte unter diesen Konfigurationen ergeben, kann gezeigt werden, welche Einflussgröße am ehesten für die verzerrten Parameterschätzwerte verantwortlich ist. 2) Weil bereits im Referenzmodell erhöhte Bias-Werte beobachtet werden können, werden die Bias-Werte der MDTs um diese Verzerrungen bereinigt. Bei dem vorliegenden Parameterbias der MDTs handelt es sich deshalb nicht mehr um den Populationsbias, sondern um einen bereinigten relativen Bias. Dieser wird auch als Samplingbias bezeichnet. Er gibt nur den Einfluss der MDTs auf die Verzerrung der Schätzung an.[99] 3) Es werden nicht alle Faktorladungen, Kovarianzen und Strukturpfade des dritten Modells aufgearbeitet. Stattdessen werden vier ausgewählt, die repräsentativ für die nicht ausgewählten sind: das ist die Faktorladung des dritten Indikators (ind3), der Strukturpfad zwischen der Dummyvariablen und dem Faktor (x3>f1), die Kovarianz zwischen einer 5er skalierten unabhängigen Variablen und der Dummyvariablen (x2-x3) sowie die Kovarianz der Dummyvariablen und der 4er skalierten Variablen (x3-x4). Der Abbildung 14 lassen sich deren Ergebnisse entnehmen.[100]

[99] Tabellen zum Populationsbias der MDTs des dritten Modells finden sich im Anhang O2.3.1.
[100] Die vollständigen Ergebnisse finden sich im Anhang A5.1.

Abbildung 14: Relativer Parameterbias der MDTs



Anmerkungen: erste Zeile: Konstanthaltung der Fallzahl auf 250 und der Verteilung auf ‚stark asymmetrisch' (skew3); Variation der Missinganteile. Zweite Zeile: Konstanthaltung der Verteilung auf ‚stark asymmetrisch' (skew3) und des Missinganteils auf 35 % (mar4); Variation der Fallzahl. Dritte Zeile: Konstanthaltung der Fallzahl auf 250 und des Missinganteils auf 35 % (mar4); Variation der Verteilungen. Die Fallzahl für EMB in der Konfiguration ‚250, skew3, mar4' beträgt 499.

8.1.1 Relativer Bias in den Faktorladungen und Strukturpfaden

Für den relativen Bias in den Faktorladungen und Strukturpfaden liegt für die einzelnen MDTs ein recht homogenes Bild vor. Größere Unterschiede lassen sich kaum erkennen. In den meisten Fällen gehen mit allen MDTs relativ unverzerrte Parameterschätzwerte einher, denn der Grenzwert von ± 10 % wird nicht überschritten. Grundsätzlich liegen demnach auch in der problematischsten Kombination von Einflussgrößen immer noch zufriedenstellende Parameterschätzwerte für die Faktorladungen und Strukturpfade mit den einzelnen MDTs vor. Neben dem nur geringen Bias, ist die Schätzung der beiden Modellteile zudem auch äußerst robust gegenüber den modellierten Bedingungen: Es lassen sich zwar Tendenzen erkennen, die zeigen, dass mit zunehmendem Missinganteil oder dass mit zunehmender Asymmetrie der Bias etwas größer wird, allerdings ist diese Zunahme vernachlässigbar gering (erste und dritte Reihe[101]). Denn obwohl eine leichte Tendenz erkennbar ist, bleibt der Grenzwert unterschritten. Ein Einfluss der Samplegröße lässt sich dagegen für beide Modellteile nicht beobachten (zweite Reihe). Demnach schaffen es die MDTs, die fehlenden Werte dementsprechend zu handhaben, dass damit unverzerrte Schätzungen der Faktorladungen und der Strukturpfade möglich werden.

8.1.2 Relativer Bias in den Kovarianzen

Außer FCS liefern alle MDTs Bias-Werte, die nur annähernd an die 10 %-Grenze herankommen und diese nicht überschreiten. Dabei macht es keinen Unterschied, um welche der Kovarianzen es sich handelt. Etwas größere Verzerrungen können zwischen der Dummyvariablen und der 4er skalierten Variablen beobachtet werden (x3-x4). Das könnte aber daran liegen, dass in dieser Kovarianz für beide Variablen Missing Values vorlagen. Weil auch hier die 10 %-Grenze nicht überschritten wird, liegen auch dafür zufriedenstellende Ergebnisse vor. Für die Kovarianzen lassen sich leichte Tendenzen erkennen, wonach mit zunehmendem Missinganteil und zunehmender Asymmetrie auch der Bias zunimmt. Ein Einfluss der Samplegröße lässt sich dagegen nicht beobachten: Zu- oder Abnahmen im Bias bei der Variation der Fallzahl lassen sich nicht erkennen. In allen Fällen gilt, dass die Bias-Werte unproblematisch bleiben. Damit schaffen es Direct-ML, EM, EMB, H0, MNV und PMM die fehlenden Werte so zu handhaben, dass damit unverzerrte Parameterschätzwerte einhergehen. Zugleich zeigen sich diese Techniken auch robust gegenüber den modellierten Bedingungen.

[101] Verweise auf Reihen in den Kapiteln 8.1.1 und 8.1.2, beziehen sich immer auf Abbildung 14.

Im Grunde gelten diese Ausführungen auch für FCS, denn auch hierbei können in den allermeisten Fällen unverzerrte Schätzungen für die Kovarianzen erwartet werden. Gleichzeitig nimmt der Bias mit dem Missinganteil und der Asymmetrie zu, aber nicht in dem Maße, dass nicht mehr von zufriedenstellenden Schätzungen auszugehen ist. Im Gegensatz zu den anderen Techniken kann FCS aber in einem Fall die Kovarianz nicht wie gewünscht replizieren: Für die Kovarianz zwischen der Dummyvariablen (x3) und der 5er skalierten unabhängigen Variablen (x2) fällt FCS gegenüber den anderen MDTs ab. Hierbei überschreiten die Bias-Werte die festgelegten Grenzen, wobei aber nur in der Konfiguration mit kleiner Fallzahl, stark asymmetrischer Verteilung und sehr hohem Missinganteil eine deutliche Überschreitung mit 32.3 % auftritt. Anzumerken ist, dass sich Überschreitungen des Grenzwertes für die anderen Kovarianzen oder den Strukturpfad mit der Dummyvariablen nicht beobachten lassen. Wenn demnach FCS für die binäre Variable Probleme mit der logistischen Regressionsschätzung während des Imputationsprozesses hätte, weil aufgrund der Asymmetrie gering besetzte Zellen gegeben sind, dann würden auch für die anderen Kovarianzen mit der Dummyvariablen sowie deren Strukturpfad erhöhte Bias-Werte vorliegen. Letztendlich sind die Kovarianzen eines SE-Modells meist von nachgeordnetem Interesse, sodass dieses Ergebnis für FCS nicht unbedingt problematisch ist, zumal es nur eine einzige Konfiguration dieser Kovarianz betrifft und alle anderen Konfigurationen derselben, sowie alle Modellteile korrekt wiedergegeben werden. Demnach ist es durchaus angebracht, FCS zu verwenden.

Generell ist der Bias, der durch die MDTs den Schätzwerten der MLR-Schätzung hinzugefügt wird, also vernachlässigbar. Damit liefern alle MDTs unter denselben Bedingungen gleich gute Ergebnisse und in nahezu allen Fällen werden die Faktorladungen, die Strukturpfade und die Kovarianzen zufriedenstellend reproduziert.

8.2 Modellbasierte Analyse: Absoluter Parameterbias

Der Fokus dieses Abschnittes richtet sich auf die Identifikation der größten Einflussfaktoren auf die Verzerrung der Parameterschätzwerte. Anders als zuvor werden hierbei nicht mehr nur einzelne Konfigurationen betrachtet, sondern Analysen getätigt, die alle Replikationen und Simulationskonfigurationen gleichzeitig miteinbeziehen. Damit werden in die linearen Regressionsmodelle alle Replikationen aufgenommen.[102] Aus diesem Grund kann in den folgenden Analysen nicht mehr der relative Bias berücksichtigt werden, sondern es muss der absolute Bias

[102] Wiederum werden nur die Ergebnisse zum dritten Modell berichtet, da diese repräsentativ für das erste und zweite Modell sind. Die Ergebnisse zu Modell 1 und Modell 2 finden sich im Anhang O2.4.

herangezogen werden. Bei diesem handelt es sich um ein Maß, das die absolute Verzerrung eines jeweiligen Parameters in einer jeweiligen Replikation angibt.[103] Gleichzeitig liegen mit diesem Maß auch nur positive Parameterverzerrungen vor, was letztlich dazu führt, dass keine Effekte verdeckt werden können, weil sich negative und positive Bias-Werte nicht mehr ausgleichen können. Dadurch sollten dann auch die einflussreichsten Effekte sichtbar werden.

Zudem, so zeigt sich in den deskriptiven Analysen, sind die relativen Bias-Werte für die Faktorladungen, Strukturpfade und Kovarianzen kaum unterschiedlich, weswegen für die folgenden Analysen die absoluten Bias-Werte der einzelnen Modellteile gemittelt werden. Dadurch liegt für jede Replikation ein absolutes Bias-Maß für die Faktorladungen, Strukturpfade und Kovarianzen vor, welches die durchschnittliche Verzerrung im jeweiligen Modellteil angibt. Wie beim relativen Bias, wird auch der absolute Bias um die MLR-Verzerrung bereinigt. Damit handelt es sich auch bei diesem Bias-Wert nicht um den Populations-, sondern um den Samplingbias. Somit dient der durchschnittliche, absolute Samplingbias in den linearen Regressionsmodellen als abhängige Variable, wohingegen die einzelnen Simulationsgrößen mit den, bereits für die logistischen Regressionsmodelle zu den Ablehnungsraten, dargelegten Referenzkategorien die unabhängigen Variablen darstellen. Durch dieses Vorgehen können die durchschnittlichen Effekte der Simulationsgrößen geschätzt werden. Dies erlaubt dann nicht nur spezifische Aussagen hinsichtlich einzelner Konfigurationen, sondern allgemeine Aussagen im Hinblick auf die Verzerrungen in den Parametern. Mit dem Abschluss der modellbasierten Analysen sollte ein umfassendes Bild der Einflussgrößen auf den Bias vorliegen.

Wie bei der modellbasierten Analyse zu den Ablehnungsraten stellt sich auch für die linearen Regressionsmodelle die Frage, was ein interpretierbarer Effekt ist und was nicht. Während bei den logistischen Regressionen vor allem die Stärke des Zusammenhangs und die Ausprägung der AMEs als Indikatoren für die Wichtigkeit der Effekte herangezogen wurden, müssen für die linearen Regressionen andere Maße gefunden werden. Wiederum liegen den Regressionsmodellen sehr viele Fälle zugrunde, sodass auch in diesem Fall kleinste Differenzen bereits signifikant werden können. Dies beschränkt sich nicht nur auf einzelne Effekte im Modell, sondern auch auf die Bewertung des Gesamtmodells anhand des F-Wertes. Gleichzeitig ist auch ein Modellvergleich mithilfe der F-Werte problematisch, da meist auch dabei signifikante Differenzen zu beobachten sind, obwohl die hinzugekommenen Variablen nur bedingt zur Erklärungsleistung beitragen. Aus diesem Grund wird der Fokus verstärkt auf die Änderung von R^2

[103] Damit liegen bspw. für eine Faktorladung in einer Simulationskonfiguration 500 absolute Bias-Werte vor.

gelegt. Cohen (1988) ordnet einen Effekt als gering ein, wenn dieser knapp 2 % an Varianz in der abhängigen Variablen bindet (vgl. ebd.: 413). Effekte, die weniger Erklärungsleistung erbringen, sind demzufolge trivial. Sollte demnach die Hinzunahme einer Variablen in das Modell zu einer Steigerung der Erklärungsleistung um 2 % führen, wird diese Variable als inhaltlich bedeutsam kategorisiert und auch interpretiert. Alle Variablen, die weniger als 2 % an Erklärungsleistung bringen, werden dagegen, auch wenn die Effekte signifikant sein sollten, als nicht bedeutsam erachtet. Gleichzeitig wird ein nicht signifikanter Effekt als nicht interpretationswürdig angesehen. Die inhaltliche Interpretation erfolgt durch die b-Koeffizienten.

Anders als bei den logistischen Regressionen wird hier für jede MDT ein eigenes Modell geschätzt. Das liegt daran, dass für jede MDT Hypothesen abgeleitet wurden. Fraglich ist dann, ob sich die Effekte der Einflussgrößen zwischen den MDTs unterscheiden, oder ob für alle dieselben Effekte vorliegen. Um dies zu prüfen, werden die b-Koeffizienten und deren Konfidenzintervalle herangezogen. Gibt es zwischen den Konfidenzintervallen der interessierenden Effekte keine Überschneidungen, ist der Unterschied signifikant. Überschneiden sich die Konfidenzintervalle, wird der Unterschied in den b-Koeffizienten mit der folgenden Formel[104] auf Signifikanz geprüft:

$$t = \frac{b_1 - b_2}{\sqrt{SE_{b_1}^2 + SE_{b_2}^2}}. \tag{8.1}$$

Sollte der berechnete t-Wert kleiner als -1.96 oder größer als 1.96 sein, dann ist die Differenz zwischen den Werten mit 5 %iger Irrtumswahrscheinlichkeit signifikant. Ist in den nachfolgenden linearen Regressionsmodellen demnach von signifikanten Unterschieden zwischen den Effekten die Rede, so bezieht sich dies immer auf die 5 %ige Irrtumswahrscheinlichkeit und den Unterschied in den b-Koeffizienten.[105]

Bevor nun die Ergebnisse präsentiert werden, ist es sinnvoll auf zwei Problempunkte der Schätzungen einzugehen. Zunächst zeigt sich auch für die Interaktionsmodelle der linearen Regressionen erhöhte Multikollinearität. Diese wird, weil sie durch die Interaktionen künstlich entstanden ist, ignoriert (siehe Kapitel 7.2). Zusätzlich treten in einzelnen Regressionen auch Verletzungen der Homoskedastizitäts- und der Normalverteilungsannahme der Residuen auf.[106]

---

[104] Vgl. Urban/Mayerl (2018: 333).

[105] Weiterhin bestehen bleibt die Problematik der zugrundeliegenden hohen Fallzahl, sodass bei diesen Vergleichen auch die Schätzwerte der b-Koeffizienten berücksichtigt werden.

[106] Heteroskedastizität kann auch auf einen Spezifikationsfehler hindeuten, was verzerrte b-Koeffizienten nach sich ziehen kann. Im vorliegenden Fall kann aber eine Verzerrung der b-Koeffizienten ausgeschlossen werden, denn die b-Koeffizienten einer Regressionsschätzung werden nur dann verzerrt, wenn Variablen in den Modellen

Um dieser Problematik zu entgegnen, werden robuste Standardfehler nach Ecker-Huber-White berechnet (vgl. Urban/Mayerl 2018: 280).

8.2.1 Identifikation der Einflussgrößen: Änderungen im R^2

Das Hauptaugenmerk liegt zunächst auf der Identifikation der interpretationswürdigen Effekte. Hierzu werden die Änderungen im R^2 herangezogen. Die nachfolgende Tabelle 12 stellt den Gewinn an Erklärungsleistung mit Hinzunahme einer jeder weiteren Variablen für jedes geschätzte Regressionsmodell zusammen (für jede der drei abhängigen Variablen ein Modell pro MDT). Dabei werden nur Interaktionen erster Ordnung gelistet. Das liegt daran, dass in allen Modellen kein Zugewinn an Erklärungsleistung mehr vorhanden ist, nachdem die Interaktion der zweiten Ordnung hinzugenommen wird. Um einen Vergleich der verschiedenen Regressionsmodelle durchführen zu können, berücksichtigen alle Modelle alle Interaktionen und damit die gleiche Anzahl an Variablen, auch wenn in einzelnen Modellen die Interaktionen nicht bedeutsam sind.

Zwischen den Modellteilen, abgesehen von der absoluten Höhe im (korrigierten) R^2 und im ΔR^2, gibt es kaum Unterschiede in den erklärungskräftigen Variablen. Demnach sind die Einflussgrößen auf die Verzerrungen in den Faktorladungen, den Strukturpfaden und den Kovarianzen dieselben. Demzufolge ist es ausreichend die Modelle mit dem durchschnittlichen Bias in den Faktorladungen als abhängige Größe zu betrachten. Die Ausführungen können entsprechend übertragen werden.

Werden die Einflussstärken (ΔR^2) der unabhängigen Variablen für EMB, FCS, H0, MNV und PMM betrachtet, wird folgendes offensichtlich: Es lassen sich nur Einflüsse der Haupteffekte beobachten. Der größte Anteil an gebundener Varianz geht dabei auf die Samplegröße zurück, gefolgt von den Verteilungen und den Missinganteilen. Die letzteren Haupteffekte fallen gegenüber der Samplegröße ab und besitzen in etwa die gleiche Erklärungskraft im Hinblick auf die Varianzanteile. Bedeutsame Interaktionen liegen dagegen keine vor.

---

nicht berücksichtigt werden, die hoch mit den Modellvariablen korrelieren. Da allerdings alle unabhängigen Variablen im vorliegenden Fall bekannt sind, weil die Daten anhand dieser definierten Größen simuliert werden und die einzelnen Simulationen unabhängig voneinander sind, existiert keine weitere unabhängige Variable, die mit den Modellvariablen hoch korrelieren könnte. Weiterhin gibt es in den Modellen auch keine Änderung in den b-Koeffizienten, wenn alle möglichen Interaktionsterme berücksichtigt werden. Zusätzlich werden die Regressionen auf den Einfluss von Ausreißern (absolutes standardisiertes Residuum von 2) und einflussreiche Beobachtungen (Cooks Distance mit einem Cut-Off von 4/n; vgl. Bollen/Jackman 1990: 266) getestet. Deren Ausschluss führt zwar zur Aufhebung der Heteroskedastizität, zwischen den Modellen mit gleich verteilten Residuen und denjenigen, die hier betrachtet werden (demnach Regressionen bei denen die Homoskedastizitätsannahme verletzt ist), gibt es aber keine substantiellen Unterschiede in den b-Koeffizienten. Damit verletzen etwaige Modelle zwar die Homoskedastizitätsannahme, ein Spezifikationsfehler und damit verzerrte b-Koeffizienten können aber ausgeschlossen werden.

Tabelle 12: Einflussgrößen auf den Parameterbias: R^2 und Änderung im R^2

abh. Variable: durchschnittlicher absoluter Bias in den Faktorladungen

	EM			EMB			FCS			FIML			H0			MNV			PMM		
+ Variable	**R^2**	**ΔR^2**	**korr. R^2**	**R^2**	**ΔR^2**	**korr. R^2**	**R^2**	**ΔR^2**	**korr. R^2**	**R^2**	**ΔR^2**	**korr. R^2**	**R^2**	**ΔR^2**	**korr. R^2**	**R^2**	**ΔR^2**	**korr. R^2**	**R^2**	**ΔR^2**	**korr. R^2**
250	0.129	**0.129**	0.129	0.279	**0.279**	0.279	0.285	**0.285**	0.285	0.11	**0.110**	0.11	0.262	**0.262**	0.262	0.283	**0.283**	0.283	0.29	**0.290**	0.29
skew	0.202	**0.073**	0.202	0.334	**0.055**	0.334	0.338	**0.053**	0.338	0.205	**0.094**	0.205	0.31	**0.048**	0.31	0.337	**0.054**	0.337	0.343	**0.053**	0.343
mar	0.562	**0.360**	0.562	0.368	**0.034**	0.368	0.374	**0.035**	0.373	0.576	**0.371**	0.576	0.337	**0.027**	0.337	0.37	**0.032**	0.369	0.379	**0.036**	0.379
250:skew	0.566	0.004	0.566	0.37	0.002	0.37	0.376	0.002	0.376	0.58	0.004	0.58	0.34	0.002	0.339	0.372	0.002	0.372	0.382	0.003	0.381
250:mar	0.591	**0.025**	0.591	0.372	0.002	0.372	0.379	0.003	0.378	0.602	**0.022**	0.602	0.341	0.001	0.34	0.374	0.002	0.374	0.385	0.003	0.384
skew:mar	0.608	0.016	0.607	0.376	0.003	0.375	0.382	0.003	0.381	0.622	**0.020**	0.621	0.344	0.003	0.343	0.377	0.003	0.377	0.388	0.002	0.387

abh. Variable: durchschnittlicher absoluter Bias in den Strukturpfaden

250	0.151	**0.151**	0.151	0.291	**0.291**	0.291	0.286	**0.285**	0.286	0.13	**0.130**	0.13	0.27	**0.270**	0.27	0.291	**0.291**	0.291	0.29	**0.290**	0.29
skew	0.196	**0.045**	0.196	0.322	**0.031**	0.322	0.317	**0.032**	0.317	0.191	**0.061**	0.191	0.298	**0.029**	0.298	0.323	**0.031**	0.322	0.318	**0.029**	0.318
mar	0.514	**0.318**	0.514	0.351	**0.029**	0.351	0.345	**0.027**	0.344	0.529	**0.338**	0.529	0.319	**0.020**	0.319	0.35	**0.027**	0.349	0.344	**0.026**	0.344
250:skew	0.518	0.004	0.517	0.353	0.002	0.352	0.347	0.002	0.346	0.534	0.005	0.534	0.322	0.003	0.321	0.351	0.002	0.351	0.346	0.001	0.345
250:mar	0.547	**0.029**	0.547	0.356	0.003	0.355	0.349	0.002	0.348	0.56	**0.026**	0.56	0.323	0.001	0.322	0.354	0.002	0.353	0.348	0.002	0.347
skew:mar	0.557	0.009	0.556	0.358	0.002	0.357	0.352	0.002	0.351	0.572	0.012	0.572	0.324	0.000	0.323	0.356	0.002	0.356	0.349	0.001	0.349

abh. Variable: durchschnittlicher absoluter Bias in den Kovarianzen

250	0.156	**0.156**	0.156	0.326	**0.326**	0.326	0.32	**0.320**	0.32	0.142	**0.142**	0.142	0.337	**0.336**	0.336	0.324	**0.324**	0.324	0.331	**0.331**	0.331
skew	0.212	**0.056**	0.212	0.356	**0.030**	0.356	0.353	**0.033**	0.353	0.213	**0.071**	0.213	0.36	**0.023**	0.36	0.354	**0.030**	0.354	0.352	**0.021**	0.351
mar	0.577	**0.364**	0.576	0.381	**0.025**	0.381	0.379	**0.026**	0.379	0.59	**0.376**	0.589	0.384	**0.024**	0.384	0.377	**0.023**	0.377	0.368	0.016	0.368
250:skew	0.581	0.004	0.58	0.383	0.002	0.382	0.381	0.002	0.381	0.594	0.004	0.594	0.385	0.001	0.384	0.379	0.002	0.379	0.369	0.001	0.369
250:mar	0.602	**0.021**	0.602	0.384	0.001	0.384	0.382	0.001	0.382	0.617	**0.023**	0.616	0.386	0.001	0.385	0.38	0.001	0.379	0.37	0.000	0.369
skew:mar	0.611	0.009	0.611	0.386	0.001	0.385	0.385	0.002	0.384	0.63	0.013	0.629	0.387	0.001	0.386	0.382	0.001	0.381	0.37	0.000	0.369

Anmerkungen: Fettdruck: Erklärungskraft der Variablen ≥ 2 %.

Für Direct-ML und EM kann ein anderes Ergebnis beobachtet werden. Zum einen ist der durchschnittliche absolute Bias der Faktorladungen von den Missinganteilen beeinflusst. Mit dieser Variablen wird der größte Anteil an Varianz in der abhängigen Variablen gebunden. Zum anderen sind bei beiden die Samplegrößen und die Variablenverteilungen Einflussgrößen, wobei die Verteilungsvariable von den Haupteffekten am wenigsten Varianzanteile bindet. Neben den Haupteffekten lässt sich für Direct-ML und EM auch die Interaktion zwischen der Fallzahl und den Missinganteilen als bedeutsam einstufen. Durch deren Hinzunahme zum Modell kann nochmals etwas über 2 % an Varianz gebunden werden. Der Einfluss der Interaktion zwischen der Fallzahl und den Verteilungen sowie den Verteilungen und den Missinganteilen, obwohl der Grenzwert für Direct-ML knapp erreicht wird, ist dagegen vernachlässigbar. Für alle MDTs geht der Hauptanteil an Erklärungskraft auf die Haupteffekte zurück und nur für Direct-ML und EM gibt es eine substantielle Interaktion: diejenige zwischen der Fallzahl und den Missinganteilen.

8.2.2 Inhaltliche Interpretation der Einflussgrößen

In diesem Unterkapitel werden die identifizierten Einflussgrößen inhaltlich interpretiert. Weil es dabei in den inhaltlichen Schlussfolgerungen zwischen den einzelnen abhängigen Variablen keine Unterschiede gibt, werden in diesem Kapitel nur die Modelle mit dem Bias in den Faktorladungen als abhängige Größe berücksichtigt. Wiederum können die Ausführungen auf die anderen Modelle übertragen werden.[107]

Anhand der F-Werte in Tabelle 13 wird ersichtlich, dass die Gesamtmodelle allesamt signifikant sind ($p < .001$).[108] Am korrigierten R^2 lässt sich zudem erkennen, dass die Varianz durch die einzelnen unabhängigen Variablen unterschiedlich stark gebunden werden kann. Im vorliegenden Fall bedeutet das, dass Modelle mit geringerem R^2 robuster gegenüber den Simulationskonfigurationen sind, als Modelle mit höheren Werten. In letzteren kann durch die unabhängigen Variablen mehr Varianz in der abhängigen Variablen gebunden werden. Ein direkter Vergleich zwischen den R^2-Werten einzelner linearer Regressionsmodelle ist allerdings problematisch, da R^2 von den Varianzen der x- als auch der y-Variablen und den Einflussstärken abhängt und diese für Regressionsschätzungen mit gleichen Variablen in verschiedenen Datensätzen jeweils anders sein können. Eine Interpretation des Wertes sollte demnach nur vor-

[107] Die Modelle für die Strukturpfade und Kovarianzen sowie die Tests auf Unterschiede in den b-Koeffizienten lassen sich dem Anhang A5.2 entnehmen.
[108] Anmerkung: F-Statistik in allen Modellen: *** ($p < 0.001$).

sichtig zwischen Modellschätzungen erfolgen (vgl. Urban/Mayerl 2018: 56 ff.). Nichtsdestotrotz zeigt sich, dass Direct-ML und EM etwas weniger robust sind als EMB, FCS, H0, MNV und PMM. Für letztere lassen sich zwischen 34 % und 39 % an Varianz in der abhängigen Variablen binden, während für Direct-ML und EM über 60 % an Varianz gebunden werden.

Tabelle 13: Ergebnisse der modellbasierten Analyse zum Parameterbias

	abh. Variable: durchschnittlicher absoluter Bias in den Faktorladungen						
	b (95 % KI)						
	EM	EMB	FCS	FIML	H0	MNV	PMM
Constant	0.77***	2.82***	2.81***	0.56***	2.79***	2.82***	2.83***
	(0.73, 0.81)	(2.71, 2.92)	(2.71, 2.91)	(0.53, 0.59)	(2.69, 2.89)	(2.71, 2.92)	(2.73, 2.93)
250	0.34***	2.05***	2.05***	0.26***	1.97***	2.05***	2.04***
	(0.27, 0.41)	(1.90, 2.20)	(1.90, 2.20)	(0.20, 0.31)	(1.81, 2.12)	(1.89, 2.20)	(1.88, 2.19)
skew2	0.14***	0.23**	0.24**	0.14***	0.20**	0.23**	0.23**
	(0.08, 0.20)	(0.08, 0.37)	(0.10, 0.39)	(0.10, 0.19)	(0.05, 0.34)	(0.08, 0.37)	(0.08, 0.37)
skew3	0.22***	0.70***	0.70***	0.25***	0.60***	0.70***	0.69***
	(0.15, 0.29)	(0.54, 0.86)	(0.54, 0.86)	(0.20, 0.31)	(0.44, 0.76)	(0.54, 0.86)	(0.53, 0.85)
mar3	0.75***	0.20**	0.21**	0.60***	0.24**	0.19*	0.19*
	(0.69, 0.82)	(0.06, 0.35)	(0.06, 0.36)	(0.55, 0.65)	(0.09, 0.39)	(0.05, 0.34)	(0.05, 0.34)
mar4	1.36***	0.45***	0.47***	1.15***	0.43***	0.44***	0.48***
	(1.28, 1.45)	(0.29, 0.60)	(0.31, 0.63)	(1.08, 1.22)	(0.28, 0.59)	(0.29, 0.60)	(0.32, 0.63)
250:skew2	0.18***	0.24**	0.24**	0.16***	0.29**	0.25**	0.29**
	(0.08, 0.28)	(0.07, 0.42)	(0.06, 0.41)	(0.08, 0.25)	(0.12, 0.47)	(0.08, 0.42)	(0.11, 0.46)
250:skew3	0.57***	0.56***	0.58***	0.47***	0.56***	0.57***	0.61***
	(0.45, 0.68)	(0.36, 0.75)	(0.39, 0.78)	(0.37, 0.56)	(0.37, 0.75)	(0.37, 0.76)	(0.41, 0.80)
250:mar3	0.72***	0.11	0.12	0.55***	0.02	0.10	0.18*
	(0.64, 0.81)	(-0.07, 0.29)	(-0.06, 0.30)	(0.48, 0.62)	(-0.16, 0.20)	(-0.07, 0.28)	(0.002, 0.36)
250:mar4	1.40***	0.51***	0.60***	1.11***	0.34***	0.51***	0.63***
	(1.28, 1.52)	(0.32, 0.70)	(0.41, 0.80)	(1.02, 1.21)	(0.15, 0.53)	(0.32, 0.70)	(0.43, 0.82)
skew2:mar3	0.23***	0.11	0.10	0.28***	0.07	0.12	0.14
	(0.14, 0.32)	(-0.09, 0.32)	(-0.10, 0.30)	(0.20, 0.35)	(-0.14, 0.27)	(-0.08, 0.32)	(-0.06, 0.34)
skew3:mar3	0.68***	0.31**	0.30**	0.72***	0.31**	0.30**	0.34**
	(0.58, 0.79)	(0.08, 0.53)	(0.07, 0.52)	(0.63, 0.81)	(0.09, 0.53)	(0.07, 0.52)	(0.12, 0.57)
skew2:mar4	0.42***	0.29**	0.21	0.49***	0.26*	0.25*	0.25*
	(0.30, 0.55)	(0.08, 0.51)	(-0.01, 0.43)	(0.39, 0.59)	(0.04, 0.47)	(0.04, 0.46)	(0.03, 0.47)
skew3:mar4	1.35***	0.81***	0.75***	1.28***	0.73***	0.77***	0.73***
	(1.21, 1.50)	(0.57, 1.06)	(0.51, 0.99)	(1.16, 1.41)	(0.49, 0.98)	(0.54, 1.01)	(0.49, 0.97)
Fallzahl	9000	8999	9000	9000	9000	9000	9000
F-Statistik	1188.242	403.069	406.957	1297.816	357.584	404.251	416.173
R^2	0.608	0.376	0.382	0.622	0.344	0.377	0.388
korr. R^2	0.607	0.375	0.381	0.621	0.343	0.377	0.387
SEE	1.133	1.887	1.89	0.948	1.875	1.872	1.895

Note: *p<0.05; **p<0.01; ***p<0.001

Für alle MDTs gilt, dass für kleine Fallzahlen im Vergleich zu großen Fallzahlen höhere Bias-Werte vorliegen, sofern alle anderen Variablen konstant gehalten werden. Gleiches gilt auch für die Verteilungen. Gegenüber der symmetrischen Kategorie nimmt der Bias mit zunehmend asymmetrischen Verteilungen zu. Auch der Missinganteil hat einen Einfluss auf die Höhe des

Bias. Dieser erhöht sich für die 20 %-Missingkategorie bzw. 35 %-Missingkategorie im Vergleich zur 5 %-Kategorie. Damit werden die Verzerrungen in den Faktorladungen bei kleiner werdenden Fallzahlen, stärker asymmetrischen Verteilungen oder größeren Anteilen an fehlenden Werten größer. Für Direct-ML und EM kann zudem die Interaktion zwischen der Fallzahl und den Missinganteilen als bedeutsam eingestuft werden. Das bedeutet, dass die Fallzahl den Einfluss der Missinganteile moderiert; dieser ist bei kleinen Fallzahlen größer als bei großen Fallzahlen. Dementsprechend ist die Schätzung mittels Direct-ML und EM bei größeren Fallzahlen und höheren Missinganteilen weniger verzerrt.

Zuletzt werden die Effekte auf Unterschiede getestet (Tabelle 14). Hier zeigt sich, dass für alle MI-Techniken keine signifikanten Differenzen vorliegen. Demnach ist die Performanz der Parameterschätzung für alle fünf MI-Techniken in gleichem Maße durch die Fallzahl, die Verteilungen und die Missinganteile beeinflusst. Im Vergleich der MI-Techniken zu den ML-Verfahren lassen sich aber Unterschiede ausmachen. Der b-Wert ist für die Samplegröße und für die stark asymmetrischen Verteilungen höher. Somit ist der Bias bei kleineren Fallzahlen oder stark asymmetrischen Verteilungen für die MI-Techniken höher als bei Direct-ML oder EM. Andererseits liegen für die Missinganteile weniger ausgeprägte Effekte vor. Damit fällt der Bias für EMB, FCS, H0, MNV und PMM unter höheren Missinganteilen kleiner aus als mit Direct-ML und EM. Der Koeffizientenvergleich zwischen Direct-ML und EM zeigt zudem, dass signifikante Unterschiede für die Missinganteile und die Interaktion zwischen der Fallzahl und den Missinganteilen vorliegen. Alle Effekte sind für EM ausgeprägter. Das bedeutet, dass der Bias für EM durchschnittlich etwas höher ausfällt als für Direct-ML.

Tabelle 14: Test auf Unterschiede in den b-Koeffizienten

Differenzen der b-Koeffizienten zwischen den MDTs mit 5 %iger Irrtumswahrscheinlichkeit signifikant?

Effekt	MNV vs. EMB	MNV vs. FCS	MNV vs. H0	MNV vs. PMM	MI (MNV) vs. EM	MI (MNV) vs. FIML	FIML vs. EM
250	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
skew2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
skew3	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
mar3	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
mar4	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
250:skew2	--	--	--	--	--	--	--
250:skew3	--	--	--	--	--	--	--
250:mar3	--	--	--	--	--	--	TRUE
250:mar4	--	--	--	--	--	--	TRUE
skew2:mar3	--	--	--	--	--	--	--
skew3:mar3	--	--	--	--	--	--	--
skew2:mar4	--	--	--	--	--	--	--
skew3:mar4	--	--	--	--	--	--	--

Anmerkungen: FALSE: Unterschied nicht signifikant; TRUE: signifikanter Unterschied; -- kein als substantiell einzuordnender Effekt (ΔR^2) und deshalb auch nicht geprüft bzw. irrelevant.

8.2.3 Zusammenfassung der modellbasierten Analyse

Zusammenfassend lässt sich der modellbasierten Analyse folgendes entnehmen:

- Die MI-Techniken sind gegenüber den modellierten Bedingungen robuster als die ML-Verfahren. Für die letzteren Techniken werden in allen Modellen mehr Varianzanteile durch die unabhängigen Variablen gebunden.
- Als Haupteinflussgrößen können für die ML-Verfahren, mit abnehmender Erklärungskraft, die Missinganteile, die Samplegröße und die Verteilungen sowie die Interaktion zwischen der Fallzahl und den Missinganteilen identifiziert werden.
- Für die MI-Techniken bindet die Fallzahl die größten Anteile an Varianz, gefolgt von den Verteilungen und den Missinganteilen.
- Es gilt für alle MDTs, dass mit abnehmender Fallzahl der Bias größer wird, womit ein Nachweis für die Konsistenz der MDTs vorliegt. Zudem nimmt der Bias zu, wenn die Missinganteile größer oder die Verteilungen asymmetrischer werden. Für Direct-ML und EM gilt weiterhin, dass der Einfluss der Missings bei größeren Fallzahlen weniger ausgeprägt ist, als bei kleinen.
- Keine Unterschiede in den Einflussgrößen können für die fünf MI-Techniken festgestellt werden. Alle MI-Techniken sind in gleichem Maße durch die Simulationskonfigurationen beeinflusst. Das wiederum lässt den Schluss zu, dass diese Techniken unter denselben Bedingungen zu annähernd gleichen Ergebnisse führen.
- Die Parameterschätzung mit Direct-ML ist im Vergleich zu EM etwas besser; EM weist in allen Fällen stärker ausgeprägte b-Koeffizienten auf als Direct-ML.

8.3 Deskriptive Analyse: Relative Effizienz

Welche der getesteten MDTs unter den modellierten Bedingungen die effizienteren/genaueren Ergebnisse erbringt, lässt sich mithilfe der relativen Effizienz erfassen. Bevor allerdings die Ergebnisse präsentiert werden können, müssen hierzu drei Punkte diskutiert werden: die zu wählende Konstante, die Interpretation und die Berechnung der relativen Effizienz.

1) Für die relative Effizienz muss eine Konstante gewählt werden, zu welcher die MDTs in Relation gesetzt werden können, da ansonsten immer nur ein Vergleich zweier MDTs zueinander möglich ist. Als Konstante wird in diesem Fall Direct-ML gewählt. Dies hat mehrere Gründe: Erstens produzieren die einzelnen MDTs kaum unterschiedliche Ergebnisse, sodass jede MDT gewählt werden könnte. Zweitens erweist sich Direct-ML zusammen mit H0 in Be-

zug auf die Fit-Indices als beste MDT. Drittens liegt für Direct-ML im Gegensatz zu den verschiedenen MI-Varianten mehr Forschungsarbeit vor, sodass sich die Ergebnisse der vorliegenden Arbeit besser einordnen lassen. Viertens kann direkt Bezug auf die Ausführungen in Kapitel 6.3.3 genommen werden.

2) Es ist auf die Interpretation der relativen Effizienz einzugehen, da im vorliegenden Fall nicht komplett unverzerrte Parameterschätzungen vorliegen. Aus diesem Grund gibt der, für die Berechnung notwendige MSE einer MDT nicht deren Effizienz wieder, sondern deren Genauigkeit. Da sich in den Analysen zum relativen und absoluten Bias allerdings zeigt, dass die MDTs ähnliche Verzerrungen aufweisen und die Unterschiede zwischen ihnen minimal sind, können Unterschiede in den MSE-Werten der MDTs eher auf die Varianz der Schätzwerte zurückgeführt werden. Dementsprechend ist in den folgenden Analysen, trotz vorliegendem Bias, nicht von der Genauigkeit der Schätzungen, sondern von deren Effizienz die Rede.

3) Zuletzt ist anzumerken, dass auch in diesem Fall der Samplingbias in der Berechnung des MSE berücksichtigt wird und nicht der Populationsbias. Gleichzeitig wird nicht die relative Effizienz für jeden Parameter untersucht, sondern die durchschnittliche relative Effizienz der Parameter in einem jeweiligen Modellteil. Das liegt daran, dass die einzelnen Werte relativ konstant sind und sich in ihrer Richtung und Höhe kaum voneinander unterscheiden. Ebenfalls werden wieder nur die Ergebnisse für Modell 3 berichtet. Sie gelten gleichermaßen für das erste und zweite Modell.[109]

8.3.1 Ergebnisse zur relativen Effizienz der MDTs

Tabelle 15 können die Ergebnisse zur relativen Effizienz der MDTs für die Modellteile des dritten Modells entnommen werden. Für die Faktorladungen zeigt sich, dass Direct-ML gegenüber EM im Durchschnitt effizienter ist, wobei die Unterschiede größer werden, wenn der Missinganteil ansteigt: Wird dieser höher, so sollte von EM auf Direct-ML gewechselt werden. Um mit EM unter erhöhten Missinganteilen gleich effiziente Schätzungen der Faktorladungen zu erzielen, müsste die Fallzahl um über 20 % größer werden. Zudem ist Direct-ML auch etwas besser als FCS und PMM. In allen Fällen liegt mit Direct-ML eine effizientere Schätzung vor; Grenzwertüberschreitungen lassen sich aber nur bei kleinen Fallzahlen beobachten. Demnach ist es unter kleinen Fallzahlen besser, Direct-ML einzusetzen. Bei größeren Fallzahlen gibt es dagegen keine Effizienzgewinne mit Direct-ML gegenüber diesen MDTs. Die Relation

[109] Modell 1 und Modell 2: Anhang O2.5.

Tabelle 15: Relative Effizienz der MDTs

			relative Effizienz (in Relation zu Direct-ML)																	
			Faktorladungen						Strukturpfade						Kovarianzen					
			EM	**EMB**	**FCS**	**H0**	**MNV**	**PMM**	**EM**	**EMB**	**FCS**	**H0**	**MNV**	**PMM**	**EM**	**EMB**	**FCS**	**H0**	**MNV**	**PMM**
		mar2	1.04	1.00	1.00	1.00	1.00	1.00	1.04	1.00	1.00	1.00	1.00	1.01	1.05	1.00	1.00	1.00	1.00	1.00
	skew1	mar3	**1.22**	1.02	1.02	1.00	1.00	1.02	**1.16**	1.02	1.01	1.00	1.01	1.02	**1.17**	1.01	1.01	1.01	1.00	1.00
		mar4	**1.21**	1.03	1.05	1.01	1.02	1.06	**1.20**	1.03	1.03	1.00	1.02	1.05	**1.21**	1.01	1.02	1.00	1.00	1.01
		mar2	1.07	1.00	1.01	1.00	1.00	1.02	1.05	1.00	1.00	0.99	1.01	1.01	1.07	1.00	1.00	1.00	1.00	0.98
750	skew2	mar3	**1.17**	1.01	1.01	1.00	1.00	1.05	**1.16**	1.02	1.02	0.98	1.01	1.03	**1.14**	1.01	1.00	1.00	1.00	0.95
		mar4	**1.21**	1.04	1.02	0.98	1.00	1.07	**1.17**	1.04	1.04	0.96	1.02	1.03	**1.20**	1.01	1.01	1.01	0.99	0.92
		mar2	1.06	1.00	1.00	1.00	1.00	1.02	1.06	1.00	1.00	0.99	1.01	1.01	1.06	1.00	1.01	1.00	1.00	0.98
	skew3	mar3	**1.15**	1.02	0.99	0.98	1.00	1.01	**1.15**	1.03	1.05	0.96	1.02	1.03	**1.14**	1.01	1.04	1.01	1.01	0.91
		mar4	**1.18**	1.03	1.00	0.99	1.00	1.03	**1.18**	1.04	1.07	0.93	1.04	1.03	**1.18**	1.01	1.08	1.01	1.00	**0.85**
		mar2	1.05	1.01	1.00	1.00	1.01	1.02	1.06	1.00	1.00	1.00	1.00	1.00	1.08	1.01	1.00	1.00	1.00	1.00
	skew1	mar3	**1.18**	1.01	1.04	1.01	1.01	1.06	**1.20**	1.02	1.00	0.99	1.01	1.02	**1.17**	1.01	1.00	1.00	1.00	1.00
		mar4	**1.24**	1.04	**1.12**	1.01	1.04	**1.15**	**1.24**	1.06	1.01	0.98	1.02	1.04	**1.19**	1.02	0.99	1.01	1.00	0.99
		mar2	1.05	1.00	1.01	1.00	1.01	1.03	1.08	1.01	1.01	0.99	1.00	1.01	1.05	1.00	0.99	1.00	1.00	0.99
250	skew2	mar3	**1.17**	1.01	1.03	0.99	1.00	1.08	**1.17**	1.03	1.01	0.98	1.01	1.01	**1.20**	1.01	0.99	0.99	0.99	0.95
		mar4	**1.21**	1.04	1.09	0.98	1.02	**1.15**	**1.25**	1.06	1.03	0.95	1.02	1.01	**1.16**	1.02	0.97	0.99	0.97	**0.89**
		mar2	1.08	1.00	1.01	1.00	1.01	1.03	1.07	1.00	1.00	0.99	1.01	1.00	1.07	1.00	0.99	0.99	1.00	0.97
	skew3	mar3	**1.15**	1.01	1.04	0.99	1.00	1.07	**1.19**	1.04	1.02	0.95	1.03	1.00	**1.14**	1.00	1.01	0.99	0.99	**0.89**
		mar4	**1.23**	1.04	**1.11**	0.96	1.04	**1.13**	**1.21**	1.08	1.05	**0.90**	1.04	0.97	**1.16**	0.99	1.02	0.98	0.96	**0.82**

Anmerkungen: Fettdruck: Unterschiede in der Effizienz zwischen den MDTs um ± 10 %.

zwischen Direct-ML und EMB oder Direct-ML und MNV übersteigt den gesetzten Grenzwert nicht, sodass mit allen Techniken durchschnittlich gleich effiziente Schätzungen einhergehen. H0 liefert dagegen bei erhöhten Anteilen an Missing Values (ab 20 %) etwas weniger Varianz in den Schätzwerten als Direct-ML (die Grenzwerte werden aber nicht überschritten).

Für die Strukturpfade lässt sich ähnliches beobachten. Mit zunehmendem Missinganteil liegt mittels Direct-ML im Durchschnitt eine effizientere Schätzung vor als mit EM. In Relation zu EMB, FCS, MNV und PMM lassen sich dagegen keine Effizienzgewinne durch Direct-ML erzielen. In keiner Konfiguration wird der gesetzte Grenzwetz überschritten (auch wenn die Relationen eher für Direct-ML sprechen; die Werte sind meist größer als 1.0). Mit H0 wird bei zunehmenden Missinganteilen das Verhältnis immer kleiner, sodass H0 tendenziell effizienter als Direct-ML ist. Auch im Hinblick auf die Kovarianzen ergibt sich ein vergleichbares Bild. EM schneidet durchschnittlich etwas schlechter ab als Direct-ML, wobei auch hier erst bei 20 % an Missing Values der Grenzwert überschritten wird. Keine Unterschiede gibt es von Direct-ML zu EMB, zu FCS sowie zu MNV oder auch zu H0. Anders als zuvor ist H0 damit nicht effizienter als Direct-ML. Dagegen lassen sich für PMM aber Effizienzgewinne beobachten: Bei großen Missinganteilen schneidet PMM besser ab als Direct-ML.

Zusammenfassend lassen sich sechs Erkenntnisse ableiten: 1) Direct-ML erweist sich durchgehend besser als EM. Ab 20 % an Missing Values sollte Direct-ML anstatt EM verwendet werden, da in solchen Fällen die Fallzahl für gleich effiziente Schätzungen mit EM um ca. 20 % größer sein müsste. 2) Für alle MDTs gilt, dass die Verteilungen keinen Einfluss auf die relative Effizienz haben. In kaum einer Konfiguration kann eine nennenswerte Änderung beobachtet werden, wenn die Verteilungen asymmetrischer werden. 3) Weiterhin gilt für alle MDTs (außer EM), dass unter großen Fallzahlen in etwa gleich effiziente Schätzungen erzielt werden können. 4) Zwischen Direct-ML und EMB sowie Direct-ML und MNV gibt es auch bei kleinen Fallzahlen keine Unterschiede. In allen Fällen wird der Grenzwert nicht überschritten. FCS liefert bei kleinen Fallzahlen eher gleich effiziente Schätzungen wie Direct-ML, auch wenn der Grenzwert für die Faktorladungen in zwei Fällen knapp überschritten wird. 5) H0 liefert tendenziell effizientere Schätzungen: die Relationen liegen durchgehend bei 1.0 und darunter. 6) PMM ist bei kleinen Fallzahlen wenig effizient bei der Schätzung der Faktorladungen, dafür aber bei den Kovarianzen. Weil die Faktorladungen von inhaltlich größerer Bedeutsamkeit sind, stellt dieses Ergebnis aber keinen Vorteil für PMM dar. Für die Strukturpfade gibt es dagegen keinen Unterschied zwischen Direct-ML und PMM.

8.4 Parameterbias und Effizienz: Ergebnisdiskussion und Einordnung

Zwischen den einzelnen MDTs bestehen nur geringe Unterschiede. Zudem lassen sich diese Ergebnisse, unabhängig von den getesteten Modellen und unabhängig von den jeweiligen Modellteilen beobachten: Für die MDTs ist es demnach unerheblich, welches Modell nach der Behandlung der fehlenden Werte geschätzt wird. Das gilt für Direct-ML und H0 aber nur, wenn das Modell zur Behandlung der fehlenden Werte korrekt ist; wie die Parameter ausfallen, wenn dieses Modell missspezifiziert ist, kann nicht beurteilt werden. Im Gegensatz dazu treffen die anderen MDTs eine solche Annahme nicht. Wenn nur die Parameter einer Modellschätzung im Fokus des Interesses stehen und das spezifizierte Modell bei der Behandlung der fehlenden Werte unbekannt sein sollte, ist es angebracht MDTs zu verwenden, die keine Annahmen über das Modell treffen. Das wäre der Fall, wenn der Imputations- vom Analyseprozess losgelöst ist und die am jeweiligen Prozess beteiligten Forschenden keinen Zugriff auf die Daten haben, die jeweils dem anderen Prozess zugrunde liegen.

Zwar liefern alle MDTs in nahezu allen Fällen unverzerrte und effiziente Schätzungen, zwei Ausnahmen liegen jedoch vor: Das betrifft die Schätzung der Kovarianzen mittels FCS und in einigen Fällen die Effizienz der MDTs. Denn während die Verzerrungen in den Faktorladungen und in den Strukturpfaden mit FCS zufriedenstellend gering sind, können in Bezug auf die Kovarianzen Unterschiede zwischen FCS und den anderen MDTs ausgemacht werden, denn mit FCS können dabei größere Verzerrungen auftreten. In der vorliegenden Arbeit kann eine Überschätzung einer Kovarianz der Dummyvariablen beobachtet werden. Einzuschränken ist allerdings, dass eine erhebliche Überschätzung nur in einer einzigen Konfiguration auftritt. Dass dies allerdings ein Ergebnis ist, das auf die binär logistischen Regressionen von FCS zurückzuführen ist, kann nicht behauptet werden. Ansonsten wäre dieses Ergebnis, vor allem bei hohen Missingquoten, auch in den anderen Konfigurationen dieser Kovarianz aufgetreten. Weiterhin wären auch die anderen Kovarianzen der Dummyvariablen und deren Strukturpfad davon betroffen gewesen; diese werden aber korrekt wiedergegeben. Demzufolge ist dieses Ergebnis wenig problematisch. Wird zudem berücksichtigt, dass es für die Hypothesentests in SE-Modellen die Faktorladungen und die Strukturpfade sind, die von Interesse sind, und diese mit FCS zufriedenstellend geschätzt werden können, dann kann auch FCS empfohlen werden.

Weiterhin können Unterschiede zwischen den MDTs in der Effizienz der Schätzungen ausgemacht werden. Hierbei schneidet EM schlechter ab als Direct-ML. Denn ab 20 % an Missing Values liefert EM größere Varianzen in den Schätzwerten, auch wenn die durchschnittlichen

Verzerrungen vernachlässigbar bleiben. Dementsprechend könnten bei wiederholten Imputationen mit EM andere Ergebnisse beobachtet werden. Es ist folglich sinnvoll, die Imputation mittels EM bei erhöhten Quoten an Missings mit unterschiedlichen Startwerten für den Imputationsprozess vorzunehmen, um auszuschließen, dass es sich bei den geschätzten Parametern nicht um Ausreißer handelt. Neben EM weisen auch mit PMM die Schätzwerte für die Faktorladungen etwas größere Varianzen auf als mit Direct-ML. Demzufolge ist die Imputation der fehlenden Werte auf Indikatoren in Modellen mit latenten Faktoren, sollten diese Indikatoren quasi-metrisch skaliert sein, mittels linearer Regressionen in Bezug auf die Effizienz etwas besser als mit PMM, da größere Varianzen mit FCS nicht auftreten. Soll für die Imputationen aber PMM beibehalten werden, ist eine Möglichkeit deren Effizienz zu steigern, die Anzahl an zu imputierenden Datensätzen (*m*) zu erhöhen. Dadurch werden in einem Anwendungsfall die durchschnittlichen Parameter einer Modellschätzung nach dem Imputationsprozess durch eine größere Anzahl an Datensätzen bestimmt, was den Einfluss von potentiellen Ausreißern reduzieren und dadurch, zu weniger verzerrten Schätzungen für die Faktorladungen führen könnte. Wird dies auf den MSE im Kontext dieser Arbeit übertragen, dann führt die Erhöhung von *m* zu einer Reduktion des MSEs und damit zu einem Effizienzgewinn. Der Gedanke dahinter ist Folgender: Weil PMM mit der gewählten Anzahl an *m* Datensätzen relativ unverzerrte Parameterschätzwerte liefert, führt deren Erhöhung für jede Replikation zu genaueren Schätzwerten. Das resultiert in einer geringeren Varianz der Schätzwerte zwischen den Replikationen, was auf Konfigurationsebene dazu führt, dass weniger Varianz und damit kleinere Werte für den MSE vorliegen. Das wieder bedeutet dann, dass effizientere Schätzungen gegeben sind.

Auch wenn keine Simulationskonfiguration vorhanden ist, die von einer problematischen Schätzung der Parameter zeugt, können den Meta-Modellen einige Einflussfaktoren entnommen werden, welche die Verzerrungen in den Parameterschätzungen vergrößern können. Das ist für die MI-Techniken zunächst die Samplegröße und für Direct-ML und EM sind es die Missinganteile (deren Effekt kann über die Vergrößerung der Fallzahl aber abgemildert werden). Sie binden jeweils den größten Anteil an Varianz in der abhängigen Variablen und sind somit auch am einflussstärksten. Grundsätzlich ist davon auszugehen, dass mit einem noch kleineren Sample (womit ein Nachweis für die Konsistenz vorliegt, weil die Schätzung der Parameter bei größeren Fallzahlen weniger verzerrt ist), oder mit noch höheren Anteilen an Missing Values, als den hier getesteten, auch die Verzerrungen größer ausfallen werden. Das gleiche gilt für noch stärker asymmetrische Verteilungen. Denn auch hierbei zeigt sich ein positiver Effekt: Je asymmetrischer die Verteilungen werden, desto größer wird der Bias. Im Vergleich

zu den anderen Variablen wird damit meist weniger, nicht aber mehr an Varianz gebunden, weswegen die Verteilungen in der Einflussstärke den anderen Faktoren nachzuordnen sind.

Sowohl der Einfluss der Missinganteile als auch der Einfluss der Verteilungen sind bereits in den deskriptiven Analysen des relativen Bias aufgetreten. Der Einfluss der Samplegröße dagegen nicht. Dies ist jedoch nicht widersprüchlich, da es sich um verschiedene Maße handelt. Beim absoluten Bias handelt es sich um ein Maß auf Replikationsebene, der relative Bias ist dagegen ein Maß auf Konfigurationseben. Demzufolge gibt der relative Bias nur eine Konfiguration im Durchschnitt wieder. Weil bei der Berechnung des relativen Bias zudem keinen Vorzeichen beachtet werden, könnte es sein, dass sich negative und positive Bias-Werte ausgleichen. Das kann dazu führen, dass der Einfluss der Samplegröße verdeckt wird. Es ist also durchaus möglich, dass für den relativen Bias über die Konfigurationen hinweg kein Effekt der Samplegröße vorliegt, dieser aber auftritt, wenn sich einzelne Bias-Werte aufgrund ihrer Vorzeichen nicht mehr ausgleichen können, wie es beim absoluten Bias der Fall ist.

Wird neben den Einflussfaktoren das korrigierte R^2 als Robustheitsmaß gegenüber den modellierten Bedingungen herangezogen, zeigt sich, dass die MI-Techniken etwas weniger stark durch diese beeinflusst sind als Direct-ML und EM. Weiterhin können für die MI-Techniken auch keine signifikanten Differenzen in den Einflussgrößen der Meta-Modell-Analysen beobachtet werden. Für die empirische Praxis bedeutet dies, dass für die Performanz der MI-Techniken einerseits die vorliegenden Dateneigenschaften etwas weniger ausschlagegebend sind als für die ML-Schätzungen und andererseits, dass unter denselben Bedingungen in etwa die gleichen Ergebnisse mit den verschiedenen MI-Techniken erwartet werden können.

Alles in allem werden durch die MDTs die Verzerrungen der MLR-Schätzungen nur minimal größer und das Niveau des Bias wird durch die identifizierten Einflussfaktoren kaum beeinflusst. Es existiert keine Konfiguration und Konstellation an Eigenschaften, die Anlass dazu gibt, eine MDT der anderen vorzuziehen; solange das Interesse auf den Parameterschätzwerten liegt und die obigen Ausführungen zu EM, FCS und PMM beachtet werden. Grundsätzlich ist aufgrund der Ergebnisse der Meta-Analyse davon auszugehen, dass größere Verzerrungen in den Schätzungen für alle MDTs vorliegen werden, wenn extremere Simulationskonfigurationen gewählt werden. Allerdings zeigen die Ergebnisse auch, dass zumindest bis zu 250 Fällen, bei stark asymmetrischen Verteilungen und 35 % Missing Values mit zufriedenstellenden Parameterschätzungen zu rechnen ist. Geht es um die Schätzung der Parameter, dann sind alle MDTs ähnlich und alle können empfohlen werden. Tabelle 16 fasst diese Ausführungen zusammen.

Tabelle 16: Zusammenfassung der Performanz bzgl. der Parameterschätzung

	Direct-ML	**EM**	**EMB, MNV**	**FCS**	**H0**	**PMM**
Bias in der Paramterschätzung						
	vernach-lässigbar	vernach-lässigbar	vernach-lässigbar	vernach-lässigbar	vernach-lässigbar	vernach-lässigbar
Einflussfaktoren auf den Bias in der Parameterschätzung						
Kleine Fallzahl	+	+	++	++	++	++
Zunehmende Asym.	+	+	+	+	+	+
Steigender Missinganteil (MA)	++	++	+	+	+	+
Effizienz (in Relation zu Direct-ML)	Konstante	deutlich gr. Varianz ab 20 % MA	gleich	gleich	tendenziell effizienter	schlechter: Faktorladung; besser: Kovarianzen
Empfehlung	unter allen Bedingungen zur Paramterschätzung geeignet					
Einschränkung	korrektes Modell	Imputationsprozess wegen gr. Varianz ggf. wiederholen	keine	ggf. überschätzte Kovarianzen	korrektes Modell	ggf. für (quasi-)metrische Indikatoren lineare Regressionen verwenden oder *m* erhöhen

Anmerkungen: x: kein Effekt; +/- positiver/negativer Effekt; ++/-- stark positiver/negativer Effekt.

8.4.1 Einordnung der Ergebnisse

Zunächst werden die Ergebnisse dieser Arbeit zu FCS und PMM mit anderen Studien verglichen. Dabei zeigt sich, dass viele der Ergebnisse mit der bisherigen Forschung übereinstimmen:

1. Die Performanz der beiden ist relativ unabhängig von den zugrundeliegenden Populationsmodellen (Raghunathan u. a. 2001; Yu u. a. 2007; Zhang u. a. 2017).
2. Die Performanz von FCS und PMM ist durch die Missinganteile beeinflusst: Je höher diese ausfallen, desto verzerrter werden die Ergebnisse (Jia/Wu 2019; Lang/Wu 2017; Lee/Carlin 2010; van Buuren u. a. 2006; White/Carlin 2010).
3. Sowohl kleine Fallzahlen als auch asymmetrische Verteilungen beeinflussen die Schätzungen der Parameter.
4. Unter den meisten der getesteten Bedingungen gehen mit diesen beiden MDTs aber unverzerrte und effiziente Schätzungen einher.

Während demnach die Performanz im Hinblick auf die Schätzung der Parameter und deren Effizienz zum größten Teil mit der bisherigen Forschung übereinstimmt, können Unterschiede ausgemacht werden, wenn die Konfiguration der IMs von FCS mit anderen Studien verglichen

wird. Während sich nämlich bei Pritikin u. a. (2018) zeigt, dass FCS Probleme hat, wenn im Imputationsprozess verschiedene Modelle herangezogen werden, kann dies für diese Arbeit nicht gelten. Zum einen ist FCS für jede Replikation einer jeden Konfiguration konvergiert und zum anderen weisen auch die Parameterwerte und deren Effizienz keine Systematik auf, als dass von einem inkompatiblen IM zu sprechen wäre, also dass die verschiedenen Verteilungen, aufgrund der unterschiedlichen Schätzverfahren für die Missings im Imputationsprozess, nicht zu einer gemeinsamen Verteilungen konvergieren. Damit zeigt sich nicht, dass die logistischen Regressionsmodelle problembehaftet sind, wenn asymmetrische Verteilungen und/oder hohe Missinganteile vorliegen, wie es bei anderen Studien zum Teil der Fall ist (McNeish 2017; Wu u. a. 2015). Zudem kann nicht beobachtet werden, dass sich PMM besser eignet als die anderen getesteten MDTs (Kleinke 2017; 2018). Sich wegen der Ergebnisse zu den Parametern für PMM und gegen andere MDTs zu entscheiden ist damit nicht unbedingt gerechtfertigt.

Die Ausführungen von eins bis vier gelten auch für die anderen MDTs. Für diese zeigt sich,

- dass der Bias mit zunehmenden Missinganteilen, asymmetrischen Verteilungen und abnehmender Samplegröße ansteigt, dabei aber in den meisten Fällen noch mit zufriedenstellenden Ergebnissen zu rechnen ist und Direct-ML, EM, EMB und MNV in den meisten Fällen ähnliche Ergebnisse generieren (Enders 2001c; Gold/Bentler 2000; Graham u. a. 1996; Honaker/King 2010; King u. a. 2001),
- dass Direct-ML effizienter ist als EM und mit MNV in etwa die gleiche Effizienz erzielt werden kann (Enders 2001b; Enders/Bandalos 2001; Li 2010; Wang 2007),
- dass MNV oder EMB bei kleinen Fallzahlen etwas schlechter abschneiden als Direct-ML oder EM (erstere weisen bei kleinen Fallzahlen mehr Bias auf, die Unterschiede sind aber gering) (Kropko u. a. 2013; Newman 2003; Olinsky u. a. 2003),
- oder dass sich Direct-ML, EMB und MNV auch für nicht metrische Variablen eignen (Asparouhov/Muthén 2010c; Finch 2010; Leite/Beretvas 2010; Lin 2010).

Obwohl sich die Ergebnisse der Arbeit kaum von anderen Studien unterscheiden, liegt zum ersten Mal ein Befund vor, wonach zwischen den fünf getesteten MI-Varianten kaum Unterschiede vorliegen. Folglich liegen keine Gründe vor, eine der MI-Varianten vorzuziehen: Es zeigt sich, dass Missing Values mit den weniger oft untersuchten Varianten EMB, FCS, H0 und PMM ebenso gut handhabbar sind, wie mit Direct-ML, EM oder MNV. Alle Varianten liefern unter denselben Bedingungen ähnliche und zudem zufriedenstellende Ergebnisse.

9 Ergebnisse Parameterebene: Standardfehlerbias

Wie die Verzerrung der Parameter, wird auch die Verzerrung der Standardfehler mit zwei Maßen untersucht. Das ist zum einen der relative Standardfehlerbias (SE-Bias) und zum anderen der absolute SE-Bias. Während es aber für die Parameterschätzungen Populationswerte als Referenzgrößen gibt, muss für den Standardfehler eine empirische Größe als ‚Populationswert' dienen. Als Referenz wird die Standardabweichung der Parameterschätzung vorgeschlagen (vgl. Bandalos/Leite 2013: 644; siehe Kapitel 6.3.4). Da der Populationswert in diesem Fall bereits durch die MLR-Schätzung verunreinigt ist, kann kein bereinigter SE-Bias berechnet werden. Der SE-Bias (ob relativ oder absolut) gibt demnach die Verzerrung des Standardfehlers als eine Kombination aus der gewählten MDT und der MLR-Schätzung wieder.

9.1 Deskriptive Analyse: Relativer Standardfehlerbias

Die Analysen zum SE-Bias orientieren sich an denjenigen zum Parameterbias. Das bedeutet zunächst, dass nur das dritte Modell betrachtet wird und die Analysen für die MDTs nur zu den oben ausgewählten Faktorladungen, Strukturpfaden und Kovarianzen erfolgen.[110] Die Ergebnisse können auf die jeweils anderen Faktorladungen, Kovarianzen und Strukturpfade des dritten sowie des ersten und zweiten Modells übertragen werden. Wiederum steht bei den Analysen das Abschneiden der MDTs zueinander im Fokus, wie die MDTs im Hinblick auf die verschiedenen Modellteile zu bewerten sind und welche Einflussgrößen[111] vorliegen. Bevor die Ergebnisse der einzelnen MDTs vorgestellt werden, wird zunächst der relative SE-Bias des Referenzmodells diskutiert (Tabelle 17).

Tabelle 17: Relativer Standardfehlerbias des Referenzmodells

Modellkonfiguration			Faktorladungen				Strukturpfade				Kovarianzen					
			ind2 .8	ind3 .7	ind4 .6	ind5 .5	x1>f1 .1	x2>f1 .3	x3>f1 .3	x4>f1 .5	x1-x2 .2	x1-x3 .2	x2-x3 .2	x1-x4 .4	x2-x4 .4	x3-x4 .2
750	skew1	mar1	2.7	-1.7	0.7	1.3	1.0	1.1	-5.1	6.5	4.6	5.3	2.0	1.1	4.4	1.4
	skew2		0.7	0.4	2.4	0.2	-0.9	-0.6	-3.5	4.4	1.7	2.6	0.6	-0.1	-2.2	5.1
	skew3		0.9	4.5	0.2	-1.5	1.1	-5.0	0.0	2.8	4.8	-3.2	-3.0	-0.5	0.9	1.0
250	skew1		2.0	-0.5	3.7	4.1	-3.8	0.1	-2.6	0.6	1.9	-0.3	-4.1	-4.5	0.8	5.3
	skew2		-2.5	-3.4	2.1	2.2	-4.3	-4.0	-1.6	2.4	-2.7	-2.2	-3.6	-6.0	-0.4	-2.8
	skew3		-1.0	0.6	0.2	2.7	-1.1	-3.1	-2.8	1.7	-1.4	-2.0	-1.3	-5.9	1.1	-2.8

[110] Analysen zu Modell 1 und Modell 2 finden sich im Anhang O2.6.

[111] Zur Identifikation der Einflussgrößen werden einzelne Konfigurationen konstant gehalten. Da sich die SE-Bias-Analysen an denen zum Parameterbias orientieren, sei auf die Ausführungen in Kapitel 8.1 verwiesen.

Alle relativen SE-Bias-Werte des Referenzmodells sind über die Konfigurationen hinweg relativ konstant. Eine klare Tendenz lässt sich nicht ablesen und keiner der Werte für die Faktorladungen, die Strukturpfade oder die Kovarianzen erreicht die Grenze von 10 % respektive 15 %. Damit liegen für das Referenzmodell unverzerrte Standardfehler vor und die inferenzstatistischen Schlüsse sind in diesem Fall unproblematisch. Berücksichtigt werden sollte, dass das Referenzmodell die festgelegten Populationsparameter der Parameter nur in Teilen zufriedenstellend schätzt (bspw. die Faktorladungen) und in einigen Fällen größere Verzerrungen derselben vorliegen (bspw. für die Kovarianzen). Da das Referenzmodell über alle Konfigurationen und Fit-Indices hinweg sehr gute Anpassungswerte liefert und zudem noch die Standardfehler unverzerrt sind, bedeutet dies für den Anwendungsfall, dass einerseits Modelle mit verzerrten Parameterschätzwerten akzeptiert werden und andererseits, dass diese verzerrten Parameterschätzwerte auch entsprechend inhaltlich eingeordnet und interpretiert werden können. Durch die Akzeptanz des Modells und durch die nicht beeinträchtigten inferenzstatistischen Schlüsse, könnte es demnach aufgrund der verzerrten Parameterschätzungen zu falschen Einordnungen der Effekte kommen, sollte die absolute Höhe der geschätzten Koeffizienten von Interesse sein. Wenn allerdings die Modellanpassung und die Prüfung von Zusammenhangshypothesen im Vordergrund stehen, so sind mit dem MLR-Schätzer korrekte Ergebnisse zu erwarten, obwohl etwas verzerrte Parameterschätzwerte vorliegen.

Auch für die MDTs sollten sich im Folgenden unverzerrte Standardfehler ergeben, sofern es diese schaffen die fehlenden Werte zufriedenstellend zu handhaben, weil mit der MLR-Schätzung unverzerrte Standardfehler geschätzt werden. Liegen hingegen verzerrte Standardfehler vor, so ist dies nicht auf den MLR-Schätzer, sondern auf die MDTs zurückzuführen. Dass das Maß des SE-Bias für die MDTs durch den MLR-Schätzer verunreinigt ist, stellt in diesem Fall also kein Problem dar.

9.1.1 Relativer SE-Bias der MDTs

Der nachfolgenden Abbildung 15 lassen sich die Ergebnisse der einzelnen MDTs für die ausgewählten Konfigurationen entnehmen.[112] Hierbei zeigt sich dreierlei: Zum einen liegen relativ homogene Ergebnisse für die verschiedenen Modellteile vor, unabhängig davon welche MDT benutzt wird. Für die Standardfehlerschätzung ist es damit unerheblich, ob es sich um Faktorladungen, Strukturpfade oder Kovarianzen handelt. Demzufolge gelten die nachfolgenden Ausführungen für alle Modellteile. Zum anderen fällt auf, dass EM gegenüber den anderen MDTs

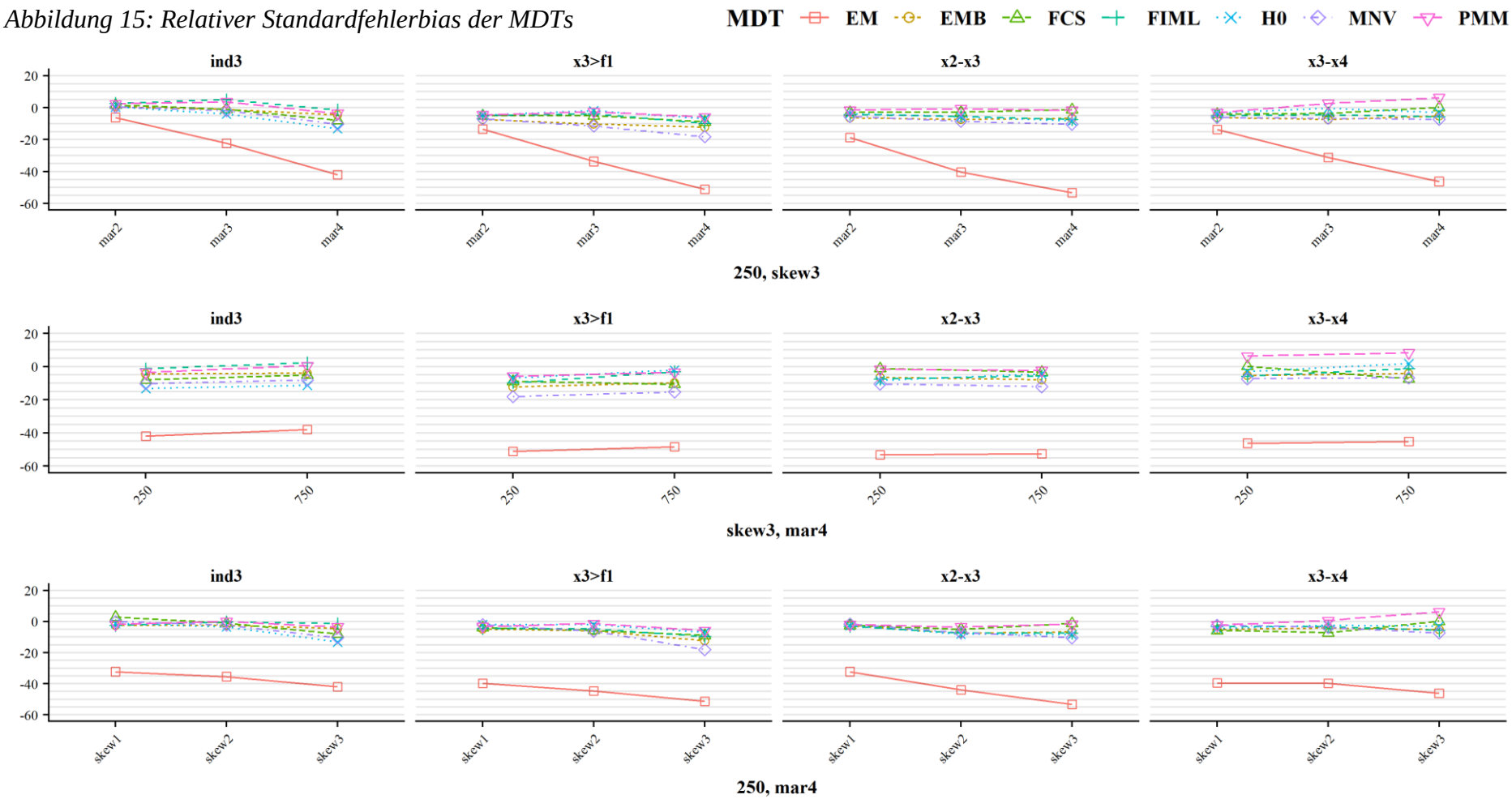
[112] Für die vollständigen Ergebnisse siehe Anhang A6.1.

(erwartungsgemäß) schlechter abschneidet: Sofern der Missinganteil nicht mehr als 5 % beträgt, sind zwar auch mit EM relativ unverzerrte Standardfehler zu erwarten (± 15 %), beträgt der Anteil der fehlenden Werte aber 20 % oder mehr, dann werden die Standardfehler in allen Fällen erheblich unterschätzt (erste Reihe[113]). Der relative SE-Bias bei 35 % an Missing Values beträgt ca. 50 %; damit wird der Standardfehler bei solch hohen Anteilen nahezu halbiert. Zusätzlich zur Zunahme des SE-Bias mit ansteigendem Missinganteil, zeigt die Abbildung, dass kein Effekt der Samplegröße[114] vorliegt und die SE-Bias-Werte relativ konstant bleiben (zweite Reihe). Dagegen nimmt der SE-Bias etwas zu, wenn die Verteilungen asymmetrischer werden. Im Vergleich zu den Missinganteilen sind die Zunahmen aber weniger ausgeprägt (dritte Reihe). Weil die Standardfehlerverzerrung in allen Fällen negativ ist, erhöht sich mit dem Einsatz von EM die Wahrscheinlichkeit den Alpha-Fehler zu begehen. Letztendlich muss durch die Anwendung von EM, sofern mehr als 5 %[115] an fehlenden Werten vorliegen, immer die Möglichkeit beachtet werden, dass bedeutsame Ergebnisse aufgrund der imputierten Werte zustande kommen.

Als drittes lassen sich der Abbildung für die sechs anderen getesteten MDTs relativ einheitliche Ergebnisse entnehmen. Größere Unterschiede zwischen ihnen lassen sich kaum erkennen und in vielen Fällen liegen auch unverzerrte Standardfehler vor.[116] Anders als beim Parameterbias wird der Grenzwert von ± 10 % für den SE-Bias mit diesen MDTs allerdings überschritten, wohingegen nur selten eine Überschreitung der etwas weicheren Grenze von ± 15 % vorliegt (mit MNV für den Strukturpfad). Neben der zufriedenstellenden Schätzung des Standardfehlers ist diese auch äußerst robust gegenüber den modellierten Bedingungen: Es lassen sich nur leichte Tendenzen erkennen, die zeigen, dass mit zunehmendem Missinganteil oder mit zunehmender Asymmetrie der SE-Bias etwas größer wird (erste und dritte Reihe). Ein Einfluss der Samplegröße lässt sich nicht beobachten (zweite Reihe).

[113] Verweise auf Reihen in diesem Kapitel beziehen sich auf Abbildung 15.
[114] Hierzu ist anzumerken, dass sowohl die Standardabweichung, die als Populationswert dient, als auch der geschätzte Standardfehler sehr wohl von der Samplegröße beeinflusst sind: Bei einer größeren Fallzahl sind beide eindeutig kleiner als bei kleinen Fallzahlen.
[115] Zu beachten ist, dass bereits bei 5 % an Missing Values der SE-Bias zum Teil 15 % beträgt und damit die obere Grenze nur knapp unterschreitet.
[116] Eine Ausnahme liegt für die Kovarianz zwischen x1 und der ordinalen Variablen (x4) für FCS vor (Abbildung A9). Hierauf wird in der Ergebnisdiskussion in Kapitel 9.3 eingegangen.

Abbildung 15: Relativer Standardfehlerbias der MDTs



Anmerkungen: erste Zeile: Konstanthaltung der Fallzahl auf 250 und der Verteilung auf ‚stark asymmetrisch' (skew3); Variation der Missinganteile. Zweite Zeile: Konstanthaltung der Verteilung auf ‚stark asymmetrisch' (skew3) und des Missinganteils auf 35 % (mar4); Variation der Fallzahl. Dritte Zeile: Konstanthaltung der Fallzahl auf 250 und des Missinganteils auf 35 % (mar4); Variation der Verteilungen. Die Fallzahl für EMB in der Konfiguration ‚250, skew3, mar4' beträgt 499.

Zum Schluss soll noch auf das Abschneiden von Direct-ML im Vergleich zu den MI-Techniken eingegangen werden. Direct-ML überschreitet in keinem Fall den Grenzwert von ± 10 %, die MI-Techniken dagegen schon. Zusätzlich sollten mit den MI-Techniken, aufgrund der Berücksichtigung der between- und within-Varianz, in aller Regel auch eher Über-, als Unterschätzungen des Standardfehlers beobachtet werden können (siehe dazu Kapitel 3). Sowohl die negativen Verzerrungen als auch die Überschreitungen des 10 %igen-Grenzwertes könnten auf die Anzahl der *m* Datensätze zurückgeführt werden. Denn bereits Graham u. a. (2007) zeigen, dass bei größeren Missinganteilen (bzw. bei einer erhöhten FMI) ein zu geringes *m* an Datensätzen die Standardfehler beeinflussen könnte, die Parameterschätzwerte davon aber nicht betroffen sind. In diesem Fall könnte es sein, dass deren Anzahl für die Konfigurationen mit erhöhten Missinganteilen nicht ausreichend war, was zu einer Unterschätzung der between-Varianz zwischen den *m* Datensätzen führte. Dies hat dann die, im Vergleich zu Direct-ML, stärker ausgeprägte Unterschätzung der Standardfehler zur Folge. Weil aber eben nur in Einzelfällen Überschreitungen der Grenzwerte mit den MI-Techniken vorliegen, darf in aller Regel davon ausgegangen werden, dass die Handhabung der fehlenden Werte mit diesen in unverzerrten Standardfehlern resultiert.

9.2 Modellbasierte Analyse: Absoluter Standardfehlerbias

Die Meta-Modelle orientieren sich an den in Kapitel 8.2 dargelegten Kriterien. Bei der abhängigen Variablen handelt es sich um den durchschnittlichen absoluten SE-Bias der Faktorladungen, der Strukturpfade oder der Kovarianzen. Erneut werden die Regressionsmodelle für jede MDT einzeln gerechnet und die Gleichheit der Effekte wird mittels t-Test geprüft (Formel 8.1 in Kapitel 8.2). Wie beim Parameterbias liegt auch in diesen Interaktionsmodellen erhöhte Multikollinearität vor. Diese wird, weil sie künstlich entstanden ist, ignoriert (siehe Kapitel 7.2). Da auch für diese Modelle Verletzungen der Homoskedastizitäts- und der Normalverteilungsannahme der Residuen vorliegen, werden robuste Standardfehler berechnet.[117]

9.2.1 Identifikation der Einflussgrößen: Änderungen im R^2

Der Fokus liegt zunächst auf der Identifikation der interpretationswürdigen Effekte. Hierzu werden die Änderungen im R^2 herangezogen (Tabelle 18). Es werden wieder nur Interaktionen der ersten Ordnung gelistet, da in allen Modellen kein Zugewinn an Erklärungsleistung mittels

[117] Der Auschluss von Ausreißern und einflussreichen Beobachtungen ergibt keine substantiellen Unterschiede in den b-Koeffizienten (siehe Kapitel 8.2, Fußnote 106).

der Interaktion der zweiten Ordnung erzielt wird. Alle Modelle berücksichtigen zudem die gleiche Anzahl an Variablen, damit ein Koeffizientenvergleich möglich wird.

Grundsätzlich gilt, dass über alle Modellteile und Modelle hinweg in etwa derselbe Anteil an Varianz gebunden werden kann. Mit Ausnahme von EM können die unabhängigen Variablen zwischen ca. 50 % (PMM) und ca. 75 % (FCS) an Varianz binden (korrigiertes R^2). Im Vergleich dieser MDTs zeigt sich PMM am wenigsten durch die Simulationsbedingungen beeinflusst, wohingegen die Performanz der anderen MDTs etwas stärker von diesen betroffen ist. Zwischen ihnen gibt es allerdings nur kleinere Unterschiede, denn in allen Modellen wird in aller Regel ein ähnlich hoher Varianzanteil gebunden. EM hebt sich dagegen von den anderen MDTs ab. In diesen Modellen kann über 90 % an Varianz erklärt werden, was dazu führt, dass der absolute Standardfehlerbias mit den unabhängigen Variablen nahezu komplett vorhergesagt wird. Wird in den EM-Modellen die Einflussstärke der einzelnen Variablen betrachtet, zeigt sich, dass annähernd 60 % der Varianz durch die Missingvariable gebunden wird. Eine solche Erklärungskraft weist diese Variable in keinem Modell für die anderen MDTs auf. Gleichzeitig zeigt sich ein Effekt der Missinganteile für EM auch bei den Analysen zum relativen SE-Bias, für die anderen MDTs gibt es einen solchen dagegen nicht. Auch zeigt sich in der vorangegangenen Forschung, dass die Verzerrungen der Standardfehler für EM vorwiegend von den Missinganteilen abhängig sind (unabhängig davon ob EM als Indirect-ML oder als Einfachimputation verwendet wird: Li 2010; Li/Lomax 2017; Newman 2003). Das hohe (korrigierte) R^2 in den EM-Modellen erscheint demnach gerechtfertigt.[118]

Im Gegensatz zur modellbasierten Analyse des Parameterbias liegen für diese Analysen heterogenere Ergebnisse zwischen den Modellteilen vor. Auch lassen sich Unterschiede zwischen den Populationsmodellen ausmachen, die in Tabelle 18 farblich hinterlegt sind.[119] Hierbei fällt auf, dass nur für H0 keine Differenzen zwischen den einzelnen Modellen vorliegen. Die grau hinterlegten Felder zeigen dagegen an, dass diese Effekte in Modell 1 oder in Modell 2 den festgesetzten Erklärungsanteil von 2 % nicht erreichen, in Modell 3 dagegen schon. Das trifft einmal auf EM und in zwei Fällen auf FCS zu. Dabei gilt allerdings, dass die Unterschreitung des Grenzwertes von 2 % relativ knapp ist.

[118] In bivariaten Kontrollanalysen zeigt sich, dass von den drei unabhängigen Variablen vor allem die mar-Variable sehr stark mit der abhängigen Variablen zusammenhängt.
[119] Die Ergebnisse zu Modell 1 und Modell 2 finden sich im Anhang O2.7.

Tabelle 18: Einflussgrößen auf den Standardfehlerbias: R^2 und Änderung im R^2

abh. Variable: durchschnittlicher absoluter SE-Bias in den Faktorladungen

	EM			EMB			FCS			FIML			H0			MNV			PMM		
+ Variable	**R^2**	**ΔR^2**	**korr. R^2**	**R^2**	**ΔR^2**	**korr. R^2**	**R^2**	**ΔR^2**	**korr. R^2**	**R^2**	**ΔR^2**	**korr. R^2**	**R^2**	**ΔR^2**	**korr. R^2**	**R^2**	**ΔR^2**	**korr. R^2**	**R^2**	**ΔR^2**	**korr. R^2**
250	0.138	**0.138**	0.138	0.33	**0.330**	0.33	0.288	**0.288**	0.288	0.411	**0.410**	0.41	0.279	**0.279**	0.279	0.294	**0.294**	0.294	0.349	**0.349**	0.349
skew	0.242	**0.104**	0.242	0.405	**0.075**	0.405	0.381	**0.094**	0.381	0.462	**0.052**	0.462	0.393	**0.114**	0.393	0.394	**0.099**	0.394	0.414	**0.065**	0.414
mar	0.843	**0.600**	0.843	0.564	**0.159**	0.564	0.573	**0.191**	0.572	0.575	**0.113**	0.575	0.568	**0.175**	0.568	0.573	**0.179**	0.573	0.557	**0.143**	0.557
250:skew	0.851	0.009	0.851	0.575	0.011	0.574	0.577	0.004	0.576	0.586	0.011	0.586	0.574	0.006	0.574	0.579	0.006	0.579	0.567	0.009	0.566
250:mar	0.9	**0.049**	0.9	0.603	**0.029**	0.603	0.593	0.016	0.592	0.613	**0.027**	0.613	0.589	0.015	0.588	0.6	**0.021**	0.599	0.581	0.014	0.58
skew:mar	0.942	**0.042**	0.942	0.63	**0.026**	0.629	0.636	**0.043**	0.635	0.624	0.010	0.623	0.648	**0.059**	0.647	0.652	**0.052**	0.651	0.593	0.012	0.593

abh. Variable: durchschnittlicher absoluter SE-Bias in den Strukturpfaden

250	0.182	**0.182**	0.182	0.268	**0.268**	0.268	0.189	**0.189**	0.189	0.357	**0.357**	0.357	0.305	**0.305**	0.305	0.201	**0.201**	0.201	0.297	**0.297**	0.297
skew	0.26	**0.078**	0.26	0.363	**0.095**	0.363	0.314	**0.124**	0.313	0.412	**0.055**	0.412	0.362	**0.057**	0.362	0.361	**0.160**	0.361	0.349	**0.053**	0.349
mar	0.853	**0.593**	0.853	0.532	**0.169**	0.532	0.527	**0.213**	0.527	0.525	**0.113**	0.525	0.492	**0.131**	0.492	0.55	**0.189**	0.55	0.487	**0.137**	0.487
250:skew	0.857	0.003	0.857	0.547	0.015	0.547	0.531	0.004	0.531	0.546	**0.021**	0.546	0.502	0.010	0.502	0.566	0.016	0.566	0.501	0.014	0.5
250:mar	0.911	**0.054**	0.911	0.572	**0.025**	0.572	0.541	0.010	0.541	0.579	**0.033**	0.578	0.514	0.012	0.514	0.58	0.013	0.579	0.519	0.018	0.519
skew:mar	0.944	**0.033**	0.944	0.601	**0.028**	0.6	0.597	**0.055**	0.596	0.593	0.014	0.593	0.527	0.012	0.526	0.653	**0.073**	0.652	0.532	0.013	0.532

abh. Variable: durchschnittlicher absoluter SE-Bias in den Kovarianzen

250	0.179	**0.179**	0.179	0.3	**0.300**	0.3	0.191	**0.191**	0.191	0.341	**0.341**	0.341	0.292	**0.292**	0.292	0.237	**0.237**	0.237	0.273	**0.273**	0.272
skew	0.292	**0.113**	0.292	0.427	**0.126**	0.426	0.347	**0.156**	0.347	0.447	**0.106**	0.447	0.375	**0.083**	0.375	0.421	**0.184**	0.421	0.349	**0.076**	0.349
mar	0.915	**0.624**	0.915	0.626	**0.199**	0.625	0.682	**0.335**	0.682	0.588	**0.141**	0.588	0.569	**0.193**	0.568	0.619	**0.198**	0.619	0.555	**0.206**	0.555
250:skew	0.922	0.007	0.922	0.643	0.017	0.642	0.692	0.010	0.692	0.621	**0.033**	0.621	0.583	0.014	0.582	0.637	0.018	0.636	0.563	0.008	0.563
250:mar	0.957	**0.034**	0.957	0.669	**0.026**	0.669	0.702	0.009	0.701	0.664	**0.043**	0.664	0.614	**0.031**	0.614	0.653	0.016	0.653	0.58	0.017	0.58
skew:mar	0.986	**0.029**	0.986	0.692	**0.023**	0.692	0.756	**0.054**	0.755	0.696	**0.032**	0.696	0.638	**0.024**	0.637	0.702	**0.049**	0.701	0.604	**0.024**	0.604

Anmerkungen: Fettdruck: Erklärungskraft der Variablen ≥ 2 %. Grau: nicht substantiell in Modell 2, Erklärungskraft ≥ 1 % aber < 2 %. Blau: substantiell in Modell 1 oder Modell 2, Erklärungskraft ≥ 2 % aber < 3 %. Rot: substantiell in Modell 1/Modell 2, Erklärungskraft ≥ 4 %.

Die blau hinterlegten Felder zeigen dagegen Effekte an, die in Modell 3 nicht, dafür aber in Modell 1 oder Modell 2 bedeutsam sind und auf Direct-ML, FCS, MNV und PMM zutrifft. In diesem Fall gilt, dass der Grenzwert von 2 % knapp überschritten wird. Erheblichere Unterschiede zwischen den Modellen liegen für die rot gekennzeichneten Felder vor: Mit der jeweiligen Variablen werden in Modell 1 und Modell 2 größere Varianzanteile gebunden als 2 %. Während die grau und blau hinterlegten Felder, aufgrund dessen, dass die Unter- respektive Überschreitung des Grenzwertes von 2 % jeweils relativ knapp ist, noch in Einklang mit den Effektstärken des dritten Modells gebracht werden können, gilt das für die rot hinterlegten Felder nicht: Hierbei werden zu große Anteile an Varianz gebunden, als dass diese Effekte vernachlässigt werden können.

Neben den Unterschieden zwischen den Populationsmodellen gibt es weiterhin Unterschiede zwischen den Modellteilen. Das trifft besonders auf Direct-ML und H0 zu. So ist für alle Modellteile bei Direct-ML nur die Interaktion aus der Fallzahl und der Missingvariable erklärungskräftig, wohingegen für die Strukturpfade noch die Interaktion aus der Fallzahl und den Verteilungen aussagekräftig ist und für die Kovarianzen alle Interaktionen substantiell sind. Ähnlich verhält es sich bei H0. Hierfür kann die Interaktion der Missingvariable und der Fallzahl sowie die Interaktion zwischen der Missingvariable und den Verteilungen als erklärungskräftig ausgemacht werden. Letzteres kann auch für die Faktorladungen beobachtet werden, wohingegen für die Strukturpfade keine Interaktion relevant ist. Bei EM, EMB und FCS liegen einheitliche Ergebnisse zwischen den Modellteilen vor. Das kann auch für MNV und PMM gelten, da die Überschreitung der Grenzwerte für die Interaktionen in den Faktorladungen (MNV) und in den Kovarianzen (PMM) knapp ist.

Was allerdings für alle MDTs, alle Modelle und alle Modellteile beobachtet werden kann ist, dass erstens in den Regressionsmodellen sehr viel Varianz gebunden wird (> 50 %), von welcher mindestens 5 % auf die Interaktionen zurückgehen.[120] Das bedeutet, dass die Interaktionen im Gesamten eine erhebliche Erklärungskraft besitzen, die dazu führt, dass alle Interaktionen den gesetzten Grenzwert von 2 % *im Durchschnitt* knapp erreichen. Zudem kann zweitens eine Systematik in Bezug auf die Reihenfolge der Erklärungskraft der unabhängigen Variablen über die Modelle und Modellteile beobachtet werden: Die Fallzahl und die Missinganteile sind in der Regel die erklärungskräftigsten Variablen. Danach folgt in vielen Fällen die Verteilungsvariable. Als erklärungskräftigste Interaktionen lassen sich die Interaktionen zwischen der

[120] Beim Parameterbias beträgt die Zunahme durch die Interaktionen in allen Fällen weniger als 5 %.

Missingvariable und der Fallzahl sowie der Missingvariable und den Verteilungen identifizieren. Zuletzt folgt die Interaktion zwischen der Fallzahl und den Verteilungen.

Weil Unterschiede in den erklärungskräftigen Variablen zwischen den Populationsmodellen und den jeweiligen Modellteilen existieren, gleichzeitig aber mit der Gesamtheit der Interaktionen eine erhöhte Varianzbindung einhergeht, die es erlaubt, alle Interaktionen *im Durchschnitt* als bedeutungsvoll einzuordnen und die Variablen zudem in etwa dieselbe Wichtigkeit in den einzelnen Modellen und Modellteilen aufweisen (sie können in Bezug auf die Varianzbindung in eine ähnliche Reihenfolge gebracht werden), wird zunächst darauf verzichtet einzelne Einflussgrößen im nächsten Kapitel nicht zu interpretieren. Stattdessen werden als Ad-hoc-Maßnahme, alle Effekte vorerst als substantiell angesehen und anhand ihrer Signifikanz und der b-Koeffizienten beurteilt. Dieses Vorgehen, in Kombination mit den daraus resultierenden Betrachtungen und Interpretationen, lässt eine Einordnung zu, welche Einflussgrößen für die jeweilige MDT von Bedeutung sind und welche nicht.

9.2.2 Inhaltliche Interpretation der Einflussgrößen

In diesem Unterkapitel werden die Regressionsmodelle inhaltlich interpretiert. Anders als für die Effektstärken, liegen in Bezug auf die b-Koeffizienten und die Signifikanz der Effekte, für alle drei Populationsmodelle und alle abhängigen Variablen größtenteils ähnliche Ergebnisse vor. Weil sich die inhaltlichen Interpretationen nicht zwischen den einzelnen Modellen und Modellteilen unterscheiden, werden alle anhand des dritten Modells mit dem SE-Bias in den Faktorladungen als abhängige Größe getätigt.[121] Die daraus gezogenen Schlussfolgerungen gelten dann für alle Populationsmodelle und deren Modellteile.

Die F-Werte in Tabelle 19 zeigen, dass alle Modelle signifikant sind ($p < .001$).[122] Das korrigierte R^2 zeigt zudem, dass EM gegenüber den simulierten Bedingungen am wenigsten robust ist (siehe vorangegangenes Kapitel). In Bezug auf die b-Koeffizienten lassen sich zudem Unterschiede zwischen den MDTs beobachten: Die Schätzwerte der b-Koeffizienten der Fallzahl sind bei EM weniger ausgeprägt, als für die anderen MDTs. In allen Fällen liegen aber positive Effekte vor, was bedeutet, dass für alle MDTs der SE-Bias bei kleinen Fallzahlen größer ist, als bei einer größeren Fallzahl (immer unter der Bedingung, dass alle anderen Variablen konstant gehalten werden). Weiterhin lassen sich positive b-Koeffizienten für die Missinganteile ausmachen, wonach mit ansteigenden Missinganteilen auch ein größerer SE-Bias zu erwarten ist.

[121] Die Modelle für die Strukturpfade und Kovarianzen sowie die Tests auf Unterschiede in den b-Koeffizienten lassen sich dem Anhang A6.2 entnehmen.
[122] Anmerkung: F-Statistik in allen Modellen: *** ($p < 0.001$).

Auch für diese Effekte gibt es zwischen EM und den anderen MDTs Unterschiede, denn die Koeffizienten sind bei EM ausgeprägter. Im Gegensatz zu den positiven Fallzahl- und Missingeffekten lässt sich für alle MDTs für die Verteilungsvariable eher kein Effekt beobachten. In vielen Fällen ist der Einfluss der Variablenverteilung nicht signifikant und falls doch, dann ist der Effekt, unabhängig von der Effektrichtung, wenig ausgeprägt.

Tabelle 19: Ergebnisse der modellbasierten Analyse zum Standardfehlerbias

	abh. Variable: durchschnittlicher absoluter SE-Bias in den Faktorladungen						
	b (95 % KI)						
	EM	EMB	FCS	FIML	H0	MNV	PMM
Constant	0.22***	0.14***	0.13***	0.15***	0.14***	0.14***	0.14***
	(0.21, 0.23)	(0.14, 0.15)	(0.12, 0.14)	(0.14, 0.15)	(0.13, 0.14)	(0.13, 0.14)	(0.13, 0.15)
250	0.06***	0.18***	0.22***	0.19***	0.20***	0.20***	0.20***
	(0.05, 0.08)	(0.17, 0.20)	(0.21, 0.23)	(0.18, 0.21)	(0.19, 0.21)	(0.19, 0.21)	(0.19, 0.21)
skew2	-0.02**	-0.01*	0.001	-0.003	-0.01	-0.01	0.004
	(-0.04, -0.01)	(-0.02, -0.001)	(-0.01, 0.01)	(-0.01, 0.01)	(-0.02, 0.001)	(-0.02, 0.003)	(-0.01, 0.02)
skew3	-0.04***	-0.01*	0.01*	0.0001	-0.001	0.002	0.02**
	(-0.05, -0.02)	(-0.03, -0.002)	(0.0005, 0.03)	(-0.01, 0.01)	(-0.01, 0.01)	(-0.01, 0.01)	(0.004, 0.03)
mar3	0.41***	0.02***	0.04***	0.01*	0.03***	0.04***	0.03***
	(0.40, 0.43)	(0.01, 0.03)	(0.03, 0.05)	(0.0003, 0.02)	(0.02, 0.04)	(0.03, 0.05)	(0.02, 0.04)
mar4	0.81***	0.05***	0.09***	0.04***	0.06***	0.06***	0.09***
	(0.79, 0.83)	(0.03, 0.06)	(0.08, 0.10)	(0.03, 0.05)	(0.05, 0.07)	(0.04, 0.07)	(0.08, 0.11)
250:skew2	0.23***	0.07***	0.03**	0.05***	0.05***	0.05***	0.05***
	(0.21, 0.25)	(0.05, 0.08)	(0.01, 0.05)	(0.04, 0.07)	(0.04, 0.07)	(0.04, 0.07)	(0.03, 0.07)
250:skew3	0.46***	0.16***	0.10***	0.15***	0.12***	0.12***	0.14***
	(0.43, 0.48)	(0.14, 0.18)	(0.08, 0.12)	(0.13, 0.17)	(0.10, 0.14)	(0.10, 0.14)	(0.12, 0.16)
250:mar3	0.30***	0.06***	0.08***	0.10***	0.05***	0.06***	0.09***
	(0.28, 0.32)	(0.05, 0.08)	(0.06, 0.09)	(0.09, 0.12)	(0.04, 0.07)	(0.05, 0.08)	(0.08, 0.11)
250:mar4	1.04***	0.25***	0.21***	0.24***	0.18***	0.22***	0.17***
	(1.02, 1.07)	(0.23, 0.27)	(0.18, 0.23)	(0.22, 0.25)	(0.16, 0.20)	(0.20, 0.24)	(0.15, 0.19)
skew2:mar3	0.20***	0.04***	0.04***	0.04***	0.03***	0.03***	0.05***
	(0.18, 0.22)	(0.03, 0.06)	(0.02, 0.05)	(0.02, 0.05)	(0.02, 0.05)	(0.01, 0.04)	(0.04, 0.07)
skew3:mar3	0.55***	0.13***	0.14***	0.08***	0.15***	0.12***	0.10***
	(0.52, 0.57)	(0.11, 0.15)	(0.12, 0.16)	(0.06, 0.10)	(0.13, 0.16)	(0.11, 0.14)	(0.08, 0.12)
skew2:mar4	0.44***	0.10***	0.09***	0.07***	0.08***	0.09***	0.09***
	(0.41, 0.46)	(0.08, 0.12)	(0.07, 0.11)	(0.05, 0.09)	(0.06, 0.10)	(0.07, 0.11)	(0.07, 0.11)
skew3:mar4	1.21***	0.30***	0.39***	0.18***	0.42***	0.41***	0.20***
	(1.18, 1.24)	(0.27, 0.33)	(0.36, 0.42)	(0.16, 0.20)	(0.40, 0.45)	(0.38, 0.44)	(0.17, 0.22)
Fallzahl	9000	8999	9000	9000	9000	9000	9000
F-Statistik	12847.496	1139.852	1294.917	1073.932	1244.017	1261.424	1095.006
R^2	0.942	0.63	0.636	0.624	0.648	0.652	0.593
korr. R^2	0.942	0.629	0.635	0.623	0.647	0.651	0.593
SEE	0.239	0.193	0.201	0.181	0.188	0.193	0.189
Note:							*p<0.05; **p<0.01; ***p<0.001

Von Bedeutung ist nun, dass das Modell Interaktionsterme enthält. Damit ändert sich grundsätzlich die Interpretation von nicht signifikanten Haupteffekten: Da sowohl der Interaktionsterm der Verteilung mit der Fallzahl, als auch der Interaktionsterm mit den Missinganteilen

signifikant ist, bedeutet der nicht signifikante Verteilungseffekt, dass kein Einfluss vorliegt, wenn ein großes Sample oder ein Missinganteil von 5 % gegeben ist. Im Gegensatz dazu, kann ein Effekt beobachtet werden, wenn ein kleines Sample gegeben ist oder, wenn die Missinganteile erhöht sind. In beiden Fällen nimmt der SE-Bias zu, wenn anstatt symmetrischen, (stark) asymmetrische Verteilungen vorliegen.

Da es sich bei den unabhängigen Variablen um dichotome Variablen handelt, können alle Interaktionen auch jeweils anders verstanden werden. Wird der Interaktionseffekt der Verteilungen mit der Fallzahl im Hinblick auf den Effekt der Fallzahl interpretiert, zeigt sich, dass dieser in beiden Fällen ausgeprägter ist. Demnach verstärkt sich der Effekt der Fallzahl, wenn (stark) asymmetrische Verteilungen gegeben sind. Zudem kann für die Fallzahl auch eine signifikante Interaktion mit den Missinganteilen beobachtet werden. Hierbei zeigt sich, dass der Effekt der Samplegröße, wenn ein 20 %iger Missinganteil vorliegt, größer ist als bei einem 5 %igen Missinganteil. Gleiches gilt auch für den Interaktionseffekt bei 35 % Missinganteil. Es kann also ein durchschnittlicher Fallzahleffekt für alle MDTs beobachtet werden, womit bei sinkender Fallzahl auch der SE-Bias größer wird, verstärkt wird dieser aber, wenn 20 % oder wenn 35 % an Missing Values gegeben sind.

Als letztes werden die Interaktionen zwischen der Missingvariable und der Fallzahl, sowie zwischen der Missingvariable und den Verteilungen im Hinblick auf die Haupteffekte der Missinganteile besprochen. Für die Interaktion mit der Fallzahl bedeutet das, dass der Einfluss der Missinganteile bei kleinen Fallzahlen deutlicher ausfällt als bei großen Fallzahlen. Demzufolge fallen bei kleinen Fallzahlen die Verzerrungen im Standardfehler bei erhöhten Missinganteilen auch größer aus. Ähnliches gilt auch für die Verteilungen: Liegen (stark) asymmetrische Verteilungen vor, dann ist von einem größeren Effekt der Missingvariable auszugehen.

Werden nun die b-Koeffizienten für die einzelnen MDTs miteinander verglichen (Tabelle 20[123]), zeigt sich, dass die Verzerrungen im Standardfehler, sollte EM eingesetzt werden, in nahezu allen Fällen ausgeprägter sind, als wenn die anderen MDTs zum Einsatz kommen. Für Direct-ML, EMB, FCS, H0, MNV und PMM liegen einheitliche Ergebnisse vor, auch wenn in der Tabelle signifikante Unterschiede zwischen den b-Koeffizienten vorliegen. Das ist aber wohl der großen Fallzahl geschuldet, denn die b-Koeffizienten sind in ihrer Ausprägung zueinander relativ ähnlich und deuten auch keine andere Einordnung an. An dieser Stelle deshalb von Unterschieden in den Einflussgrößen zu sprechen ist nicht angebracht. Letztlich ist die

[123] Tabelle 20 zeigt ausgewählte Vergleiche, alle anderen lassen sich dem Anhang A6.2 entnehmen.

Performanz der Schätzung des Standardfehlers für alle diese Techniken in gleichem Maße durch die unabhängigen Variablen beeinflusst. Demzufolge können unter denselben Bedingungen mittels dieser MDTs auch ähnliche Schätzungen für die Standardfehler erwartet werden.

Tabelle 20: Test auf Unterschiede in den b-Koeffizienten (SE-Bias)

Differenzen der b-Koeffizienten zwischen den MDTs mit 5 %iger Irrtumswahrscheinlichkeit signifikant?

Effekt	MNV vs. EMB	MNV vs. FCS	MNV vs. H0	MNV vs. PMM	MI (MNV) vs. EM	MI (MNV) vs. FIML	FIML vs. EM
250	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE
skew2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
skew3	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
mar3	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
mar4	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE
250:skew2	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE
250:skew3	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
250:mar3	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
250:mar4	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE
skew2:mar3	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
skew3:mar3	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
skew2:mar4	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
skew3:mar4	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE

Anmerkungen: FALSE: Unterschied nicht signifikant; TRUE: signifikanter Unterschied.

9.2.2.1 Schlussfolgerungen in Bezug auf die Einflussgrößen

In Rückbezug auf Tabelle 18 werden jetzt die einzelnen Einflussgrößen auf den SE-Bias diskutiert. Neben der Effektstärke wird dabei auch die inhaltliche Einordnung aus den vorangegangenen Ausführungen herangezogen sowie das Abschneiden der MDTs zueinander (Vergleich der b-Koeffizienten), um die nicht gleichartigen Ergebnisse der Effektstärken zu systematisieren. Da sich im vorangegangenen Kapitel zeigte, dass es eher keine Unterschiede zwischen Direct-ML, EMB, FCS, H0, MNV und PMM gibt, werden diese gemeinsam besprochen.

Für EM können aus Tabelle 18 relativ ähnliche Ergebnisse entnommen werden. Dabei zeigt sich, dass vor allen Dingen die Missinganteile, gefolgt von der Samplegröße die größten Varianzanteile binden. Zudem können auch die Interaktionen aus dem Missinganteil und der Fallzahl sowie dem Missinganteil und den Verteilungen als erklärungskräftig angesehen werden, die Interaktion aus der Fallzahl und den Verteilungen ist dagegen für kein Modell und keinen Modellteil substantiell, sodass diese an dieser Stelle auch nicht weiter relevant ist. Letztendlich zeigt sich für die Fallzahl anhand der b-Koeffizienten (Fallzahlkoeffizient im Vergleich zu den anderen MDTs weniger ausgeprägt) und für die Verteilungen durch die fehlende Signifikanz, dass die Einflüsse zwar relevant sind, allerdings erst, wenn diese über die Missinganteile vermittelt werden. Die Missinganteile sind demzufolge für EM vor allem für die Verzerrungen im Standardfehler verantwortlich und es ist für die Performanz wichtig, wie hoch dieser ausfällt. Ob kleine Fallzahlen oder asymmetrische Verteilungen vorliegen ist dabei nebensächlich, denn

unter sonst optimalen Bedingungen (einer großen Fallzahl und symmetrischen Verteilungen) ist aufgrund der Haupteffekte bereits von einem erhöhten SE-Bias auszugehen. Erst dann interessiert, ob zusätzlich kleine Fallzahlen oder asymmetrische Verteilungen vorliegen; beides führt zu zusätzlichen Verzerrungen.

Im Gegensatz zu EM liegen für die anderen MDTs eher keine einheitlichen Ergebnisse in Bezug auf die Einflussstärke einzelner Variablen vor. Grundsätzlich zeigen sich dort die Samplegröße und die Missinganteile, gefolgt von den Verteilungen am einflussstärksten. Ein uneinheitliches Bild kann aber für die Interaktionen beobachtet werden. Denn für einzelne MDTs, für einzelne Modellteile und zwischen den Populationsmodellen ergeben sich unterschiedliche Ergebnisse. Es stellt sich also die Frage, welche der Interaktionen relevant sind und welche nicht. Zunächst zeigt sich, dass diese MDTs bei einem genügend großen Sample relativ robust gegenüber den Missinganteilen und den Verteilungen sind, da die Haupteffekte der Missings im Vergleich zu den EM-Effekten sehr gering und die Effekte der Verteilungen nicht bedeutsam sind. Einflüsse der beiden zeigen sich erst, wenn diese über die Fallzahl vermittelt werden. Es ist für die Performanz dieser MDTs also unwichtig, wie hoch die Missinganteile sind oder ob asymmetrische Verteilungen vorliegen, sofern ein großes Sample gegeben ist.

Neben der Fallzahl ist aufgrund der Effektstärke auch der Missinganteil eine entscheidende Einflussgröße. Anhand der inhaltlichen Interpretation zeig sich aber, dass weniger die Haupteffekte eine Rolle spielen, als vielmehr die Interaktionen. Demnach wird der Effekt einerseits über die Fallzahl vermittelt, andererseits über die vorliegenden Verteilungen. Sind kleine Fallzahlen oder asymmetrische Verteilungen gegeben, dann verstärkt sich der Effekt. Für beide Interaktionseffekte liegen aber nicht immer Grenzwertüberschreitungen vor. In Anbetracht der Tatsache, dass aber der Haupteffekt der Missingvariablen im Vergleich zu EM eher zu vernachlässigen ist, dieser aber größer wird, wenn kleine Fallzahlen oder asymmetrische Verteilungen vorliegen, können beide Interaktionen, trotz einiger Unterschreitungen des Grenzwertes, als relevant eingestuft werden.

Zuletzt muss noch auf das widersprüchliche Ergebnis hinsichtlich der Interaktion aus der Fallzahl und den Verteilungen eingegangen werden. Diese überschreitet zwar nur in sehr wenigen Fällen den Grenzwert von 2 %, dafür aber deutlich. Liegt demnach ein kleines Sample vor, dann kann auch ein Einfluss der Verteilungen wahrgenommen werden. Allerdings erweisen sich die, mittels der Interaktionsterme errechenbaren b-Koeffizienten für die Verteilungen bei kleinen Fallzahlen, als weniger ausgeprägt, als diejenigen für EM. Demzufolge ist der Einfluss bei EM ausgeprägter als bei den anderen MDTs. Da einerseits der Effekt für die anderen

MDTs weniger ausgeprägt ist und sich andererseits zeigt, dass der stärker ausgeprägte Effekt für EM in keinem Fall relevant ist, kann auch für erstere davon ausgegangen werden, dass dieser nicht relevant ist. Letztendlich liegen also auch für Direct-ML, EMB, FCS, H0, MNV und PMM in etwa ähnliche Einflussgrößen vor wie für EM: das sind die Samplegröße, der Missinganteil und, über die Missinganteile vermittelt, die Verteilungen sowie die Interaktion aus der Fallzahl und den Missinganteilen.

9.2.3 Zusammenfassung der modellbasierten Analyse

Zusammenfassend lassen sich der modellbasierten Analyse folgende Ergebnisse entnehmen:

- Für EM werden in allen Modellen mehr Varianzanteile durch die unabhängigen Variablen gebunden. Die anderen MDTs sind gegenüber den modellierten Bedingungen zwar robuster, dennoch aber von diesen beeinflusst.
- Als Haupteinflussgröße für EM kann der Missinganteil genannt werden. Erhöht sich dieser, ist nicht mehr mit unverzerrten Standardfehlern zu rechnen, unabhängig von der Fallzahl oder den Verteilungen. Weitere Einflussgrößen, mit abnehmender Erklärungskraft, sind: die Fallzahl, die Verteilungen (nur über die Missinganteile vermittelt) sowie die Interaktion zwischen der Fallzahl und den Missinganteilen.
- Für die anderen MDTs sind die Fallzahl und die Missinganteile die Haupteinflussgrößen. Auch weisen die Verteilungen einen Effekt auf, der allerdings nur über die Missinganteile vermittelt wird. Zudem kann die Interaktion zwischen der Fallzahl und den Missinganteilen als Einflussgröße identifiziert werden.
- Es gilt für alle MDTs, dass mit abnehmender Fallzahl der SE-Bias größer wird, womit ein Nachweis für die Konsistenz der Standardfehlerschätzung vorliegt. Zudem nimmt der SE-Bias tendenziell zu, wenn die Missinganteile größer werden. Erst bei erhöhten Missinganteilen ergibt sich ein positiver Effekt für die Verteilungen. Zudem ist der Einfluss der Missings bei kleineren Fallzahlen ausgeprägter.
- Keine Unterschiede in den Einflussgrößen können für Direct-ML, EMB, FCS, H0, MNV und PMM festgestellt werden. Alle sind in gleichem Maße durch die Simulationskonfigurationen beeinflusst. Das wiederum lässt den Schluss zu, dass diese MDTs unter denselben Bedingungen, zu annähernd gleichen Ergebnisse führen.
- EM liefert gegenüber den anderen MDTs in der Regel stärker ausgeprägte b-Koeffizienten. Das bedeutet, dass der SE-Bias für EM durchschnittlich höher ausfällt.

9.3 Standardfehlerbias: Ergebnisdiskussion und Einordnung

Wie für die Parameterschätzungen liegen auch für die Standardfehler mit Direct-ML, EMB, H0, MNV und PMM kaum verzerrte Schätzungen vor. In den meisten Fällen wird der Grenzwert für den relativen SE-Bias nicht erreicht. Einzige Ausnahme ist FCS: In der Konfiguration mit hohem Missinganteil und stark asymmetrischen Verteilungen liegen für die Kovarianz zwischen der ordinalen Variablen (x4) und x1 (eine der metrischen unabhängigen Variablen) stark negativ verzerrte Standardfehler vor. Die Standardfehler der anderen Kovarianzen von x4 sowie deren Strukturpfad werden unverzerrt geschätzt. Wiederum handelt es sich also um ein einmaliges Ergebnis. Dadurch kann nicht auf ein Problem mit den zugrundeliegenden Modellen zur Ersetzung der fehlenden Werte geschlossen werden, zumal es sich in diesem Fall um eine andere Kovarianz handelt als beim Parameterbias. Weil allerdings sowohl bei den Parametern als auch bei den Standardfehlern stärkere Verzerrungen mit FCS vorliegen, stellt sich die Frage, ob der zugrundeliegende computergesteuerte Zufallsprozess zu diesem Ergebnis führt, also ob mit einem anderen Startwert beim Einsatz von FCS andere Ergebnisse für die Kovarianzen erzielt werden können, oder ob es generell problematisch ist, *alle* Kovarianzen mittels FCS wiederherzustellen, wenn logistische Regressionsmodelle eingesetzt werden. Für letzteres spricht, dass auch in Modell 2 nicht alle Kovarianzen unverzerrt geschätzt werden. Grundsätzlich erscheint das Ergebnis von FCS für die Kovarianzen jedoch weniger problematisch zu sein, da eben nur wenige Konfigurationen in jeweils nur einer einzigen Modellkomponente von diesen Verzerrungen betroffen sind.

Zwar liegen für alle MI-Techniken eher keine verzerrten Standardfehler vor. Allerdings ist zu beobachten, dass diese bei höheren Quoten an Missings (35 %) in aller Regel unterschätzt werden. Das deutet darauf hin, dass die Anzahl der *m* Datensätze für die vorliegende Arbeit für diese Konfigurationen nicht hoch genug war und die between-Varianz unterschätzt wurde. Das resultiert in unterschätzten Standardfehlern. Da allerdings diese Unterschätzung in den meisten Fällen nicht bedeutsam ist, die MI-Techniken hinsichtlich des relativen SE-Bias in etwa dieselben Ergebnisse erbringen wie Direct-ML und dies nur zutrifft, wenn *m* hoch genug ist (siehe Graham u. a. 2007), ist das gewählte *m* in diesem Fall ausreichend. Für die empirische Praxis empfiehlt es sich aber *m* möglichst hoch zu wählen, zumal in dieser Arbeit mehr Datensätze imputiert wurden ($m = 50$), als Anteile an Missing Values vorlagen (35 %). Um in der empirischen Praxis prüfen zu können, ob *m* hoch genug ist, könnte ein Vergleich zwischen den Ergebnissen einer Direct-ML-Schätzung und der Schätzung auf MI-imputierten Daten vorgenommen werden. Ist *m* ausreichend, sollten sich die Ergebnisse eher nicht unterscheiden, denn wie

diese Arbeit zeigt, gibt es in Bezug auf die Standardfehler nur geringfügige Unterschiede zwischen den MI-Techniken und Direct-ML.

Es können in den Ergebnissen zwar keine Verzerrungen für die Standardfehler ausgemacht werden. Allerdings können mittels der modellbasierten Analyse einige Einflussfaktoren identifiziert werden, welche die Chance von verzerrten Standardfehlern erhöhen. Zu nennen ist dabei zunächst die Samplegröße[124] (womit ein Nachweis für die Konsistenz vorliegt und die Schätzung der Standardfehler bei größeren Fallzahlen weniger verzerrt ist), gefolgt von den Missinganteilen und den Verteilungen. Zwar wird mit der Verteilungsvariablen immer genügend Varianz gebunden, sodass der Grenzwert überschritten wird, allerdings zeigt sich, dass sich für diese Variable kein signifikanter Effekt ergibt, sondern der Verteilungseffekt über die Missinganteile vermittelt wird. Damit ist die Variablenverteilung nur einflussreich, wenn höhere Quoten an Missings vorliegen. Gleichzeitig kann zwar auch ein Effekt der Missinganteile beobachtet werden, das gilt allerdings nur, wenn eine kleine Fallzahl vorliegt. Im Umkehrschluss bedeutet dies, dass Direct-ML und die MI-Techniken, sollte eine größere Fallzahl gegeben sein, robust gegenüber den Missinganteilen sind. Es kann dementsprechend davon ausgegangen werden, dass die Missinganteile bei genügend hohen Fallzahlen eher keinen Effekt aufweisen und nur in Kombination mit kleinen Fallzahlen oder asymmetrischen Verteilungen problematisch sind. Zudem sind die Unterschiede in den b-Koeffizienten zwischen diesen MDTs auch allesamt eher unbedeutend. Für die Praxis bedeutet dies, dass unter denselben Bedingungen, relativ ähnliche Ergebnisse bei der Schätzung der Standardfehler zu erwarten sind, sofern *m* für die MI-Techniken hoch genug ist und sofern das spezifizierte Modell für Direct-ML und H0 korrekt ist (siehe dazu die Diskussion in Kapitel 8.4). Sollten also nur die Standardfehler von Interesse sein, dann können alle diese MDTs empfohlen werden. Das liegt auch daran, weil keine Bedingungen existieren, nach denen eine MDT einer anderen vorzuziehen wäre.

Für die Schätzungen mit EM liegen stark negativ verzerrte Standardfehler vor. Bereits bei einem 5 %igen Missinganteil können dabei negative Verzerrungen beobachtet werden, wobei der SE-Bias die gesetzten Grenzen von ± 15 % hierbei noch nicht überschreitet. Ab 20 % an Missing Values kann kein relativer SE-Bias mehr beobachtet werden, der als unverzerrt einzuordnen ist. Wird EM eingesetzt, können signifikante Effekte aufgrund der ersetzten Werte zu-

[124] Wie schon bei den Parametern, kann in den Meta-Modellen ein Einfluss der Samplegröße identifiziert werden, welcher in der deskriptiven Auswertung zum relativen SE-Bias nicht aufgetreten ist. Die Gründe dafür, können der Diskussion in Kapitel 8.4 entnommen werden.

stande kommen, da der Standardfehler kleiner wird. Der Zusammenhang zwischen den Missinganteilen und der Zunahme im SE-Bias kann auch mit den Meta-Modellen bestätigt werden. Hier zeigt sich, dass die Missinganteile den größten Anteil an Varianz binden (über 60 % der Varianz in der abhängigen Variablen). Damit besitzt diese Variable die größte Erklärungskraft. Zudem zeigt sich in der modellbasierten Analyse, dass einerseits von den Verteilungen wieder nur ein vermittelter Effekt ausgeht und andererseits, dass der Effekt der Fallzahl bisweilen weniger ausgeprägt ist als für die anderen MDTs (erneut ein Nachweis der Konsistenz). Dementsprechend ist für EM zunächst entscheidend, wie viele Missing Values vorliegen. Ob dabei kleine Fallzahlen oder asymmetrische Verteilungen gegeben sind ist nebensächlich, weil aufgrund der Haupteffekte der Missings bereits von einem erhöhten SE-Bias auszugehen ist. Liegen zusätzlich zu hohen Missingquoten noch kleine Fallzahlen oder asymmetrische Verteilungen vor, dann erhöht sich der SE-Bias nochmals. Gleichzeitig zeigt sich, dass EM mehr Verzerrungen produziert als die anderen MDTs: Nahezu alle Effekte der Meta-Modelle sind für EM signifikant größer als für die anderen MDTs. Insgesamt ist von einem Einsatz von EM abzuraten, es sei denn es liegen nur geringe Anteile an fehlenden Werten vor. Tabelle 21 fasst die Ergebnisse zur Performanz der MDTs hinsichtlich der Standardfehler zusammen.

Tabelle 21: Zusammenfassung der Performanz bzgl. der Schätzung der Standardfehler

	EM	**Direct-ML**	**H0**	**EMB, FCS, MNV, PMM**
Verzerrungen in der Schätzung der Standardfehler				
	stark neg. verzerrt wenn MA > 5 %	vernachlässigbar	vernachlässigbar	vernachlässigbar
Einflussfaktoren auf den SE-Bias				
Kleine Fallzahl (FZ)	+	+	+	+
Zunehmende Asym.	x	x	x	x
	robust gegenüber Verteilungen bei geringem MA; Beeinträchtigung durch Verteilungen erst bei erhöhtem MA (≥ 20 %)			
Steigender Missinganteil (MA)	++	x	x	x
	wird durch kl. FZ / (starke) Asym. vergrößert	tendenziell robust gegenüber Missinganteilen bei gr. FZ; Beeinträchtigung durch Missinganteil erst bei kl. FZ / (starker) Asym.		
Empfehlung	einsetzbar bis 5 % MA	unter allen Bedingungen zur Schätzung der Standardfehler geeignet		
Einschränkung	Standardfehler ist auch bei 5 % MA neg. verzerrt	korrektes Modell	korrektes Modell; gr. *m* bei hohem MA wählen	gr. *m* bei hohem MA wählen **FCS**: ggf. signifikante Kovarianzen aufgrund unterschätzter SE

Anmerkungen: x: kein Effekt; +/- positiver/negativer Effekt; ++/-- stark positiver/negativer Effekt.

9.3.1 Einordnung der Ergebnisse

Mit den Ergebnissen zu den Standardfehlern sind die Analysen der einzelnen MDTs abgeschlossen. Es zeigt sich dabei, dass mittels FCS und PMM relativ unverzerrte Standardfehler zu erwarten sind, was bereits die Studien von van Buuren u. a. (2006), Kleinke (2017) oder Zhang u. a. (2017) zeigen. Eine Über- oder Unterlegenheit von FCS gegenüber EMB und MNV, wie sie sich in den Ergebnissen von Wu u. a. (2015) respektive Lee/Carlin (2010) zeigt, lässt sich in dieser Arbeit nicht beobachten. Erneut zeigen sich die logistischen Regressionen für FCS als gute Möglichkeiten, um Missing Values zu ersetzen. Probleme, wie es bei Jia/Wu (2019) der Fall ist, konnten nicht repliziert werden.

Auch die Ergebnisse zu den anderen MDTs wie Direct-ML, EMB oder MNV können reproduziert werden. Zu H0 liegen bisher aufgrund der geringen Anwendungszahl keine Ergebnisse vor. Enders (2001a) zeigt, dass Direct-ML unverzerrte Standardfehler generiert. Er weist zudem nach, dass die Verteilungen der Variablen für Direct-ML einen Einfluss auf die Verzerrungen im Standardfehler nehmen. Zwar kann dieser Effekt bei geringen Anteilen an Missing Values in dieser Arbeit nicht nachgewiesen werden, wohl aber für größere Anteile. Liegen demnach höhere Anteile an fehlenden Werten vor, dann ist die Performanz von Direct-ML davon abhängig, ob zudem noch asymmetrische Verteilungen gegeben sind. Weiterhin lässt sich anderen Studien entnehmen (Enders 2001c; Ferro 2014; Gold/Bentler 2000; Graham u. a. 1996; Honaker/King 2010; King u. a. 2001), dass für Direct-ML, EMB und MNV der SE-Bias mit zunehmenden Missinganteilen und/oder abnehmender Samplegröße ansteigt, in den meisten Fällen aber mit zufriedenstellenden Ergebnissen zu rechnen ist und diese MDTs auch ähnliche Ergebnisse generieren. Auch dies wird mit der vorliegenden Arbeit bestätigt.

Während demnach die MI-Techniken überzeugen, zeigt sich für EM ein wenig zufriedenstellendes Bild. Sowohl Li (2010), Li/Lomax (2017) als auch Newman (2003) können nachweisen, dass EM im Vergleich zu Direct-ML oder MNV bei hohen Missinganteilen weniger zuverlässig ist. Zudem weisen diese Studien eine Abhängigkeit der Performanz von EM in Bezug auf die Variablenverteilungen nach. Wenn demnach nicht normalverteilte Variablen vorliegen, verschlechtert sich die Performanz von EM zusehends. Dies bestätigt die vorliegende Arbeit. Es zeigt sich, dass Direct-ML und MNV vergleichbar gute Ergebnisse erbringen, und dass sie weniger durch die Missinganteile beeinflusst sind als es bei EM der Fall ist. In der vorliegenden Arbeit kann mit den Meta-Modellen zudem ein eindeutiger Zusammenhang zwischen den Verzerrungen der Standardfehler und der Missingvariable nachgewiesen werden.

Zusätzlich zeigt sich, dass diese Sensitivität gegenüber dem Anteil an fehlenden Werten zunimmt, wenn asymmetrische Verteilungen vorliegen. Hinsichtlich EM zeigt die Listung des Forschungsstandes allerdings auch, dass damit in anderen Fällen korrekte Standardfehler geschätzt werden können (Olinsky u. a. 2003). Da die vorliegenden Ergebnisse für EM den analytischen Annahmen entsprechen, wonach mit einer Einfachimputation die Standardfehler unterschätzt werden, könnte es sein, dass mit dem Ergebnis von Olinsky u. a. (2003) ein nur für deren Forschungsdesign spezifisches Ergebnis gegeben ist, zumal in allen Modellen dieser Arbeit die Standardfehler mittels EM unterschätzt werden.

Diese Arbeit indiziert, dass die Schätzung der Standardfehler weniger davon abhängig ist, welche Dateneigenschaften gegeben sind, oder welche Modelle geschätzt werden, als vielmehr davon, welche MDT angewendet wird: Denn unabhängig von den Gegebenheiten, so zeigt sich auch in anderen Studien, werden mit EM zu kleine Standardfehler, mit den anderen MDTs aber unverzerrte Standardfehler geschätzt.

10 Hypothesentest und Ergebniszusammenfassung

Insgesamt wurden in dieser Arbeit sieben Hypothesen abgeleitet und drei Fragestellungen zu den Fit-Indices formuliert. Zunächst wird der Hypothesentest durchgeführt, danach werden die Ergebnisse zusammengefasst, um im Anschluss die Fragen zu den Fit-Indices zu beantworten.

ad Hypothese 1: Liegen nur (quasi-)metrische Variablen vor (Modell 1), die zudem (symmetrisch) normalverteilt sind, dann lassen sich nur geringe Unterschiede in den Schätzungen zwischen den MDTs, bezogen auf die Parameter und die Standardfehler, beobachten. In kaum einer Simulationskonfiguration können Verzerrungen beobachtet werden, welche die Grenzwerte überschreiten. Demzufolge ist Hypothese 1 für alle MDTs für die Schätzung der Parameter und der Standardfehler zu bestätigen (zweites mit Ausnahme für EM; siehe Hypothese 6 und 6.1).

ad Hypothese 2: Auch wenn die (quasi-)metrischen Variablen die (multivariate) Normalverteilungsannahme verletzen, werden in den Schätzungen keine gravierenden Verzerrungen erzeugt. Zwar lassen sich Verteilungseinflüsse beobachten, wonach die Verzerrungen in den Parameterschätzwerten und Standardfehlern (über die Missinganteile vermittelt) etwas größer werden, wenn keine (symmetrisch) normalverteilten Variablen vorliegen. Jedoch sind diese Einflüsse nur wenig bedeutsam. Zudem liegen für EM zwar verzerrte Standardfehler vor, diese Verzerrungen sind aber laut den Meta-Modellen nicht auf die Verteilungen, sondern auf die Missinganteile zurückzuführen (siehe Hypothese 6 und 6.1). Damit kann Hypothese 2 für PMM widerlegt und für die anderen MDTs bestätigt werden, weil bei asymmetrischen Verteilungen

mit etwas größeren Verzerrungen zu rechnen ist. Es sollte aber beachtet werden, dass die Zunahmen der Verzerrungen durch asymmetrische Verteilungen meist bedeutungslos sind und keine praktischen Folgen[125] für die letztlichen Modellschätzungen haben.

ad Hypothese 3: Werden SE-Modelle mit symmetrisch (normal) verteilten, unterschiedlich skalierten Variablen geschätzt, welche neben (quasi-)metrischen auch diskrete Skalenniveaus aufweisen (Modell 2 und Modell 3), ergeben sich für alle MDTs nur unwesentlich verzerrte Schätzungen für die Parameter und nur unwesentlich verzerrte Standardfehler (zweites mit Ausnahme für EM; siehe Hypothese 6 und 6.1). In den meisten Fällen werden keine Grenzwerte überschritten, sodass Hypothese 3 für PMM zurückzuweisen ist, für alle anderen MDTs aber bestätigt werden kann.

ad Hypothese 4: Auch wenn die für das SE-Modell verwendeten, unterschiedlich skalierten Variablen zunehmend asymmetrisch verteilt sind, ergeben sich keine unverhältnismäßigen Verzerrungen. Es lassen sich zwar Verteilungseinflüsse auf das Ausmaß der Verzerrungen der Parameter und – vermittelt über die Missinganteile – der Standardfehler erkennen, allerdings sind diese Einflüsse kaum problematisch. Für EM liegen wiederum häufig verzerrte Standardfehler vor, diese lassen sich aber auf die Missinganteile zurückführen und nicht auf die Verteilungen der Variablen (siehe Hypothese 6 und 6.1). Dementsprechend werden mit zunehmend asymmetrisch verteilten, gemischten Daten, die Verzerrungen für alle MDTs zwar etwas größer, allerdings sind diese Zunahmen bedeutungslos und haben keine Folgen für die letztlichen Modellschätzungen. Damit ist Hypothese 4 für FCS und PMM zurückzuweisen und für die anderen MDTs zu bestätigen.

ad Hypothese 5: Für alle MDTs lässt sich mittels der Meta-Modell-Analysen ein Effekt der Samplegröße nachweisen. Dementsprechend nehmen die Verzerrungen für alle MDTs zu, egal ob es sich dabei um die Verzerrungen der Parameter oder Standardfehler handelt, wenn die Fallzahl kleiner wird. Demzufolge kann Hypothese 5 für alle MDTs bestätigt werden: Bei kleinen Fallzahlen liegen stärker verzerrte Ergebnisse vor, als bei größeren Fallzahlen. Für die MI-Techniken (bzgl. der Parameter und Standardfehler) und für Direct-ML (bzgl. der Standardfehler) ist es dann auch die Samplegröße, welche diejenige Einflussgröße darstellt, die entscheidend ist für das Ausmaß an Verzerrungen. Für EM ist das für die Parameter und die Standard-

[125] ‚Folgen' in dem Sinne, als dass die inferenzstatistischen Schlüsse, beruhend auf den Modellschätzungen, die auf den imputierten Daten bzw. auf den Daten mit fehlenden Werten bei Direct-ML vorgenommen werden, nicht beeinträchtigt sind und die Schätzergebnisse korrekte inhaltliche Schlussfolgerungen erlauben, weil unverzerrte Parameter und Standardfehler geschätzt werden.

fehler der Missinganteil. Weil aber die Verzerrungen in den Parametern und den Standardfehlern für alle MDTs auch bei kleinen Fallzahlen vernachlässigbar bleiben und die MDTs durch die Fallzahl nicht derart beeinflusst werden, dass dies Folgen für die Modellschätzungen hat, können die Zunahmen in den Verzerrungen durch die Abnahme der Samplegröße auch vernachlässigt werden. Bis zu Samplegrößen von ca. 250 Fällen bleiben die Verzerrungen unproblematisch.

ad Hypothese 6: Der Missinganteil kann für Direct-ML (bzgl. der Parameter) und für EM (bzgl. der Parameter und Standardfehler) durch die Meta-Modelle als einflussreichste Größe auf das Ausmaß an Verzerrungen identifiziert werden. Zudem ist der Missinganteil auch für die MI-Techniken eine einflussreiche Größe, wenngleich für diese MDTs die Samplegröße bedeutender ist. Letztlich lassen sich mit zunehmendem Missinganteil größere Verzerrungen beobachten, was darin resultiert, dass Hypothese 6 für alle MDTs bestätigt werden kann. Allerdings bleiben die Verzerrungen in den Parametern und den Standardfehlern (zweites mit Ausnahme für EM; siehe Hypothese 6.1) durch den zunehmenden Missinganteil ohne Folgen für die Modellschätzung. Demnach ist zwar bei hohen Missinganteilen mit etwas stärker verzerrten Schätzungen zu rechnen, da diese Verzerrungen aber empirisch unbedeutend bleiben, können mit den MDTs, auch bei höheren Anteilen an fehlenden Werten, zufriedenstellende Schätzungen der Parameter und Standardfehler erwartet werden.

ad Hypothese 6.1: Die Standardfehler mit EM werden teilweise sehr stark unterschätzt. Ab einem Missinganteil von 20 % kann keine unverzerrte Schätzung derselben mehr beobachtet werden. Der Anteil an fehlenden Werten ist für EM in den Meta-Modellen zum SE-Bias zudem auch stärkster Prädiktor. Demzufolge kann es, aufgrund der unterschätzten Standardfehler, beim Einsatz von EM vermehrt zu falschen inhaltlichen Schlüssen kommen. Das hat die Bestätigung von Hypothese 6.1 zur Folge.

ad Hypothese 7: Tendenziell liegen für alle MDTs etwas größere Verzerrungen für die Parameter und die Standardfehler vor, wenn zwei oder drei Kombinationen an Einflussgrößen gegeben sind. Das zeigen neben den deskriptiven Analysen zum Parameter- und Standardfehlerbias auch die Meta-Modell-Analysen, wonach sich viele der Effekte der unabhängigen Variablen verstärken oder erst dann einflussreich werden, wenn zusätzlich eine Interaktion aus einer weiteren Einflussgröße berücksichtigt wird (wie der, durch die Missinganteile vermittelte, Verteilungseffekt auf den SE-Bias). Demnach ist Hypothese 7 für alle MDTs zu bestätigen: Mit Kombinationen an Einflussgrößen gehen tendenziell stärkere Verzerrungen einher. Wieder haben die Verzerrungen nur für EM Folgen für die Modellschätzungen. Denn durch die negativ

verzerrten Standardfehler könnten die Modellschätzungen signifikante Effekte aufweisen, die nicht signifikant sein sollten. Für alle anderen MDTs liegt keine Kombination an Einflussgrößen vor, mit der die Modellschätzung beeinflusst wird. Demnach ist Hypothese 7 zwar zu bestätigen, allerdings gilt auch, dass selbst unter kleinen Fallzahlen, bei stark asymmetrischen Verteilungen und hohen Missinganteilen keine derartigen Verzerrungen vorliegen, als dass Folgen für die Modellschätzungen eintreten. Es können selbst unter diesen wenig optimalen Bedingungen, zufriedenstellende Schätzungen der Parameter und Standardfehler erwartet werden (zweites mit Ausnahme für EM; siehe Hypothese 6 und 6.1).

Diskussion zu den widerlegten Hypothesen: Letztendlich entsprechen viele der Ergebnisse den aufgestellten Erwartungen, auch wenn die Einflussgrößen nicht in dem Maße Wirkung zeigen, als dass damit verzerrte Ergebnisse einhergehen. Für PMM können allerdings die aufgestellten Hypothesen in insgesamt drei Fällen widerlegt werden. Das betrifft die Performanz von PMM bei diskreten Variablen (Hypothese 3) sowie deren Robustheit gegenüber den Verteilungen (Hypothese 2 und 4). Hier zeigt sich einerseits, dass PMM keine Probleme hat, fehlende Werte auf diskreten Variablen zu ersetzen und andererseits, dass PMM trotz dessen, dass damit beobachtete Fälle imputiert werden, anfällig gegenüber asymmetrisch verteilten Variablen ist. Zum ersten Punkt lässt sich anführen, dass mit PMM (wenn PMM mit dem Standardwert für das Spenderset von fünf Spendern verwendet wird) plausible Imputationen generiert werden, auch wenn durch die wenigen Skalenpunkte auf diskreten Variablen potentiell nur wenige Spender zur Verfügung stehen. Demzufolge wiederspricht das vorliegende Ergebnis zwar der aufgestellten Erwartung, entspricht allerdings der Diskussion in der Literatur, wonach sich PMM für alle Skalenniveaus eignet. Mit der vorliegenden Arbeit werden diese Aussagen erstmalig durch MC-Ergebnisse gestützt (siehe viertes Desiderat in Kapitel 5.4). Zum zweiten Punkt kann angemerkt werden, dass zwar (vermittelte) Verteilungseinflüsse für PMM vorliegen, diese allerdings nicht dazu beitragen, dass sich die Performanz von PMM im Hinblick auf die Schätzung der Parameter und der Standardfehler so stark verschlechtert, als dass damit keine zufriedenstellenden Modellschätzungen mehr einhergehen. Es ist davon auszugehen, dass PMM robust gegenüber den Variablenverteilungen ist, zumal deren Einflussstärke im Vergleich zur Samplegröße und zu den Missinganteilen weniger bedeutsam ist.

Zusätzlich liegt auch eine Zurückweisung der Hypothese 4 für FCS vor. Denn für FCS kann die Erwartung nicht bestätigt werden, wonach durch die logistischen Regressionsmodelle, aufgrund asymmetrisch verteilter Daten, Probleme im Schätzprozess der fehlenden Werte entstehen. Zum einen sind alle Modellschätzungen auf den FCS-imputierten Daten konvergiert, was

davon zeugt, dass der Imputationsprozess selbst erfolgreich ist und die logistischen Regressionsmodelle zumindest Imputationen generieren. Das wiederum bedeutet, dass aufgrund von gering besetzten Zellen, welche durch die asymmetrischen Verteilungen entstehen können, die logistischen Regressionsmodelle nicht beeinträchtigt werden. Zum anderen erweisen sich die imputierten Werte dann in den meisten Fällen auch als plausibel, denn die Parameter und Standardfehler der Modellschätzungen sind in der Regel unverzerrt. Somit führen die logistischen Regressionsmodelle nicht zu unplausiblen Imputationen, welche die Modellschätzungen verzerren, sodass keine korrekten Ergebnisse mehr vorliegen. Demzufolge können mit den logistischen Regressionsmodellen Imputationen auf diskreten Variablen vorgenommen werden.

Zusammenfassung: Aus den Meta-Modell-Analysen geht hervor, dass für die Verzerrungen in den Parameterschätzwerten vor allem die Samplegröße und der Missinganteil verantwortlich sind, und nur teilweise die Variablenverteilungen. Auch in Bezug auf die geschätzten Standardfehler können die Meta-Modell-Analysen zeigen, dass die Samplegröße und der Missinganteil für die Verzerrungen entscheidend sind. Vernachlässigbar erscheinen dagegen die Verteilungen, weil diese nur einen, über die Missinganteile, vermittelten Effekt ausüben. Es können also durch die Meta-Modell-Analysen Zusammenhänge zwischen den unabhängigen Variablen auf die Verzerrungen in den Parametern und den Standardfehlern nachgewiesen werden. Damit ist auch anzunehmen, dass sich mit noch kleineren Fallzahlen und/oder noch höheren Missinganteilen (und/oder noch stärker asymmetrisch verteilten Variablen) auch größere Verzerrungen ergeben. Ob diese Verzerrungen, wie es in dieser Arbeit in der Regel für die Parameter und die Standardfehler der Fall ist, vernachlässigbar bleiben, oder ob dann letztlich doch mit problematischen Verzerrungen zu rechnen ist, kann an dieser Stelle nicht beantwortet werden. Schlussendlich gilt für die MDTs in Bezug auf die Schätzung der Parameter und Standardfehler aber, dass diese auch bei nicht optimalen Bedingungen, wie etwa Fallzahlen von 250, stark asymmetrischen Verteilungen und hohen Missinganteilen von 35 %, zufriedenstellend arbeiten und plausible Imputationen generieren, sodass die auf den imputierten Daten durchgeführten Modellschätzungen (bzw. die direkten Modellschätzungen mittels Direct-ML) unverzerrt sind und korrekte Ergebnisse liefern.

10.1 Antworten auf die Fragen zu den Fit-Indices

ad Frage 1: Von den MI-Techniken kann in Bezug auf die Fit-Indices nur H0 überzeugen. In allen Fällen, unabhängig vom gewählten Fit-Index und unabhängig von den simulierten Bedingungen, werden mittels H0 die Modelle korrekt beurteilt. Eine korrekte Modellbewertung mit den anderen MI-Techniken ist unter allen Bedingungen dagegen nur mit dem SRMR möglich.

Werden andere Fit-Indices als das SRMR herangezogen, sollten die vorliegenden Daten darüber entscheiden, welcher Fit-Index zur Modellbewertung benutzt wird. Denn nachdem die fehlenden Werte mit den MI-Techniken ersetzt wurden, ist die korrekte Modellbewertung mit dem p-Wert, RMSEA und dessen Konfidenzintervall und auch mit dem CFI davon abhängig, welche Dateneigenschaften gegeben sind.

ad Frage 2.1: Wie mit H0 werden auch mit Direct-ML unter den meisten Umständen korrekte Modellbewertungen vorgenommen. Direct-ML zeigt sich dabei auch robust gegenüber den Dateneigenschaften. Es ist demnach für diese MDT unerheblich, welcher Fit-Index zur Modellbewertung herangezogen wird und ob ein kleines Sample, asymmetrische Verteilungen oder erhöhte Missinganteile vorliegen.

Im Gegensatz zu Direct-ML und H0, sind korrekte Modellbewertungen (mit Ausnahme durch das SRMR) nach Imputation der fehlenden Werte mit EM, EMB, FCS, MNV und PMM nur unter bestimmten Bedingungen möglich. So können mit dem p-Wert bereits bei einem Anteil von 5 % an Missing Values inkorrekte Schlussfolgerungen bei der Modellbewertung auftreten und mit RMSEA gehen bei 5 % an Missings nur bei großen Fallzahlen zufriedenstellende Schlussfolgerungen einher. Ab 20 % an Missing Values sollte auf den p-Wert und auf RMSEA verzichtet werden. Hier empfiehlt es sich das Konfidenzintervall von RMSEA oder den CFI heranzuziehen. Beide liefern bei größeren Fallzahlen und 20 % Missings noch zufriedenstellende Modellbewertungen (für kleine Fallzahlen gilt aber auch dies nicht). Liegt der Missinganteil noch höher, sollte nur der CFI eingesetzt werden. Auch dieser ist aber nur bedingt empfehlenswert, weil die Chance einer Zurückweisung des Modells auch mit dem CFI erhöht ist.

ad Frage 2.2: Wie die Ausführungen zu Frage 2.1 zeigen, sind die Haupteinflussgrößen für die korrekte Modellbewertung durch den p-Wert, RMSEA und dessen Konfidenzintervall sowie durch den CFI, nachdem die Imputationstechniken eingesetzt wurden (mit Ausnahme von H0), der Missinganteil und die Samplegröße. Für beide zeigen sich in den Meta-Modellen derart starke Effekte, dass der Einfluss der Variablenverteilungen kaum mehr ins Gewicht fällt. Zudem lässt sich den Meta-Modell-Analysen auch entnehmen, dass der Effekt der Missinganteile zwar über die Erhöhung der Fallzahl reduziert werden kann, aber nicht in dem Maße, als dass er komplett kompensiert wird. Grundsätzlich ist es vor allem bei erhöhten Missinganteilen notwendig eine größere Fallzahl als Grundlage zu nutzen. Auch das garantiert aber keine korrekte Modellbewertung. Dagegen lassen sich für Direct-ML, H0 sowie für das SRMR keine Einflussgrößen identifizieren. Sie sind robust gegenüber den Dateneigenschaften.

11 Exemplifizierung der MC-Ergebnisse

Zur Verdeutlichung der Ergebnisse folgt an dieser Stelle die Anwendung der MDTs bei der Schätzung eines SE-Modells auf empirischen Daten. Gleichzeitig soll das vorliegende Kapitel als Hinweis für die Übertragbarkeit und Gültigkeit der MC-Ergebnisse dienen. Für diese beispielhafte Anwendung müssen allerdings einige Punkte angesprochen werden. Sie betreffen das zu schätzende Modell, die Datengrundlage und die Bedeutung für die empirische Forschung.

Im Grunde sind Ergebnisse aus MC-Studien immer auf die modellierten Bedingungen und auf die untersuchten Populationsmodelle (für dieses Kapitel: MC-Modelle) zurückzuführen. Sie können also nur teilweise verallgemeinert werden. Um für das vorliegende Beispiel ähnliche Ergebnisse zu erhalten, müsste einerseits das zu schätzende Modell den MC-Modellen entsprechen (in der Struktur und in den Parameterschätzwerten) und andererseits müssten die empirischen Daten den Simulationsbedingungen möglichst nahekommen (die Skalierungen und Verteilungen der Variablen, die Fallzahl und der Anteil an fehlenden Werten). Wie sich allerdings in den MC-Ergebnissen zeigt, macht es für die MDTs keinen Unterschied, welche Modellstruktur vorliegt, denn für alle drei MC-Modelle sind die Ergebnisse ähnlich. Demzufolge nimmt die Struktur des zugrundeliegenden Modells keinen Einfluss auf die Performanz der MDTs. Aus diesem Grund wird an dieser Stelle keines der MC-Modelle reproduziert, sondern ein Modell geschätzt, das eine andere Struktur aufweist. Wichtig dabei ist, dass es sich nicht um ein willkürliches SE-Modell handelt, sondern dass dieses analytisch begründbar ist. Weiterhin bedeutet dies, dass damit analytisch abgeleitete Hypothesen getestet werden könnten. Weil die vorliegende Arbeit aber primär das methodische Ziel verfolgt, verschiedene MDTs zu evaluieren, dieses Beispiel für die Beantwortung der Forschungsfragen keine Relevanz hat und es sich nicht um eine Arbeit handelt, die eine inhaltlich bedeutsame Fragestellung anhand empirischer Daten untersucht, wird kein eigenes Modell abgeleitet. Stattdessen wird aus der Literatur ein Modell herangezogen, das bereits auf empirischen Daten geschätzt und für Hypothesentests benutzt wurde. Da das Modell aus der bisherigen Forschung entlehnt wird, wird sich dieses auch in den Parameterschätzwerten von den MC-Modellen unterscheiden. Dies ist aber auch dann der Fall, wenn eines der MC-Modelle mit empirischen Daten geschätzt wird. Letztendlich könnte nur über die analytische Begründung und die Operationalisierung des empirischen Modells Einfluss auf die Ausprägung der Parameterschätzwerte genommen werden. Ein bereits getestetes, empirisches Modell, das zudem noch unterschiedliche Parameterschätzwerte gegenüber den MC-Modellen aufweist, sollte einen Hinweis geben, ob die MC-Ergebnisse der MDTs auf die Empirie und andere Modelle übertragen werden können.

Während sich das Modell von den MC-Modellen unterscheidet, wird die Datengrundlage an den Simulationskonfigurationen ausgerichtet. Denn hier zeigt sich, dass vor allem kleine Fallzahlen und hohe Missinganteile das Ergebnis, insbesondere im Hinblick auf die Fit-Indices, beeinflussen können. Demzufolge werden für dieses Beispiel die Fallzahl und der Missinganteil an die Simulationskonfigurationen angepasst. Auf die Verteilungen der Variablen kann hingegen kein Einfluss genommen werden (wie die MC-Ergebnisse aber zeigen, sind diese auch eher bedeutungslos). Weiterhin sollen für das Modell gemischte Daten verwendet werden. Das hat zur Folge, dass die Variablenskalierungen der verwendeten Variablen angepasst werden, sodass diese binäre, ordinale und quasi-metrische Skalenniveaus aufweisen.

Aufgrund der Tatsache, dass die empirischen Daten, was die Fallzahl, die Missinganteile und die Variablenskalierungen betrifft, auf die Simulationskonfigurationen zugeschnitten werden, besitzen die Schätzergebnisse des SE-Modells keine praktische Relevanz für die empirische Forschung. Stattdessen sollen die Analysen zeigen, welche Auswirkungen der Einsatz der MDTs auf die Schätzung eines SE-Modells mit empirischen Daten haben kann. Im Hinblick auf diese Auswirkungen lassen sich, aufgrund der MC-Ergebnisse, vier Erwartungen ableiten:

1. Die Fit-Indices werden sich zwischen den MDTs, die bei der Behandlung der fehlenden Werte die Modellstruktur berücksichtigen (Direct-ML und H0), von den anderen MDTs unterscheiden; das kann zu unterschiedlichen Ergebnissen in der Modellbewertung führen (E1).
2. Die Parameterschätzungen werden für alle MDTs ähnlich sein (E2).
3. Mit Ausnahme von EM dürften auch alle MDTs relativ ähnliche Schätzungen für die Standardfehler liefern (E3).
4. EM dürfte kleinere Standardfehler schätzen, sodass es zu anderen inhaltlichen Schlussfolgerungen als mit den anderen MDTs kommt (E4).

11.1 Empirisches Beispielmodell

Im Fokus der Anwendung steht ein Modell aus Herrmann (2001). Aus der Arbeit der Autorin wird ein Modell gewählt, da sie erstens eine breit gestreute empirische Analyse mittels SE-Modellen durchführt, da sie zweitens in ihrer Arbeit mit Daten des ALLBUS arbeitet, sodass das Modell und die Operationalisierung direkt übernommen werden können und da drittens ihre bearbeitete Thematik, nämlich die Erklärung von Ethnozentrismus, weiterhin von großer soziologischer Relevanz ist. Das zeigen sowohl die Forschungsergebnisse zur gruppenbezogenen Menschenfeindlichkeit, als auch Ergebnisse neuerer Studien zu autoritären und rechtsextremen

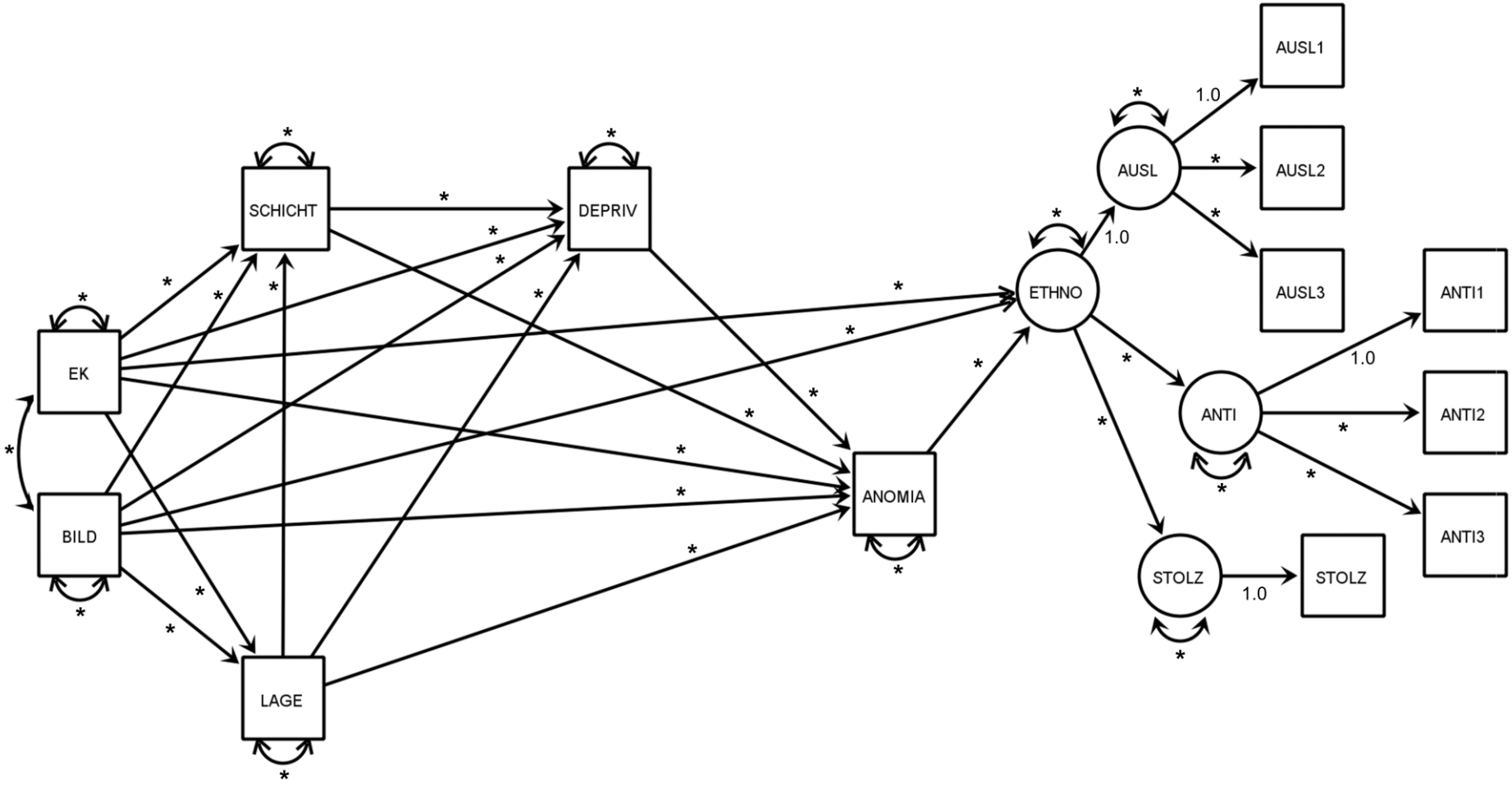
Einstellungen. Hierbei zeigt sich, dass in Deutschland die Ausländerfeindlichkeit, die Zustimmung zu rechtsextremistischem Gedankengut oder die negative Wahrnehmung von Muslimen zunehmen und dass zudem ein nach wie vor weit verbreiteter Antisemitismus vorherrscht (siehe Brähler/Decker 2018).

Die angesprochenen Einstellungen sind einzelne Dimensionen des gemeinsamen latenten Konstruktes Rechtsextremismus. Dieser kennzeichnet sich erstens durch ablehnende und abwertende Einstellungen gegenüber allen als fremd konstruierten Gruppen, zweitens durch die Überbewertungen der Eigengruppe (z. B. durch einen ausgeprägten Nationalstolz) und drittens durch spezielle politisch-weltanschauliche Vorstellungen, wie die Verherrlichung des Nationalsozialismus (vgl. Heyder/Gaßner 2012: 280 ff.). Ethnozentrismus berücksichtigt im Gegensatz zum Rechtsextremismus keine politisch-weltanschaulichen Vorstellungen, sondern nur die Abwertung der Fremd- und Aufwertung der Eigengruppe (vgl. Rippl/Baier 2005: 653).

Die Ursachen für rechtsextremistische Einstellungen oder Ethnozentrismus sind empirisch gut belegt. Zu nennen ist dabei das Konzept der Anomie, der Autoritarismus, die Kontakthypothese oder bestimmte sozio-demographische und -ökonomische Merkmale wie der Wohnort (Ost- oder Westdeutschland), das Alter, die Bildung oder das Deprivationskonzept. Herrmann richtet den Fokus ihrer Analysen, unter anderem, auf die Erklärung der Anomie und die Anomie als Ursache des Ethnozentrismus. Anstatt allerdings die Bezeichnung Anomie zu verwenden, wird explizit die Bezeichnung der Anomia gebraucht. Das soll verdeutlichen, dass damit „das fehlende individuelle Gefühl, ausreichend in die Gesellschaft integriert zu sein“ (Herrmann 2001: 90) gemeint ist und nicht „ein Problem der sozialen Ordnung, der sozialen Regulierung und der sozialen Integration“ (ebd.: 90). Entsprechend wird diese Bezeichnung auch für diese Arbeit übernommen. Als Ursachen der Anomia führt Herrmann, Durkheim folgend, die „individuelle Unzufriedenheit mit den ökonomischen Bedingungen und die Zunahme einer nicht mehr gerechtfertigten sozialen Ungleichheit“ (ebd.: 108) an. Operationalisiert wird dies mit der subjektiven Schichteinstufung, der wahrgenommenen Deprivation und der Einschätzung der wirtschaftlichen Lage. Abbildung 16 stellt das Modell dar.

In Bezug auf die Zusammenhänge wird folgendes erwartet: Je geringer die Schichteinstufung (SCHICHT), je höher die Deprivation (DEPRIV) und je schlechter die Einschätzung der eigenen wirtschaftlichen Lage (LAGE) ist, desto größer ist das Ausmaß der Anomia. Für den Zusammenhang zwischen der Anomia und dem Ethnozentrismus (ETHNO) wird ein positiver Effekt erwartet. Die Bildung (BILD) und das Haushaltsnettoeinkommen (EK) dienen als Kontrollvariablen.

Abbildung 16: Empirisches Beispiel-SE-Modell



Quelle: eigene Darstellung nach Herrmann (2001: 115).
Anmerkungen: Aufgrund der Übersichtlichkeit wird auf die Darstellung der Fehlervarianzen der Indikatoren verzichtet. Ein doppelseitiger Pfeil, der nur auf eine jeweilige Variable/ein jeweiliges latentes Konstrukt zeigt, stellt die Varianzen/Residualvarianzen dar. Im Gegensatz zum Modell von Herrmann werden keine Messfehlerkorrelationen und keine Fremdladungen zugelassen.

Zur Schätzung des vorgestellten Modells werden die Daten des ALLBUS 2016 (GESIS 2017) herangezogen und die entsprechenden Variablen ausgewählt. Diese lassen sich Tabelle 22 entnehmen. Für die Variablen werden entsprechende Kodierungen vorgenommen, um die Skalierungen an die MC-Modelle anzupassen. Weiterhin wird, anders als bei Herrmann, das Haushaltsnettoeinkommen als dichotome Variable in das Modell aufgenommen, um eine binäre Variable zu berücksichtigen. Im Unterschied zu Herrmann ist es zur Operationalisierung der Anomia erforderlich einen Index zu bilden, da die notwendigen Variablen nur binär skaliert sind und nicht auf einem quasi-metrischen Messniveau vorliegen. Die binär skalierten Variablen werden zu einem 5er skalierten additiven Index zusammengefasst.

Tabelle 22: Operationalisierung des Beispielmodells

Ethnozentrismus (ETHNO)	
Ausländerfeindlichkeit (AUSL) Skala: 1 *stimme gar nicht zu* – 5 *stimme voll zu* (sk.)	
AUSL1	Ausländer: Wieder heim bei knapper Arbeit
AUSL2	Ausländer: Politische Betätigung untersagen
AUSL3	Ausländer: Sollten unter such heiraten
Antisemitismus (ANTI) Skala: 1 *stimme gar nicht zu* – 5 *stimme völlig zu* (sk.)	
ANTI1	Juden haben auf der Welt zu viel Einfluss
ANTI2	Juden nutzen deutsche Vergangenheit aus
ANTI3	Juden an Verfolgung nicht unschuldig
Nationalstolz (STOLZ) Skala: 1 *gar nicht stolz* – 4 *sehr stolz*	
STOLZ	Genereller Stolz, Deutscher zu sein (rec.)
Prädiktoren[126]	
ANOMIA	Additiver Index (Skala: 0 *keiner Aussage zugestimmt* – 4 *allen Aussagen zugestimmt*) aus: • Lageverschlechterung für einfache Leute • Bei dieser Zukunft keine Kinder mehr • Politiker uninteressiert an einfachen Leuten • Mehrheit uninteressiert an Mitmenschen • Skala jeweils: 0 *bin anderer Meinung* 1 *bin derselben Meinung* (rec.)
LAGE	Einschätzung eigene wirtschaftliche Lage heute (Skala: 1 *sehr gut* – 5 *sehr schlecht*)
DEPRIV	Gerechter Anteil am Lebensstandard (Skala: 1 *mehr als gerechten Anteil* – 4 *sehr viel weniger*; rec.)
SCHICHT	Subjektive Schichteinstufung (Skala: 1 *Oberschicht* – 5 *Unterschicht*; rec.)
BILD	Allgemeinbildender Schulabschluss (Skala: 1 *HS*[127], 2 *MR*[128], 3 *FH*, 4 *Abitur*)
EK	Haushaltsnettoeinkommen: offene Abfrage (Skala 0 *bis 2500 Euro*, 1 *über 2500 Euro*) • Dichotomisierung durch Mediansplit

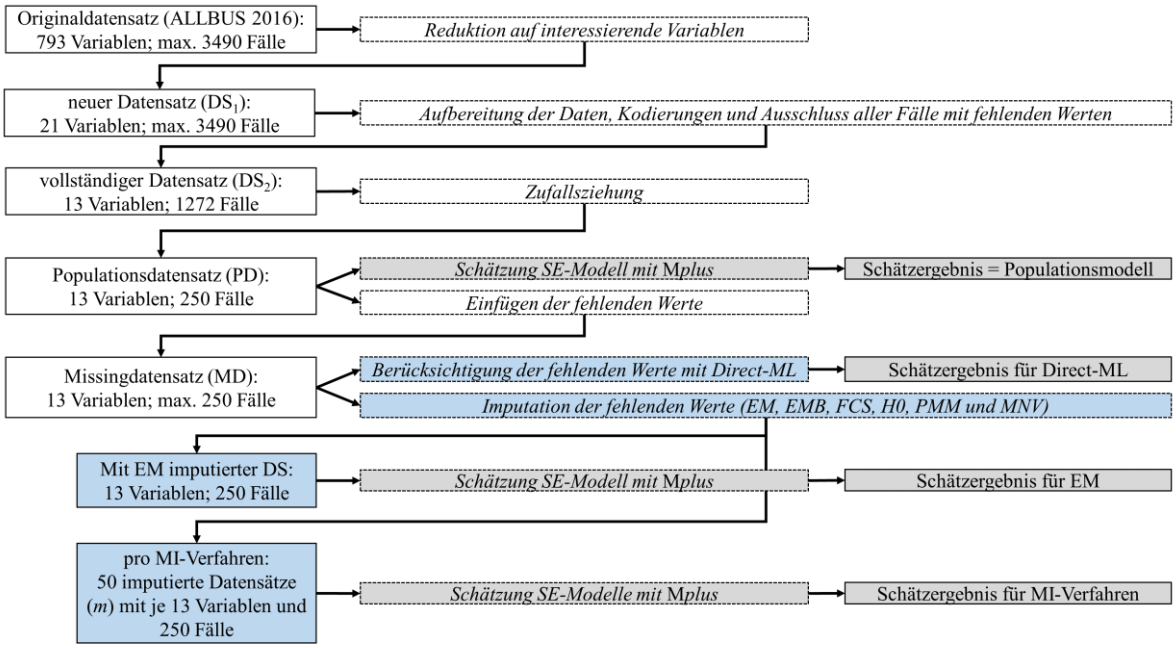
Anmerkungen: sk.: Skalierung an MC-Modelle angepasst; rec.: Skala gedreht.

[126] Die Prädiktoren sind zum Teil auch abhängige Variablen; siehe Abbildung 16.
[127] Darunter: Volks-/Hauptschulabschluss; Polytechnische Oberschule DDR, Abschluss 8. oder 9. Klasse.
[128] Darunter: Mittlere Reife, Realschulabschluss bzw. Polytechnische Oberschule mit Abschluss 10. Klasse.

11.1.1 Datenanpassung und Analyseprozess

Die Bearbeitung der Daten und der Analyseprozess sind in Abbildung 17 schematisch dargestellt. Zunächst stehen die weiß hinterlegten Felder im Fokus, denn diese veranschaulichen den Prozess der Anpassung der Ausgangsdaten an die Simulationskonfigurationen.

Abbildung 17: Aufbereitungs- und Analyseprozess



Im Anpassungsprozess wird zunächst der unbearbeitete Originaldatensatz (ALLBUS 2016) auf die interessierenden Variablen reduziert. Daraus geht ein neuer, reduzierter Datensatz hervor: DS1. Auf den Variablen in diesem Datensatz werden dann etwaige Kodierungen vorgenommen, um die Variablen zu bilden, die für das SE-Modell notwendig sind. Gleichzeitig werden dabei auch Skalen gedreht und es werden die Skalierungen der einzelnen Variablen angepasst. Nachdem alle Modellvariablen entsprechend vorliegen, werden alle Fälle mit Missing Values entfernt, wodurch ein auf die Modellvariablen reduzierter Datensatz vorliegt, der für alle Fälle gültige Werte aufweist (DS2). Per Zufallsziehung wird dieser Datensatz dann auf 250 Fälle reduziert, da die MC-Ergebnisse zeigen, dass vor allem eine kleine Fallzahl problematisch für die MDTs ist: Dieser Datensatz wird als (Quasi-)Population definiert (PD). Infolgedessen werden aus diesem Datensatz Fälle entfernt, um den gewünschten Missinganteil zu erreichen. Mithilfe des R-Pakets *simsen* (Version 0.5-14) werden ca. 35 % der Fälle aus allen Indikatoren der latenten Konstrukte Ausländerfeindlichkeit und Antisemitismus sowie aus der Anomia-, der Deprivations-, der Bildungs- und der Einkommensvariablen gelöscht. Aufgrund der Tatsache, dass die gelöschten Fälle (bzw. die nun vorliegenden fehlenden Werte) durch die Ausprägungen in den Variablen Nationalstolz, Selbsteinschätzung der wirtschaftlichen Lage und

Schichteinstufung bedingt sind, unterliegen diese einem MAR-Ausfallmechanisums.[129] Damit liegt ein Datensatz (MD) vor, dessen Anteil an fehlenden Werten, dessen Ausfallmechanismus und dessen Fallzahl den Testbedingungen der Simulationen entsprechen.

Die blau hinterlegten Felder in der Abbildung zeigen den Prozess der Behandlung der fehlenden Werte. Grundlage dafür ist jeweils der Missingdatensatz (MD). Während der Prozess mit Direct-ML direkt in einem Schätzergebnis resultiert, gehen aus dem Imputationsprozess neue Datensätze hervor. Mit EM wird ein einziger Datensatz, mit den MI-Techniken, die, wie in Kapitel 6.4.2 beschrieben, eingesetzt werden, werden jeweils 50 Datensätze generiert (*m* wird auf 50 fixiert; siehe Kapitel 6.4.2). Weiterhin kann die Konvergenz der Verfahren mit den in Tabelle 7 gelisteten Kriterien gewährleistet werden.[130]

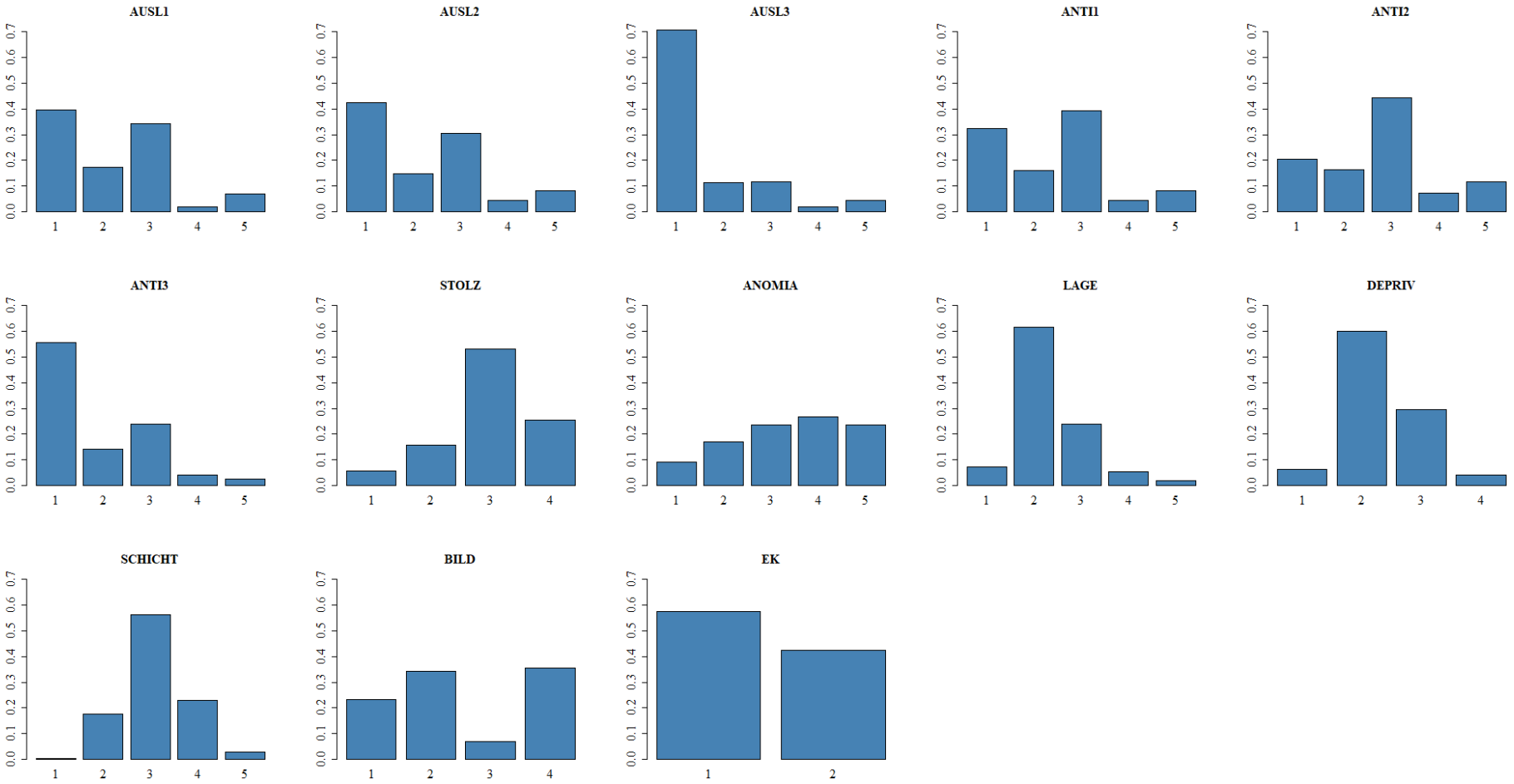
Die grau hinterlegten Felder zeigen den Analyseprozess. Die jeweilige Schätzung des SE-Modells erfolgt mit dem EDV-Programm M*plus* (Version 7.31) und dem MLR-Schätzer. Die Modellschätzung auf dem Populationsdatensatz dient in der Bewertung der MDTs als Referenzmodell und liefert die Populationsparameter; die Modellschätzung auf dem Missingdatensatz liefert das Ergebnis für Direct-ML; die Schätzungen auf den imputierten Daten liefern die Ergebnisse für EM, EMB, FCS, H0, MNV und PMM.

11.1.2 Deskription

Abbildung 18 zeigt die Verteilungen der Variablen des Populationsdatensatzes.[131] Die befragten Personen stimmen den Aussagen der einzelnen Indikatoren zu den latenten Konstrukten Ausländerfeindlichkeit und Antisemitismus eher weniger zu oder ordnen sich auf der Mittelkategorie ein. Zudem sind die befragten Personen eher stolz darauf, Deutsch zu sein. Im Durchschnitt neigen sie eher zur Anomia, schätzen ihre eigene wirtschaftliche Lage als gut ein, finden dass sie einen gerechten Anteil am Lebensstandard erhalten und ordnen sich der Mittelschicht zu. Zudem liegt ein mittleres Bildungsniveau vor und es finden sich eher befragte Personen im Datensatz wieder, die ein Haushaltsnettoeinkommen von unter 2500 Euro aufweisen.

[129] Die Variable Nationalstolz dient für die Missing Values auf den Variablen DEPRIV, AUSL1, ANTI1 als verursachende Variable; die Schichteinstufung verursacht Missings auf den Variablen EK, AUSL2, ANTI2; die LAGE-Variable auf den Variablen BILD, ANOMIA, AUSL3 und ANTI3.
[130] Im Anhang O2.8 finden sich die Konvergenzplots für EM, EMB, FCS, MNV und PMM. Die Konvergenz von Direct-ML und H0 (der Wert für den PSR) kann den jeweiligen Ausgabedateien entnommen werden.
[131] Da es zum Datensatz mit Missings keine Unterschiede gibt, sind nur die Verteilungen des vollständigen Datensatzes abgetragen.

Abbildung 18: Häufigkeitsverteilungen der empirischen Daten



Sowohl in Bezug auf den Grad der Asymmetrie, also welcher Anteil an Fällen sich in den jeweiligen Kategorien befindet, als auch in der Richtung der Schiefe, unterscheiden sich die vorliegenden Verteilungen von denjenigen, die simuliert wurden (dort wurden nur rechtsschiefe Verteilungen berücksichtigt). Von den Variablen selbst, kann nur die Schichtvariable als annähernd normalverteilt gelten. Alle anderen weichen von dieser Annahme ab, auch wenn die Werte für Skewness und Kurtosis für alle Variablen eher keine Verletzung der Normalverteilungsannahme anzeigen.[132] Der Anteil an Missing Values reicht auf den dafür vorgesehenen Variablen von 28,4 % bis 36 %.

11.2 Modellschätzungen und Ergebnisse

Wie bereits angeführt, stellt die Schätzung auf dem Datensatz ohne fehlende Werte die Populationsschätzung dar. Wenn nun die einzelnen MDTs die Missing Values zufriedenstellend handhaben können, dann sollten sich die Schätzungen für die Parameter, die Standardfehler und die Fit-Indices zwischen der Populationsschätzung und den Modellschätzungen der MDTs nur minimal unterscheiden. Die inhaltlichen Schlussfolgerungen sollten gleich sein. Die Ergebnisse aller Modellschätzungen finden sich in Tabelle 23.[133]

Zunächst kann an den Fit-Indices abgelesen werden, dass das Populationsmodell eine gute Passung an die Daten aufweist. Alle Fit-Indices liegen über oder unter den üblichen Grenzwerten und auch der Chi²-Wert ist nicht signifikant. Zudem lassen sich für das Populationsmodell die erwarteten Einflüsse beobachten: Alle Effekte sind positiv, wenngleich sie nicht alle signifikant sind. Das bedeutet, dass mit einer geringeren Schichteinstufung, mit einer höheren Deprivation und einer schlechteren Einschätzung der eigenen wirtschaftlichen Lage auch das Ausmaß der Anomia zunimmt. Zudem liegt ein positiver Zusammenhang zwischen der Anomia und dem Ethnozentrismus vor. Auch entsprechen die Effektrichtungen der Bildungs- und Einkommensvariablen denjenigen des Herrmann-Modells. Fraglich ist nun, ob sich diese Ergebnisse auch nach dem Einsatz der einzelnen MDTs zeigen.

Wie erwartet, liegen für die Performanz der MDTs im Hinblick auf die Fit-Indices zwei differierende Ergebnisse vor. Denn mit Direct-ML und H0 gehen korrekte Modellbewertungen einher, wohingegen diese nach dem Einsatz der anderen MDTs fehlerbehaftet sind. Während mit H0 die Werte der Fit-Indices in etwa denjenigen des Populationsmodells entsprechen, werden diese mit Direct-ML etwas schlechter. Allerdings weisen die Fit-Indices auch mit Direct-

[132] Siehe dazu die deskriptiven Statistiken im Anhang A7.
[133] Auf die Darstellung der Varianzen und Fehlervarianzen, sowie der standardisierten Schätzung wird verzichtet.

ML keine derartigen Abweichungen auf, als dass dies zu anderen inhaltlichen Schlüssen führt. Mit beiden MDTs wird weder der Chi²-Wert signifikant, noch werden die Grenzwerte der anderen Fit-Indices über- respektive unterschritten. Für das 90 %ige Konfidenzintervall von RMSEA zeigt sich zudem, dass die untere Intervallgrenze den Wert .05 miteinschließt (Akzeptanz der *close-fit* Hypothese) und die obere Intervallgrenze den Wert von .10 nicht überschreitet (Zurückweisung der *not-close-fit* Hypothese). Letztlich führt der Einsatz von Direct-ML und H0 zu einer korrekten Modellbewertung; alle Fit-Indices attestieren dem Modell eine gute Passung an die Daten.

Im Gegensatz dazu, liegen für die anderen MDTs weniger gute Ergebnisse vor. So ist der Chi²-Wert nach der Imputation der fehlenden Werte hoch signifikant. Das führt zu einer Zurückweisung der Nullhypothese, wonach die geschätzte mit der beobachteten Kovarianzmatrix übereinstimmt. Gleichzeitig liegen auch Abweichungen in den Schätzwerten für RMSEA (inklusive dessen Konfidenzintervall) und für den CFI im Vergleich zu den Populationswerten vor. In allen Fällen über- bzw. unterschreiten RMSEA und der CFI die akzeptierten Grenzwerte von .05 und .95. Gleichzeitig schließt auch die untere Intervallgrenze des 90 %igen Konfidenzintervalls von RMSEA den Wert .05 nicht mit ein (Verwerfung der *close-fit* Hypothese) und der Schätzwert der oberen Intervallgrenze überschreitet den festgelegten Grenzwert von .10 (Akzeptanz der *not-close-fit* Hypothese). Lediglich das SRMR spricht für die Akzeptanz des Modells. Das gilt allerdings nur dann, wenn der Grenzwert von .08 verwendet wird. Werden in diesem Fall, wie es in der Empirie üblich ist, alle Fit-Indices zusammengenommen betrachtet, dann rechtfertigen der signifikante Chi²-Wert[134], die schlechten Werte für den CFI, für RMSEA und dessen Konfidenzintervall, die Akzeptanz des Modells nicht. In diesem Fall führt der Einsatz von EM, EMB, FCS, MNV und PMM zu einem anderen inhaltlichen Schluss; denn anstatt das Modell anzunehmen, ist es zu verwerfen.

[134] Für die empirische SEM-Forschung ist ein signifikanter Chi²-Wert meist wenig bedeutsam, denn oftmals wird dieser zwar konstatiert, aber eben auch ignoriert (weil bspw. in einem größeren Sample meist ein signifikanter Wert vorliegt). Ignoriert wird er allerdings nur dann, wenn die anderen Bewertungsparameter einen akzeptablen oder guten Modellfit anzeigen. Letztendlich wiegen damit das schlechte Abschneiden von RMSEA und dessen Konfidenzintervall sowie das des CFIs für die angewandte Forschung etwas schwerer.

Tabelle 23: Modellschätzungen vor und nach Einsatz der MDTs

		MLR	Direct-ML	EM	EMB	FCS	H0	MNV	PMM
Fit-Indices									
Chi²-Statistik[135]		54.263	58.596	239.925***	256.624***	209.468***	56.415	210.686***	183.323***
RMSEA (90 % KI)		0.016 (0.000 – 0.044)	0.024 (0.000 – 0.049)	0.122 (0.106 – 0.137)	0.126 (0.111 – 0.142)	0.111 (0.095 – 0.127)	0.019 (0.001 – 0.045)	0.111 (0.096 – 0.127)	0.101 (0.086 – 0.117)
CFI		0.995	0.982	0.803	0.800	0.826	0.991	0.830	0.849
SRMR		0.045	0.056	0.063	0.070	0.069	0.041	0.066	0.066
Faktorladungen[136]									
ETHNO >	ANTI	0.991 (0.164)***[137]	0.923 (0.205)***	0.997 (0.157)***	0.868 (0.223)***	0.904 (0.214)***	0.908 (0.186)***	0.737 (0.140)***	0.890 (0.206)***
	STOLZ	0.126 (0.098)	0.112 (0.103)	0.079 (0.105)	0.117 (0.105)	0.118 (0.110)	0.106 (0.099)	0.110 (0.099)	0.124 (0.109)
AUSL >	AUSL2	0.888 (0.136)***	1.027 (0.189)***	1.066 (0.133)***	1.019 (0.216)***	1.048 (0.205)***	1.033 (0.174)***	0.998 (0.190)***	1.019 (0.195)***
	AUSL3	0.788 (0.098)***	0.714 (0.146)***	0.626 (0.103)***	0.702 (0.181)***	0.715 (0.164)***	0.710 (0.145)***	0.685 (0.151)***	0.739 (0.182)***
ANTI >	ANTI2	0.949 (0.106)***	0.940 (0.109)***	0.893 (0.087)***	1.026 (0.150)***	1.056 (0.168)***	0.938 (0.126)***	0.992 (0.147)***	1.004 (0.138)***
	ANTI3	0.773 (0.109)***	0.692 (0.135)***	0.676 (0.088)***	0.779 (0.149)***	0.744 (0.138)***	0.704 (0.121)***	0.737 (0.140)***	0.720 (0.117)***
Kovarianz									
BILD – EK		0.100 (0.036)**	0.093 (0.047)*	0.090 (0.037)*	0.102 (0.050)*	0.088 (0.048)$^{+}$	0.101 (0.053)$^{+}$	0.099 (0.047)*	0.110 (0.055)*
ausgewählte Strukturpfade[138]									
ANOMIA >	ETHNO[139]	0.251 (0.050)***	0.258 (0.057)***	0.199 (0.050)***	0.266 (0.065)***	0.250 (0.065)***	0.267 (0.058)***	0.262 (0.061)***	0.227 (0.066)**
LAGE >	ANOMIA	0.199 (0.113)$^{+}$	0.258 (0.169)	0.287 (0.124)*	0.293 (0.151)$^{+}$	0.294 (0.160)$^{+}$	0.256 (0.163)	0.303 (0.161)$^{+}$	0.238 (0.140)$^{+}$
DEPRIV >		0.422 (0.122)**	0.330 (0.160)*	0.230 (0.122)$^{+}$	0.305 (0.174)$^{+}$	0.207 (0.165)	0.335 (0.196)$^{+}$	0.258 (0.167)	0.249 (0.171)
SCHICHT >		0.100 (0.137)	0.014 (0.162)	0.006 (0.150)	-0.007 (0.163)	0.026 (0.170)	0.042 (0.164)	0.013 (0.162)	0.038 (0.168)
EK >		0.023 (0.153)	-0.232 (0.222)	-0.366 (0.146)*	-0.213 (0.224)	-0.217 (0.222)	-0.197 (0.221)	-0.192 (0.228)	-0.228 (0.234)
EK >	DEPRIV	-0.187 (0.077)*	-0.146 (0.107)	-0.137 (0.076)$^{+}$	-0.156 (0.104)	-0.119 (0.114)	-0.142 (0.113)	-0.121 (0.106)	-0.155 (0.101)
BILD >	LAGE	-0.071 (0.038)$^{+}$	-0.073 (0.047)	-0.053 (0.041)	-0.065 (0.046)	-0.076 (0.043)$^{+}$	-0.071 (0.047)	-0.072 (0.045)	-0.065 (0.046)

Anmerkungen: grau: andere inhaltliche Schlussfolgerungen zum Populationsmodell. Blau: andere Effektrichtung im Vergleich zum Populationsmodell allerdings n.s.; $^{+}$p<0.1; *p<0.05; **p<0.01; ***p<0.001.

[135] Freiheitsgrade: 51.

[136] Die Faktorladungen von AUSL1, ANTI1 und STOLZ sowie die Faktorladungen zwischen ETHNO und AUSL werden aus Identifikationsgründen auf 1.0 fixiert.

[137] Es handelt sich um unstandardisierte Koeffizienten; in der Klammer: Standardfehler.

[138] Für die nicht dargestellten Strukturpfade ergeben sich keine Unterschiede zum Populationsmodell. Die gesamten Schätzergebnisse lassen sich den Ausgabedateien der Modellschätzungen entnehmen (Anhang O2.8).

[139] Auf die Darstellung der indirekten und totalen Effekte wird verzichtet.

11.2.1 Ergebnisse zu den Parametern und den Standardfehlern

Im Hinblick auf die Parameterschätzungen lassen sich zwischen der Schätzung auf dem Populationsdatensatz und den Schätzungen nach der Behandlung der fehlenden Werte nur geringe Unterschiede ausmachen. Es werden alle Faktorladungen und die Kovarianz mit allen MDTs korrekt wiedergegeben. Sowohl die Effektrichtung als auch die Signifikanz der Effekte lassen sich mit den MDTs replizieren. Für die Strukturpfade können dagegen kleinere Unterschiede beobachtet werden. Zunächst sei auf den blau hinterlegten Pfad zwischen dem Einkommen und Anomia verwiesen. Für alle MDTs dreht sich hierbei das Vorzeichen im Vergleich zum Populationseffekt; der Effekt bleibt aber nicht signifikant (außer für EM). Da dieser Pfad im Populationsmodell im Hinblick auf die Effektstärke, gemessen an den standardisierten Koeffizienten, wenig bedeutsam ist und für die MDT-Schätzungen wenig bedeutend bleibt, liegt für diesen Pfad im Vergleich zum Populationsmodell auch eher kein sich widersprechendes Ergebnis vor (ähnliches gilt für den Pfad zwischen der Schichteinstufung und Anomia für FCS). Das liegt auch daran, weil es sich um einen Pfad einer Kontrollvariable handelt.

Im Gegensatz dazu, zeigen die grau hinterlegten Felder nicht signifikante Effekte an, die ursprünglich signifikant waren. Das wiederum führt zu anderen inhaltlichen Schlüssen im Vergleich zum Populationsmodell und hätte Auswirkungen auf die Hypothesentests. In der Regel können für die einzelnen Pfade etwas größere Standardfehler beobachtet werden, als im Populationsmodell, sodass vor allem Pfade von diesem Phänomen betroffen sind, die im Populationsmodell nur mit 5 %iger oder 10 %iger Irrtumswahrscheinlichkeit signifikant sind. Für die Pfade zwischen dem Einkommen und der Deprivation (für Direct-ML, EMB, FCS, H0, MNV und PMM) sowie zwischen der Bildung und der Einschätzung der wirtschaftlichen Lage (für Direct-ML, EM, EMB, H0, MNV und PMM) erscheint dieses nicht signifikante Ergebnis für die inhaltliche Bedeutsamkeit aber weniger problematisch, da es sich wieder um Pfade der Kontrollvariablen handelt. Dagegen ist der nicht mehr signifikante Pfad zwischen der Einschätzung der wirtschaftlichen Lage und Anomia nach Einsatz von Direct-ML und H0 von größerer Relevanz. Allerdings muss hierbei beachtet werden, dass dieser Pfad im Populationsmodell lediglich mit 10 %iger Irrtumswahrscheinlichkeit signifikant ist. Grundsätzlich gilt für Direct-ML und die MI-Techniken, dass die inferenzstatistischen Schlüsse etwas konservativer ausfallen sollen, weil dadurch verhindert wird, dass, aufgrund der Unsicherheit, die mit den fehlenden Werten einhergeht, knapp nicht signifikante Ergebnisse nach dem MDT-Einsatz signifikant werden. Demzufolge ist auch dieses Ergebnis als eher unproblematisch zu werten, zumal das 10 %ige Signifikanzniveau in der angewandten Forschung meist nur als Indiz gewertet wird

und die Effektrichtung als auch deren Stärke mit Direct-ML und H0 repliziert werden können. Der Pfad zwischen der Deprivationsvariablen und Anomia bildet hingegen eine Ausnahme. Hierbei liegt im Populationsmodell ein signifikanter Effekt mit 1 %iger Irrtumswahrscheinlichkeit vor, wohingegen der Pfad für FCS, MNV und PMM nicht mehr signifikant ist. Weil es sich zudem um einen Pfad handelt, der von inhaltlichem Interesse ist, liegt mit diesen MDTs ein Ergebnis vor, das im Widerspruch zum Populationsmodell steht, auch wenn es in Bezug auf die Effektstärke dieses Pfades zu keinen Unterschieden kommt.

Entgegen den Erwartungen liefert die Einfachimputation mit EM, weil nur zwei Felder farblich gekennzeichnet sind, in quantitativer Hinsicht die besten Ergebnisse. Während für die anderen MDTs aber das Problem besteht, dass signifikante Effekte nach deren Einsatz nicht mehr signifikant sind, kann für EM ein Pfad beobachtet werden der signifikant ist, aber nicht signifikant sein sollte: derjenige zwischen dem Einkommen und Anomia. Zudem weicht dieser Pfad auch in seiner Stärke erheblich vom Populationsmodell ab. Wie zuvor gilt, dass es sich dabei um einen Pfad einer Kontrollvariablen handelt. Allerdings könnte dieser aufgrund der, im Vergleich zu den anderen Pfaden auf Anomia, überschätzten Effektstärke überinterpretiert werden. Weiterhin können im vorliegenden Beispiel die Unterschätzungen der Standardfehler, die sich in den MC-Ergebnissen zeigen, eher nicht beobachtet werden. Stattdessen entsprechen in vielen Fällen die Ergebnisse mit EM denjenigen der anderen MDTs.

11.2.2 Zusammenfassung und Einschränkungen

Legt man dieses Beispiel als Gradmesser für die Performanz der verschiedenen MDTs an, dann lassen sich daraus einige Erkenntnisse ableiten. Diese sollten allerdings nicht überbewertet werden, da die Aussagekraft dieses Beispiel, in Bezug auf die Empirie und hinsichtlich der Performanz der MDTs beschränkt ist (siehe dazu die Ausführungen weiter unten):

- Mit Direct-ML und H0 ist eine korrekte Modellbewertung durch alle Fit-Indices sichergestellt. Für die anderen MDTs erfolgt eine korrekte Bewertung des Modells nur durch das SRMR, die anderen Fit-Indices (der Chi^2-Wert, RMSEA, dessen Konfidenzintervall und der CFI) führen dagegen zur falschen Ablehnung des Modells.
- Die Parameterschätzungen sind in der Regel unverzerrt und entsprechen, zwar nicht in der absoluten Höhe aber in ihrer Richtung, denjenigen des Populationsmodells.
- Oftmals entsprechen auch die inferenzstatistischen Schlüsse denjenigen des Populationsmodells. Es gilt aber zu beachten, dass dies in einigen Fällen nur dann gilt, wenn das 10 %ige Signifikanzniveau verwendet wird. Dementsprechend sollten nach dem Einsatz

der MDTs Effekte auch mit 10 %iger Irrtumswahrscheinlichkeit als substantiell betrachtet werden.

- Liegen im Populationsmodell signifikante Effekte auf dem 10 %igen und zum Teil auch auf dem 5 %igen Signifikanzniveau vor, dann ist die Wahrscheinlichkeit erhöht, dass diese nach dem Einsatz der MDTs nicht mehr signifikant sind. Es muss zusätzlich zur Einordnung der Bedeutsamkeit eines Effektes auch dessen Effektstärke berücksichtigt werden, denn auch diese wird in der Regel unverzerrt geschätzt.

Schlussendlich kann dieses Bespiel drei zentrale MC-Ergebnisse replizieren, was auch mit der Bestätigung der eingangs getätigten Erwartungen E1 bis E3 einhergeht. So werden mit den imputationsbasierten MDTs ohne Berücksichtigung der Modellstruktur die Fit-Indices nicht zufriedenstellend geschätzt, wohingegen mit Direct-ML und H0 korrekte Modellbewertungen erfolgen (E1). Die Parameterschätzungen sind dagegen zwischen den MDTs ähnlich und entsprechen der Populationsschätzung. Zudem zeigen sich für alle MDTs relativ ähnliche Ergebnisse im Hinblick auf die Signifikanz und Nicht-Signifikanz einzelner Effekte. Zwar unterscheiden sich einige Pfade nach dem Einsatz der MDTs zum Populationsmodell, allerdings unterscheiden sich die MDTs kaum voneinander. Das bedeutet, dass durch den Vorzug einer MDT das Ergebnis der Modellschätzung im Hinblick auf die Parameterschätzung und die Bedeutsamkeit einzelner Effekte eher nicht beeinflusst wird: Mit allen MDTs lassen sich gleiche Ergebnisse erwarten (E2 und E3). In einem Fall liegt für EM fälschlicherweise ein signifikanter Effekt vor. Dass EM aber im Hinblick auf die Signifikanz einzelner Effekte schlechter abschneidet, lässt sich nicht belegen. Dieses Ergebnis widerspricht den MC-Ergebnissen (E4).

Anzumerken ist, dass dieses Beispiel in seiner Aussagekraft beschränkt ist. Das liegt daran, dass es nur auf einer Stichprobe beruht. Somit liegt auch nur ein einziger Schätzwert für den jeweiligen Parameter, den Standardfehler oder die Fit-Indices vor. Würde man dieses Beispiel wiederholen, dann würde sich auch das vorliegende Ergebnis unterscheiden, denn die Ziehung der 250 Fälle aus dem Originaldatensatz, der Prozess des Einfügens der fehlenden Werte oder die Imputation derselben sind allesamt computergesteuerte Zufallsprozesse, welche von den zugewiesenen Startwerten abhängig sind. Wird bei einer Wiederholung dieser Exemplifikation ein anderer Startwert für den Zufallsprozess in nur einem dieser Schritte verwendet, dann wird sich auch das Ergebnis von dem vorliegenden unterscheiden. Aus diesem Grund ist es auch möglich, dass die Ergebnisse etwas von den MC-Ergebnissen abweichen. Denn dort können in den meisten Fällen nur Unterschätzungen der Parameter und Standardfehler beobachtet werden,

wohingegen es bei dieser Anwendung auch zu Überschätzungen der Parameter und Standardfehler gekommen ist. Genau deshalb ist es in einer MC-Studie notwendig, möglichst viele Stichproben zu ziehen. Denn damit geht eine empirische Verteilung der Schätzwerte einher, womit die Berechnung von Durchschnittswerten möglich wird, die durch die große Anzahl an Stichproben gegenüber möglichen Ausreißern einer einzigen Stichprobe stabil sind. Die Abweichungen in den Schätzergebnissen dieses Beispiels zu den MC-Ergebnissen, könnten aber auch daher stammen, dass ein anderes Modell geschätzt wird, dass dieses mehr Variablen beinhaltet, die zudem noch jeweils unterschiedliche Werteverteilungen haben, oder dass die Zusammenhänge nicht denen der MC-Modelle entsprechen.

Weil nicht sichergestellt werden kann, ob die Unterschiede in den Schätzergebnissen auf Ausreißer zurückzuführen sind (weil nur eine Stichprobe vorliegt) oder ob es sich um systematische Abweichungen von den MC-Ergebnissen handelt, die auf das Beispielmodell, die Anzahl und die Verteilungen der Variablen und/oder deren Zusammenhänge zurückzuführen sind, sollten die Ergebnisse und die darauf basierenden Schlussfolgerungen nicht auf andere Sachverhalte übertragen werden. Schlussendlich liegt der Mehrwert dieses Kapitels darin, dass es einen Hinweis liefert, wonach die MC-Ergebnisse auch auf die Empirie und auf andere Modelle übertragen werden können. Denn es bestätigt die Tendenzen in den MC-Ergebnissen, wonach die Modellbewertung mit den Fit-Indices abhängig von der gewählten MDT ist, und wonach die Schätzungen der Parameter und der Standardfehler mit allen MDTs relativ ähnlich sind.

12 Zusammenfassung der Arbeit, Einschränkungen und Ausblick

In der vorliegenden Arbeit sollten verschiedene Konfigurationen von FCS unter möglichst differenzierten, empirienahen Gegebenheiten und im Vergleich zu anderen, bereits häufiger evaluierten MDTs, wie JM und Direct-ML, im SEM-Kontext mittels MC-Simulationen untersucht werden. Die herausgearbeiteten Desiderate zeigten, dass für FCS in diesem Analyserahmen kaum Forschung existiert, aber auch, dass die Performanz vieler anderer MI-Techniken darin bisher kaum untersucht wurde. Dementsprechend lagen auch kaum Erkenntnisse vor, wie gut oder schlecht die MI-Techniken im Hinblick auf die Fit-Indices zu bewerten sind. Durch die klare Festlegung der Forschungslücke konnte ein Forschungsdesign aufgestellt werden, das eine breit gestreute Evaluation der MI im Allgemeinen und FCS im Besonderen ermöglichte. Dadurch konnte sichergestellt werden, dass die ausgegebenen Ziele auch allesamt erreicht werden konnten. Denn neben der Performanzevaluation von FCS im Hinblick auf die Fit-Indices (Ziel 1; siehe Kapitel 1.1), fand eine Evaluation von FCS unter möglichst differenzierten, empirienahen Rahmenbedingungen, mit unterschiedlichen Konfigurationen (Ziel 2 und Ziel 3) und

im Vergleich zu anderen MI-Techniken und zu alternativen ML-Verfahren statt (Ziel 4). Mit den in diesem Abschlusskapitel abzuleitenden Handlungsempfehlungen, kann dann auch das letzte Ziel dieser Arbeit erreicht werden (Ziel 5).

Diese Zusammenfassung der Arbeit nimmt zunächst Bezug auf die Desiderate, die in dieser Arbeit bearbeitet wurden, bevor die Ergebnisse hinsichtlich der aufgestellten Forschungsfragen diskutiert werden, um daraus Handlungsempfehlungen für die empirische Praxis abzuleiten. Danach folgen Bemerkungen zur Übertragbarkeit der vorliegenden Ergebnisse sowie ein Ausblick für mögliche Folgeforschungen.

12.1 Bearbeitete Forschungslücke

Um Ergebnisse für FCS unter möglichst differenzierten Bedingungen zu erhalten, wurden drei verschiedene SEM-Populationsmodelle (Modell 1 bis Modell 3) unter unterschiedlichen Konfigurationen mittels MC-Simulationen getestet (fünftes Desiderat). Zu den Simulationskonfigurationen gehörten, in Anlehnung an empirisch häufig vorzufindende Datenstrukturen, einerseits Variablen mit quasi-metrischem Skalenniveau und andererseits binär sowie ordinal skalierte Variablen. Damit konnte FCS unter der Bedingung getestet werden, dass erstens nur ein einheitliches Skalenniveau in den Daten vorlag und dass zweitens gemischte Daten gegeben waren (erstes und zweites Desiderat). Der erste Punkt erlaubte es, für FCS ein IM anzulegen, das nur ein Schätzverfahren zur Ersetzung der fehlenden Werte benötigte, was für die erste FCS-Spezifikation lineare Regressionen waren und für die zweite Spezifikation PMM. Die gemischten Daten machten es dagegen notwendig, FCS unter der Bedingung zu testen, dass während des Imputationsprozesses gleichzeitig verschiedene Schätzverfahren angewendet werden mussten, um die fehlenden Werte zu ersetzen. So wurde für den Imputationsprozess mittels FCS einerseits ein IM konstruiert, das binär und multinomial logistische Regressionen sowie lineare Regressionen zur Imputation der fehlenden Werte benutzte, und andererseits ein IM, das die fehlenden Werte auf den gemischten Daten mit PMM (viertes Desiderat) und damit mit einem einzigen Schätzverfahren ersetzte. Letztendlich wurden in dieser Arbeit damit insgesamt drei Variationen von FCS getestet. Das ist zum einen eine Konfiguration, in der nur lineare Regressionen berücksichtigt wurden (Modell 1) und zum anderen eine Konfiguration, in welcher FCS mittels verschiedener Schätzverfahren im IM spezifiziert wurde (mit binär und multinomial logistischen sowie linearen Regressionen; Modell 2 und 3). Zuletzt wurde noch PMM untersucht (einmal auf quasi-metrischen Daten und einmal bei gemischten Daten; Modell 1 respektive Modell 2 und 3).

Zusätzlich zu den verschiedenen Populationsmodellen wurde die Performanz von FCS bei zwei verschiedenen Fallzahlen, unter symmetrischen und (stark) asymmetrischen Verteilungen (drittes Desiderat) sowie unter verschieden hohen Anteilen an Missing Values unter einem MAR-Ausfallmechanisums erfasst. Neben den FCS-Spezifikationen wurden noch weitere MDTs berücksichtigt. Das waren die weiteren MI-Varianten EMB, H0 und MNV (sechstes Desiderat) sowie die ML-Schätzverfahren Direct-ML und EM (achtes Desiderat). Als Performanzkriterien dienten der vorliegenden Arbeit zum einen die Fit-Indices (der p-Wert der Chi²-Statistik, das SRMR, RMSEA und dessen Konfidenzintervall sowie der CFI; siebtes Desiderat) und damit die Bewertungsebene des Gesamtmodells und zum anderen die Schätzung der Parameter, die relative Effizienz der Parameterschätzung und die Schätzung der Standardfehler. Durch diese Bewertungskriterien konnte herausgearbeitet werden, ob sich die MI-Techniken in deren Performanz voneinander (neuntes Desiderat) und gegenüber den ML-Verfahren unterscheiden (zehntes Desiderat).

Diese Arbeit erweitert in den vier folgenden Punkten den Forschungsstand:

- Zum einen wurde in dieser Arbeit ein Vergleich verschiedener populären MDTs (ob ML-Verfahren oder MI-Techniken) in einem kontrollierten Rahmen unter möglichst differenzierten, empirienahen Bedingungen durchgeführt, wohingegen in anderen Studien nur wenige Vergleiche zwischen diesen MDTs getätigt wurden.
- Zum anderen wurde eine detaillierte Analyse von MI-Techniken (EMB, H0 und PMM) durchgeführt, die in der bisherigen Forschung nur wenig Berücksichtigung fanden.
- Drittens wurde mit dieser Arbeit zum ersten Mal eine detaillierte Analyse der Fit-Indices für verschiedene MI-Varianten durchgeführt. Andere Studien konzentrierten sich dabei auf die Performanz des p-Wertes der Chi²-Statistik, nicht aber auf andere populäre Fit-Indices. Zudem stand bei diesen Studien dann auch die Performanz von Direct-ML im Vordergrund und nur in wenigen Fällen die Performanz von EM oder MNV; andere Varianten der MI wurden nicht evaluiert.
- Zuletzt stellt diese Arbeit eine der wenigen MC-Studien dar, in denen die MDTs auf quasi-metrischen/gemischten Daten unter variierten Verteilungen getestet wurden. Solche Daten sind in der empirischen Praxis weit verbreitet, wurden bisher aber kaum untersucht.

Dementsprechend liegen mit dieser Arbeit erstmals Ergebnisse vor, die zeigen, dass sich verschiedene MI-Techniken unter realitätsnahen Bedingungen kaum voneinander unterscheiden

und dass es, zumindest in Bezug auf die Parameterschätzung, deren Effizienz und die Schätzung der Standardfehler kaum einen Unterschied macht, ob MI-Techniken oder ML-Verfahren verwendet werden. Zudem liefert diese Arbeit wichtige Hinweise für die empirisch Forschenden im SEM-Kontext. Denn zum ersten Mal liegen Ergebnisse für viele praxisrelevante MDTs im Hinblick auf Bewertungskriterien vor, die bisher unberücksichtigt geblieben sind: die Fit-Indices. Das ist umso wichtiger, als dass diese einen wesentlichen Bestandteil der Modellanalyse darstellen. Beides verdeutlicht die Relevanz der vorliegenden Ergebnisse.

12.2 Beantwortung der Forschungsfragen und Handlungsempfehlungen

Nachdem die vorausgehenden Kapitel zeigen, dass die herausgearbeiteten Desiderate durch das gewählte Forschungsdesign mit der vorliegenden Arbeit bearbeitet werden konnten und dass die ausgegebenen Ziele erreicht wurden, stellt sich nun die Frage, wie die Ergebnisse im Hinblick auf die Forschungsfragen zu bewerten sind. Die Beantwortung der Fragen geht mit Handlungsempfehlungen für die Praxis einher.[140]

1. Eignet sich FCS im Rahmen der Strukturgleichungsmodellierung als Alternative zur Ersetzung von fehlenden Werten?

Integraler Bestandteil bei der Ergebnisanalyse mit SE-Modellen ist die Bewertung der Übereinstimmung zwischen der, durch das spezifizierte Modell implizierten Kovarianzmatrix mit der empirisch beobachteten Kovarianzmatrix. Kann dabei eine gute Passung nachgewiesen werden, dann können in einem Folgeschritt auch die Parameter und die Standardfehler dieser Modelle interpretiert werden. Mit allen getesteten FCS-Spezifikationen besteht nun das Problem, dass die Fit-Indices, die zur Bewertung des Modells herangezogen werden, um die Übereinstimmung des Modells mit den Daten anzuzeigen, in vielen Fällen nicht mehr zuverlässig sind. Das gilt, wenn die gängigen Grenzwerte angelegt werden, mit denen akzeptable Passungen von nicht-akzeptablen Passungen unterschieden werden.

Im Ergebnis zeigt sich, dass mit diesen Grenzwerten unter dem Einsatz der FCS-Spezifikationen in vielen Fällen zu hohe Ablehnungsraten mit dem p-Wert der Chi²-Statistik, mit RMSEA und dessen Konfidenzintervall oder dem CFI generiert werden. Das bedeutet, dass korrekte Modelle mittels dieser Fit-Indices fälschlicherweise zurückgewiesen werden. Nichtsdestotrotz lassen sich mit FCS auch korrekte Modellevaluationen erzielen. Diese sind aber von

[140] Diese verdichten die Ergebnisse der Arbeit, sodass an dieser Stelle auf die ausführlicheren Ergebnisdarstellungen in den Tabellen 9, 10, 16 und 21, verwiesen sei.

den zugrundeliegenden Eigenschaften der Daten und dem jeweils gewählten Fit-Index abhängig. So geht mit dem SRMR immer eine korrekte Modellbewertung einher, unabhängig davon wie viele Missing Values vorliegen oder wie groß das Sample ist. Mit dem p-Wert und RMSEA (sofern für RMSEA kleine Fallzahlen vorliegen) kann es aber bereits bei 5 % an Missing Values zu inkorrekten Schlussfolgerungen bei der Modellbewertung kommen. Ab 20 % an Missing Values liefern beide Indices keine zufriedenstellenden Ergebnisse mehr; eine korrekte Modellbewertung lassen diese dann nicht mehr zu. Anders dagegen das Konfidenzintervall von RMSEA und der CFI. Beide erbringen bei wenigen Missing Values bis hin zu 20 % an Missing Values zufriedenstellende Ergebnisse, für letzteres allerdings auch nur dann, wenn eine große Fallzahl gegeben ist. Bei sehr hohen Anteilen an Missing Values (35 %) lassen sich aber auch mit diesen Indices keine korrekten Modellbewertungen mehr erreichen. Unter Umständen kann dann mit dem CFI, sollte gleichzeitig eine größere Fallzahl vorliegen, noch eine korrekte Modellbewertung erzielt werden.

Die vorliegenden Ergebnisse zu den Fit-Indices zeigen, dass die MI mit FCS eine Alternative zur Ersetzung der fehlenden Werte im Zusammenhang mit SEM sein kann, unabhängig davon, welche Konfiguration für FCS gewählt wird. Aber sie zeigen auch, dass es nach Einsatz von FCS zu fehlerbehafteten Modellbewertungen kommen kann, da korrekt spezifizierte Modelle in vielen Fällen zurückgewiesen werden. Grundsätzlich erlauben es die Ergebnisse aber nicht, von FCS als Alternative zur Behandlung der fehlenden Werte im Bereich von SEM abzuraten. Dazu liegen für das SRMR und, bei großen Fallzahlen, für das Konfidenzintervall von RMSEA und für den CFI ausreichend gute Ergebnisse vor, damit FCS in vielen Fällen auch in der empirischen Praxis eingesetzt werden kann. Nicht zuletzt auch deshalb, weil mit einem Missinganteil von 35 % pro Variable ein sehr hoher Anteil an fehlenden Werten berücksichtigt wurde, der in der empirischen Praxis wohl über den gängigen Anteilswerten liegt. Unabdingbar für die Praxis ist allerdings, verschiedene Fit-Indices zu betrachten, das SRMR zu konsultieren und widersprüchliche Ergebnisse der Fit-Indices in Relation zu den analytischen Fundierungen zu setzen. Denn aufgrund der statistischen Kennwerte der Fit-Indices ist die Wahrscheinlichkeit, ein korrekt spezifiziertes Modell nach Behandlung der fehlenden Werte mit FCS fälschlicherweise zurückzuweisen, erhöht.

2. Sind mit FCS unter möglichst realitätsnahen und empirisch vorzufindenden Bedingungen plausible Imputationen zu erwarten?
3. Welche der möglichen FCS-Spezifikationen ist unter möglichst vielen Bedingungen am robustesten und liefert die besten Ergebnisse?

Liegen plausible Imputationen vor, dann sind die Schätzergebnisse der Analysemethoden (in diesem Fall sind es SE-Modelle) im Hinblick auf die Parameter und die Standardfehler unverzerrt. In einem solchen Fall lassen sich dann auch korrekte inferenzstatistische Schlüsse erwarten. Während die Imputationen mit FCS dazu führen, dass die Bewertung der Modelle durch die Fit-Indices problembehaftet ist, lässt sich ein problematisches Ergebnis für die Parameter und die Standardfehler nicht beobachten, denn sowohl die Parameter als auch die Standardfehler der Modellschätzungen sind in aller Regel unverzerrt. Zukünftige Forschung sollte aber folgendes berücksichtigen: Die FCS-Spezifikation mit den logistischen Regressionsmodellen zur Ersetzung der fehlenden Werte in den diskreten Variablen (Modell 2 und 3) liefern in wenigen Einzelfällen für Kovarianzen stärker überschätzte Parameter und unterschätzte Standardfehler, als PMM. Denn mit PMM gehen in keinem Fall stark über- oder unterschätzte Parameter oder Standardfehler für die Kovarianzen einher: PMM scheint sich also besser zu eignen, um fehlende Werte auf binär und ordinal skalierten Variablen zu ersetzen. Allerdings sollte beachtet werden, dass die Über- und Unterschreitungen bei FCS nur in wenigen Einzelfällen vorliegen. Hierbei von einer Systematik zu sprechen, erlauben die Ergebnisse nicht. Letztendlich kann nicht ausgeschlossen werden, dass es sich bei diesen Einzelfällen um Zufallsergebnisse aufgrund des computergesteuerten Zufallsprozesses handelt. Insgesamt kann an dieser Stelle deshalb nicht von den logistischen Regressionsmodellen abgeraten werden, denn alle Modellkomponenten werden in der Regel unverzerrt geschätzt. Auch kann nicht zu PMM geraten werden, weil für PMM keine weiteren MC-Ergebnisse vorliegen, die zeigen, wie sich diese Schätzmethode für Missing Values bei diskreten Variablen verhält.

Letzten Endes können mit FCS, unabhängig von der gewählten Konfiguration, also zumeist gute Ergebnisse für die Parameterschätzungen und deren Effizienz, als auch für die Schätzung der Standardfehler erzielt werden. Es können auch keine Ergebnisse beobachtet werden, welche eine FCS-Spezifikation gegenüber einer anderen bevorteilen würde. Infolgedessen kann davon ausgegangen werden, dass die MI mit FCS auch bei einer kleinen Fallzahl, stark asymmetrischen Verteilungen und sehr hohen Missinganteilen plausible Imputationen generiert, sodass damit korrekte inferenzstatistische Schlüsse möglich werden und die inhaltliche Interpretation der Modellschätzung durch diese MDT nicht verfälscht wird – sofern es nur um die Parameterebene geht. Damit eignet sich die MI mit FCS im Kontext der Analyse von SE-Modellen und unter einem MAR-Ausfallmechanisums, unabhängig davon, welche Konfiguration des Verfahrens zur Imputation der fehlenden Werte gewählt wird.

4. Wie gut schneidet FCS im Vergleich zu den konkurrierenden Verfahren JM und Direct-ML ab, da sich beide Verfahren unter verschiedensten Simulationsbedingungen zumindest als gleichwertig, wenn nicht sogar besser erwiesen als FCS?

Neben den angesprochenen FCS-Spezifikationen wurden in dieser Arbeit noch Direct-ML sowie JM in zwei verschiedenen Varianten untersucht: EMB und MNV. Aufgrund dessen, dass bisher keine Regeln vorliegen, wie die Fit-Indices beim Einsatz der MI zusammengeführt werden sollen, wurde auch eine Einfachimputation mit EM untersucht, um zu prüfen, ob es problematisch ist, den Durchschnittswert der Fit-Indices zu bilden. Zudem wurde H0 in das Forschungsdesign integriert, da dabei die Modellstruktur des Analysemodells bereits während des Imputationsprozesses berücksichtigt wird und sie deshalb eine, für Analysen mittels SE-Modellen, interessante Variante der MI darstellt.

Beide MDTs, welche die Modellstruktur bei der Behandlung der fehlenden Werte berücksichtigen (Direct-ML und H0) sind in Bezug auf die Fit-Indices allen anderen getesteten MDTs überlegen. Sie liefern akzeptable Ablehnungsraten und damit zufriedenstellende Werte für die Fit-Indices bei der Modellevaluation. Es liegt zudem keine Kombination an Dateneigenschaften vor, die für diese MDTs problematisch ist. Damit ist die Performanz der beiden unabhängig von den zugrundeliegenden Daten. Wenn demnach Missing Values vorliegen, dann kann nach deren Behandlung mit Direct-ML oder H0 davon ausgegangen werden, dass die geschätzten SE-Modelle korrekt beurteilt werden, sofern diese korrekt spezifiziert wurden. Für die MI-Varianten EMB und MNV gelten hingegen dieselben Ausführungen wie für FCS, da durch den Einsatz die Modellbewertung problembehaftet wird. Die fehlerbehaftete Modellevaluation ist allerdings nicht auf die Bildung des Durchschnittswertes der *m* Fit-Indices zurückzuführen, denn auch für EM können keine befriedigenden Ergebnisse beobachtet werden. Das Fehlen spezieller Regeln zur Zusammenführung der *m* Fit-Indices, wie sie für die Parameterschätzwerte und die Standardfehler bei der MI existieren (*Rubin´s rules*), ist demnach nicht das vordergründige Problem bei der Modellbewertung. Ansonsten könnten bei der Einfachimputation zufriedenstellende Bewertungen beobachtet werden. Stattdessen erscheinen die festgelegten Grenzwerte für die Fit-Indices, die für vollständige Daten vorgeschlagen sind, für imputierte Daten zu strikt zu sein. Im Hinblick auf die Fit-Indices sind FCS und auch JM mittels EMB und MNV sowie die Einfachimputation mit EM, weder mit Direct-ML noch mit H0 gleichrangig. In diesem Fall liegen mit Direct-ML und H0 vorzuziehende MDTs vor.

Im Gegensatz zu den Ergebnissen der Modellebene, liegen nicht nur für Direct-ML und H0 in Bezug auf die Schätzung der Parameter und der Standardfehler zufriedenstellende Ergebnisse

vor, sondern auch für JM. Zudem gibt es zwischen den einzelnen MI-Varianten und Direct-ML nur triviale Unterschiede: Unter fast allen Konfigurationen können mit diesen MDTs nahezu gleiche Ergebnisse beobachtet werden. Sie alle erzielen im Hinblick auf die Schätzung der Parameter, deren relative Effizienz und die Schätzung der Standardfehler gute Ergebnisse, sodass alle diese MDTs uneingeschränkt empfohlen werden können. Dementsprechend gibt es keinen Grund, eine dieser MDTs den anderen, und im speziellen FCS, für die Schätzung der Parameter und der Standardfehler, vorzuziehen.

Neben den MI-Varianten und Direct-ML eignet sich auch EM als Imputationsvariante für die Parameterschätzung, obwohl diese im Vergleich zu den anderen MDTs etwas weniger effizient ist und die Parameterschätzwerte eine etwas größere Variabilität aufweisen. Da dies aber nicht zu inakzeptablen Schätzwerten führt, kann auch beim Einsatz von EM von unverzerrten Parameterschätzwerten ausgegangen werden. Problematisch bei EM ist allerdings, dass damit negativ verzerrte Standardfehler einhergehen und ab 20 % an Missing Values kein Standardfehler mehr geschätzt wird, der als unverzerrt einzuordnen ist. Wird EM demnach bei erhöhten Missinganteilen eingesetzt, können signifikante Effekte aufgrund der ersetzten Werte entstehen, da der Standardfehler in der Regel zu klein, und bei sehr hohen Anteilen an fehlenden Werten nahezu halbiert wird. EM sollte deshalb eher nicht und wenn, dann nur bei geringen Missinganteilen verwendet werden.

Schlussendlich gibt es keine Vor- oder Nachteile von FCS gegenüber den anderen getesteten MDTs, wenn nur die Schätzung der Parameter und der Standardfehler von Interesse ist (unabhängig davon, wie FCS konfiguriert wird). In aller Regel werden mit allen getesteten MDTs unverzerrte Parameter und Standardfehler (mit Ausnahme für EM) geschätzt, sodass der Einsatz unproblematisch ist. Für die empirische Praxis ist das eine gute Nachricht, denn die Auswahl einer MDT entscheidet nicht über das inhaltliche Ergebnis der späteren Modellschätzung (mit Ausnahme von EM). Allerdings lassen sich eben in Bezug auf die Fit-Indices zwei MDTs identifizieren, die besser abschneiden als alle anderen getesteten: Direct-ML und H0. Mit beiden gehen unter allen Bedingungen korrekte Modellevaluationen einher, wohingegen eine solche bei den anderen MDTs von den zugrundeliegenden Daten und dem jeweils ausgewählten Fit-Index abhängig ist. Sollten demnach EM, EMB, FCS, MNV oder PMM im Kontext von SEM zum Einsatz kommen, empfiehlt es sich für die empirische Praxis, den nachfolgenden Handlungsempfehlungen zu folgen. Nicht nur reduziert deren Befolgung die Wahrscheinlichkeit ein korrektes Modell fälschlicherweise zurückzuweisen, sie fassen die Ergebnisse der Ar-

beit auch in der Hinsicht zusammen, als dass sie angeben, unter welchen empirisch vorzufindenden Bedingungen welche MDT zu wählen ist. Tabelle 24 dient den empirisch Forschenden als Orientierungshilfe.

In der Tabelle werden die Verteilungen der Variablen nicht berücksichtigt. Das liegt daran, dass diese in allen Meta-Modell-Analysen am wenigsten einflussreich sind. In allen Fällen kann entweder der Missinganteil oder die Samplegröße als entscheidend für die Performanz der MDTs ausgemacht werden. Die Verteilungen haben im besten Fall keine Auswirkungen auf die Wahrscheinlichkeit ein Modell zurückzuweisen oder auf die Verzerrungen in den Parametern und Standardfehlern und im ungünstigsten Fall werden die Wahrscheinlichkeiten der Modellablehnung und die Verzerrungen, die ohnehin aufgrund der Fallzahl oder der Missinganteile bereits vorliegen, nur vernachlässigbar verschärft. Weiterhin werden verschiedene Variablenskalierungen nicht aufgeführt, da es für alle MDTs keinen Unterschied macht, ob nur quasimetrisch skalierte Variablen vorliegen, oder ob gemischte Daten, mit unterschiedlich skalierten Variablen gegeben sind. Dementsprechend können gemischte Daten mit MDTs behandelt werden, welche die Skalenniveaus explizit berücksichtigen können (FCS oder H0), oder es können MDTs verwendet werden, welche aufgrund ihrer Verfahrenslogik für gemischte Daten eigentlich nicht geeignet sein sollten (Direct-ML, EM, EMB, MNV, PMM). Denn sowohl bei Daten mit nur quasi-metrisch skalierten Variablen, als auch bei Daten, die zusätzlich binär und ordinal skalierte Variablen aufweisen, gehen mit den MDTs in allen Fällen zufriedenstellende Ergebnisse einher.

Tabelle 24: Handlungsempfehlungen für die Praxis

		Kleine Fallzahl (N = 250)			Große Fallzahl (N = 750)		
		Wenige Missing Values (5 %)	Erhöhter Missing-anteil (20 %)	Sehr hoher Missing-anteil (35 %)	Wenige Missing Values (5 %)	Erhöhter Missing-anteil (20 %)	Sehr hoher Missing-anteil (35 %)
Modell-ebene	SRMR	alle MDTs	alle MDTs	alle MDTs	alle MDTs	alle MDTs	alle MDTs
	p-Wert	alle MDTs	Direct-ML, H0	Direct-ML, H0	alle MDTs	Direct-ML, H0	Direct-ML, H0
	RMSEA	Direct-ML, H0	Direct-ML, H0	Direct-ML, H0	alle MDTs	Direct-ML, H0	Direct-ML, H0
	RMSEA KI (90 %)	alle MDTs	Direct-ML, H0	Direct-ML, H0	alle MDTs	alle MDTs	Direct-ML, H0
	CFI	alle MDTs	Direct-ML, H0	Direct-ML, H0	alle MDTs	alle MDTs	alle MDTs
Parame-terebene	Unverzerrte Parameter	alle MDTs					
	Unverzerrte Standard-fehler	alle MDTs	alle außer EM		alle MDTs	alle außer EM	

12.3 Einschränkungen zur Übertragbarkeit der Ergebnisse

Sicherlich gibt es Grund zur Annahme, dass die Ergebnisse auch für andere Situationen Gültigkeit besitzen. Ein handfester Nachweis liegt aber nicht vor, weil die vorliegenden Ergebnisse vom gewählten Forschungsdesign und damit von dessen Repräsentativität abhängen. Dementsprechend besitzen die Ergebnisse einer MC-Studie immer nur für die vorgelegten Bedingungen Gültigkeit und können nicht ohne weiteres auf Sachverhalte übertragen werden, die nicht überprüft wurden. Damit ist die Generalisierbarkeit einer MC-Studie von vorneherein eingeschränkt. Aus diesem Grund wurde ein Simulationsdesign gewählt, das möglichst nahe an empirische Daten herankommt. Das kann allerdings immer nur bis zu einem bestimmten Maße erfolgen. Im Folgenden werden die Schwächen der Arbeit systematisiert.

Es wurden z. B. nur recht einfache Modelle untersucht. Wie die Performanz der MDTs unter noch komplexeren Strukturen zu bewerten ist, kann nicht beantwortet werden. Auch waren alle Modellvariablen bekannt und konnten im IM und im Analysemodell berücksichtigt werden. Damit liegt mit absoluter Sicherheit ein MAR-Ausfallmechanismus vor. Das kann allerdings in der Empirie nicht sichergestellt werden. Die Übertragbarkeit der vorliegenden Ergebnisse leidet demnach daran, dass für empirische Analysen ein MAR-Mechanismus nicht immer vorliegt. Andere Studien zeigen, dass Ergebnisse unter einem MAR-Ausfallmechanismus auch für MCAR-Missings Geltung besitzen, allerdings gilt dies nicht für einen NI-Ausfallmechanismus. In diesem Fall kann nicht garantiert werden, dass die vorliegenden Ergebnisse gültig bleiben. Auch wurden alle Modelle korrekt spezifiziert, was für die Empirie wohl nicht immer zutreffend ist. Zwar lässt sich festhalten, dass es für die MDTs unerheblich ist, welche Art von SE-Modellen geschätzt wird, allerdings wurde nicht geprüft, welche Auswirkung die Behandlung der fehlenden Werte hat, wenn das Analysemodell missspezifiziert ist. Besonders stellt sich diese Frage bei Direct-ML und H0. Denn beiden liegt die Annahme zugrunde, wonach das spezifizierte IM und Analysemodell korrekt ist. Fraglich ist, welche Auswirkungen es hat, wenn für Direct-ML im Analysemodell und für H0 zusätzlich noch im IM Missspezifikationen vorliegen. Sicherlich ist die Annahme gerechtfertigt, wonach Missspezifikationen Auswirkungen auf das Analyseergebnis haben.

Weiterhin wurden in der vorliegenden Arbeit auch nur rechtsschiefe Variablenverteilungen untersucht und alle Variablen hatten immer dieselben Schwellenwerte. Letztlich könnte das Ergebnis ein anderes sein, wenn alle Variablen linksschief verteilt sind oder, wenn unterschiedliche Schwellenwerte für einzelne Variablen verwendet werden. Zwar kann in dieser Arbeit ein Einfluss der Verteilungen auf die Performanzkriterien nicht nachgewiesen werden, allerdings

kann nicht ausgeschlossen werden, dass bei extremeren Verteilungen doch ein solcher vorliegt. Ähnliches gilt auch für die Samplegröße oder den Missinganteil: Es gibt keine Verzerrungen, die auch nur annähernd als problematisch einzustufen sind. Allerdings liegt in den Meta-Modellen ein Zusammenhang vor, sodass nicht ausgeschlossen werden kann, dass die Verzerrungen nach wie vor vernachlässigbar bleiben, wenn noch kleinere Fallzahlen oder noch mehr Missing Values vorliegen. Zudem wurden zwar Daten mit gemischt skalierten Variablen untersucht, allerdings wurde jeweils nur eine binär und eine ordinal skalierte Variable berücksichtigt. Werden dementsprechend gemischte Daten einer Untersuchung zugrunde gelegt, die mehrere solcher Variablen beinhalten, dann kann nicht ausgeschlossen werden, dass dies die MDTs beeinflusst.

Werden demnach andere Modelle getestet, andere Verteilungen untersucht, mehr diskrete Variablen berücksichtigt, andere Fallzahlen verwendet oder fällt der Missinganteil größer aus, dann dürften sich auch die Ergebnisse unterscheiden. Allerdings liegen, selbst in der Konfiguration mit einer Fallzahl von 250, unter stark asymmetrischen Verteilungen mit unterschiedlich skalierten Modellvariablen und 35 % Missing Values, in Bezug auf die Parameter und Standardfehler gute Ergebnisse vor. Wie realistisch es ist, dass noch weniger Fälle für die Analyse von SE-Modellen mit noch mehr Missing Values vorliegen, kann an dieser Stelle nicht beantwortet werden. Aber angesichts der Tatsache, dass selbst unter solchen Bedingungen gute Schätzergebnisse erzielt werden, kann davon ausgegangen werden, dass die getesteten MDTs in der empirischen Praxis sehr gut für die Behandlung fehlender Werte geeignet sind. Zumal sich die zentralen Tendenzen dieser Arbeit im empirischen Beispiel wiederfinden. Damit liegt ein Hinweis vor, wonach die Ergebnisse auch in anderen Kontexten Gültigkeit besitzen.

12.4 Ausblick

Letztlich verlangt eine MC-Studie einige Entscheidungen, welche das Forschungsdesign in eine bestimmte Richtung festlegen. Mit diesen Entscheidungen werden aber gleichzeitig auch andere, interessante Bereiche nicht beleuchtet. Das sind einerseits die naheliegenden Punkte, die bereits im vorherigen Kapitel angesprochen wurden und die Folgearbeiten untersuchen könnten, wie: komplexere Modelle, andere oder gemischte Variablenverteilungen, ein MCAR- oder NI-Ausfallmechanisums, kleinere Fallzahlen oder noch größere Missinganteile. Andererseits bietet sich auch die Möglichkeit neben den gewählten Konfigurationen für FCS andere zu testen, um zu prüfen, ob es keine Konfiguration von FCS gibt, die den anderen überlegen ist, wie eine Kombination aus PMM für die diskreten und linearen Regressionen für die quasi-metrischen Variablen oder die angebotenen *machine learning*-Verfahren. Zudem ist im Anschluss

an diese Arbeit auch ein Vergleich zwischen verschiedenen Statistikpaketen denkbar, der prüft, ob bspw. EQS mit deren Implementation von Direct-ML ähnliche Ergebnisse liefert wie die verwendete Implementation in M*plus* oder ob FCS in SPSS, SAS oder Stata ähnliche Ergebnisse generiert, wie FCS in R.

Weiterhin kann mit der Arbeit festgestellt werden, dass die explizite Berücksichtigung von diskreten Variablen mit FCS oder H0 nicht zwangsläufig zu besseren Ergebnissen führt. Im Grunde ergeben sich für diejenigen MDTs, welche das Skalenniveau nicht berücksichtigen, ähnliche Ergebnisse: Alle untersuchten MDTs können meist zufriedenstellende Ergebnisse erbringen. Wichtig ist allerdings, dass die explizite Berücksichtigung der Skalenniveaus im vorliegenden Fall auch eher unbedeutend ist, weil für die Analysestrategie keine diskreten Variablen benötigt werden. Wenn aber eine Analysemethode eingesetzt werden soll, die diskrete Variablen voraussetzt, wie eine binär oder multinomial logistische Regressionsanalyse, dann ist die Imputation der Missing Values mit EM, EMB oder MNV nicht mehr möglich, da mit diesen MDTs ‚nicht-realistische' Werte generiert werden, also Werte, die außerhalb der ursprünglichen Kategorien der Variablen liegen. Weil zudem das Runden dieser ‚nicht-realistischen' Werte eben keine erfolgreiche Strategie darstellt, sollten diese drei MDTs bereits im Vorhinein für solche Analysestrategien nicht angewendet werden. Wie die Performanz der anderen getesteten MDTs aber für Analysemethoden ausfällt, die diskrete Variablen voraussetzen, ist zum Teil noch ungeklärt. Vor allen Dingen könnte sich hier PMM als interessante Alternative erweisen, denn wie sich in dieser Arbeit zeigt, eignet sich diese auch bei diskreten Variablen mit wenigen Skalenpunkten. Fraglich ist dann, ob sie nicht eine bessere Alternative darstellt, als die für FCS verfügbaren logistischen Regressionsmodelle.

Für die Imputation der fehlenden Werte mit H0 wurden in dieser Arbeit die diskreten Variablen als solche ausgegeben. Weil sich nun aber für EM, EMB, MNV und PMM zeigt, dass es nicht notwendig ist, die diskreten Variablen explizit als solche zu berücksichtigen, stellt sich die Frage, ob dies auch für H0 der Fall ist. Hier könnte eine Untersuchung ansetzen, die prüft, ob eine Nicht-Berücksichtigung des Skalenniveaus von diskreten Variablen bei H0 tatsächlich keine besseren/schlechteren Ergebnisse erbringt. Infolgedessen könnte dann evaluiert werden, wie lange dies der Fall ist, oder ab wann diskrete Variablen als solche berücksichtigt werden sollten. Folglich wäre eine Fortsetzung der Evaluation dieser Technik angebracht.

Zum Schluss sollen noch zwei Punkte angesprochen werden, welche Folgeforschungen aufgrund der vorliegenden Ergebnisse priorisieren könnten. Zudem könnten diese für die empirische Praxis vielleicht auch von größerer Bedeutung sein: Das ist erstens das gute Abschneiden

von Direct-ML und H0 bei korrekt spezifizierten Modellen und das ist zweitens die fehlerbehaftete Modellbewertung von SE-Modellen, nachdem die fehlenden Werte mit MDTs ersetzt wurden, welche keine Rücksicht auf die Modellstruktur des Analysemodells nehmen.

Direct-ML und H0 sind die einzigen MDTs, die in der Lage sind, gute Anpassungswerte für SE-Modelle zu generieren, wenn Missing Values vorliegen. Das gilt allerdings nur, wenn sowohl das IM als auch das Analysemodell korrekt spezifiziert ist. Demzufolge muss untersucht werden, was passiert, wenn diese missspezifiziert sind. Für Direct-ML zeigt sich bereits, dass sie unter einem missspezifizierten Modell sehr gute Anpassungswerte liefern kann (Davey u. a. 2005). Wichtig wäre es herauszufinden, ob mit Direct-ML und sehr hohen Missinganteilen auch falsche Modelle geschätzt werden könnten. Hier stellt sich die Frage, ob der Missinganteil nur hoch genug sein muss, dass im Fall des Einsatzes von Direct-ML jedes Modell akzeptiert wird. Ähnliche Fragen stellen sich auch für H0: Was passiert, wenn das IM bereits komplett fehlspezifiziert ist? Lassen sich mit einem korrekt spezifizierten Analysemodell dann trotzdem korrekte Ergebnisse erwarten, oder sind diese nicht mehr akzeptabel? Zudem sollten die Folgen untersucht werden, wenn das IM korrekt, das Analysemodell aber missspezifiziert ist. Fraglich ist an dieser Stelle, ob sich an den Schätzergebnissen des Analysemodells ablesen lässt, dass dieses missspezifiziert ist. Bisher zeigt sich, dass H0 unter leicht fehlspezifizierten IMs trotzdem funktioniert. Allerdings liegt dazu lediglich der Nachweis von Asparouhov/Muthén (2010c) vor, bei welchem nicht näher erläutert wird, was unter einer ‚leichten Fehlspezifikation' zu verstehen ist.

Abschließend wird das problematische Ergebnis im Hinblick auf die Fit-Indices aufgegriffen. Das zentrale Ergebnis der Arbeit in Bezug darauf ist, dass mit den imputationsbasierten MDTs (mit Ausnahme von H0) in vielen Fällen keine zufriedenstellenden Werte für die Fit-Indices generiert werden, dafür aber unverzerrte Parameter und Standardfehler. Das hat zur Folge, dass in der Praxis korrekt spezifizierte Modelle nach der Imputation der fehlenden Werte, deren Schätzungen für die Parameter und Standardfehler zufriedenstellend sind und die in den inhaltlichen Schlüssen denen des Populationsmodells gleichen, abgelehnt werden müssen. Einzig anhand des SRMR lassen sich die Modelle korrekt bewerten und auch nur, wenn ein Grenzwert von .08 gewählt wird.

Weil zum jetzigen Zeitpunkt nur sehr wenig Forschung im Hinblick auf die Performanz der MDTs in Bezug auf die Fit-Indices existiert, diese sich vor allem mit Direct-ML, EM, MNV und dem p-Wert der Chi^2-Statistik bei metrischen Variablen, nicht aber mit anderen Fit-Indices

und auch nicht mit Variablen, die nicht metrisch skaliert sind, auseinandersetzt, kann nicht ausgeschlossen werden, dass es sich bei den vorliegenden Ergebnissen um forschungsdesignspezifische Ergebnisse handelt. Zumal für EM und MNV bisweilen in anderen Studien gute Ergebnisse vorliegen. Daher lassen sich zwei Fragen ableiten, welche die zukünftige Forschung zu beantworten hat:

1. Ist das vorgefundene Ergebnis lediglich durch das Forschungsdesign bestimmt, also resultieren die Ergebnisse daher, dass diskrete und quasi-metrische Variablen untersucht wurden, oder zeigt sich auch unter anderen Simulationskonfigurationen ein ähnliches Bild?
2. Sind die gängigen Grenzwerte der Fit-Indices für imputierte Daten zu streng und welche Grenzwerte führen zu korrekten Modellevaluationen?

Um eine adäquate Antwort auf diese Fragen liefern zu können, muss die Forschung an dieser Stelle anknüpfen. Das hat mit anderen Modelleigenschaften (komplexere Modelle, andere Werte für die Populationsparameter) und unter anderen Dateneigenschaften (anders skalierte Modellvariablen, andere Verteilungen, kleinere/größere Fallzahlen oder auch größere Anteile an Missing Values) zu erfolgen. Sollten sich dabei ähnliche Ergebnisse für die Fit-Indices ergeben, muss unter Variierung der Grenzwerte herausgefunden werden, welche für imputierte Daten herangezogen werden sollten.

13 Literaturverzeichnis

Allison, Paul D. 2000: Multiple Imputation for Missing Data. A Cautionary Tale. In: *Sociological Methods & Research*, Jg. 28, Nr. 3, 301–309.

Allison, Paul D. 2002: Missing Data. Thousand Oaks, CA: SAGE.

Allison, Paul D. 2003: Missing Data Techniques for Structural Equation Modeling. In: *Journal of Abnormal Psychology*, Jg. 112, Nr. 4, 545–557.

Allison, Paul D. 2009: Missing Data. In: Millsap, Roger Ellis (Hrsg.): *The Sage Handbook of Quantitative Methods in Psychology*, 1st ed., Los Angeles, CA: SAGE, 72–89.

Allison, Paul D. 2015: Imputation by Predictive Mean Matching: Promise & Peril. In: https://statisticalhorizons.com/predictive-mean-matching, zugegriffen am 06.08.2019.

Anderson, James C./Gerbing, David W. 1984: The Effect Of Sampling Error On Convergence, Improper Solutions, And Goodness-Of-Fit Indices For Maximum Likelihood Confirmatory Factor Analysis. In: *Psychometrika*, Jg. 49, Nr. 2, 155–173.

Andridge, Rebecca R./Little, Roderick J. A. 2010: A Review of Hot Deck Imputation for Survey Non-Response. In: *International Statistical Review*, Jg. 78, Nr. 1, 40–64.

Arbuckle, James L. 1996: Full Information Estimation in the Presence of Incomplete Data. In: Marcoulides, George A./Schumacker, Randall E. (Hrsg.): *Advanced Structural Equation Modeling. Issues and Techniques*, Mahwah, NJ: Lawrence Erlbaum Associates, 243–278.

Asparouhov, Tihomir/Muthén, Bengt O. 2010a: Bayesian Analysis Using Mplus: Technical Implementation. In: https://www.statmodel.com/download/Bayes3.pdf, zugegriffen am 06.08.2019.

Asparouhov, Tihomir/Muthén, Bengt O. 2010b: Chi-Square Statistics with Multiple Imputation. In: https://www.statmodel.com/download/MI7.pdf, zugegriffen am 20.12.2018.

Asparouhov, Tihomir/Muthén, Bengt O. 2010c: Multiple Imputation with Mplus. In: https://www.statmodel.com/download/Imputations7.pdf, zugegriffen am 06.08.2019.

Bandalos, Deborah L. 2014: Relative Performance of Categorical Diagonally Weighted Least Squares and Robust Maximum Likelihood Estimation. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 21, 102–116.

Bandalos, Deborah L./Gagné, Phillip 2012: Simulation Methods in Structural Equation Modeling. In: Hoyle, Rick H. (Hrsg.): *Handbook of Structural Equation Modeling*, New York, NY: The Guilford Press, 92–108.

Bandalos, Deborah L./Leite, Walter 2013: Use Of Monte Carlo Studies In Structural Equation Modeling Research. In: Hancock, Gregory R./Mueller, Ralph O. (Hrsg.): *Structural Equation Modeling. A Second Course*, 2nd ed., Charlotte, NC: Information Age Publishing, 625–666.

Bentler, Peter M. 1990: Comparative Fit Indexes in Structural Models. In: *Psychological Bulletin*, Jg. 107, Nr. 2, 238–246.

Bentler, Peter M. 2006: EQS 6 Structural Equations Program Manual. Encino, CA: Multivariate Software, Inc.

Bernaards, Coen A./Belin, Thomas R./Schafer, Joseph L. 2007: Robustness of a multivariate normal approximation for imputation of incomplete binary data. In: *Statistics in Medicine*, Jg. 26, 1368–1382.

Bodner, Todd E. 2008: What Improves with Increased Missing Data Imputations? In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 15, Nr. 4, 651–675.

Bollen, Kenneth A. 1989a: A New Incremental Fit Index for General Structural Equation Models. In: *Sociological Methods & Research*, Jg. 17, Nr. 3, 303–316.

Bollen, Kenneth A. 1989b: Structural Equations with Latent Variables. New York, NY: John Wiley & Sons.

Bollen, Kenneth A./Jackman, Robert W. 1990: Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases. In: Fox, John/Long, J. Scott (Hrsg.): *Modern Methods of Data Analysis*, Newbury Park, CA: SAGE, 257–291.

Boomsma, Anne 2013: Reporting Monte Carlo Studies in Structural Equation Modeling. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 20, Nr. 3, 518–540.

Borenstein, Michael/Hedges, Larry V./Higgins, Julian P. T./Rothstein, Hannah R. 2009: Introduction to Meta-Analysis. Chichester: John Wiley & Sons.

Brähler, Elmar/Decker, Oliver (Hrsg.) 2018: Flucht ins Autoritäre. Rechtsextreme Dynamiken in der Mitte der Gesellschaft die Leipziger Autoritarismus-Studie 2018. Gießen: Psychosozial-Verlag.

Brick, J. Michael 2013: Unit Nonresponse and Weighting Adjustments: A Critical Review. In: *Journal of Official Statistics*, Jg. 29, Nr. 3, 329–353.

Brick, J. Michael/Montaquila, Jill M. 2009: Nonresponse and Weighting. In: *Sample Surveys: Design, Methods and Applications*, 29A, 163–185.

Brooks, Stephen P./Gelman, Andrew 1998: General Methods for Monitoring Convergence of Iterative Simulations. In: *Journal of Computational and Graphical Statistics*, Jg. 7, Nr. 4, 434–455.

Brooks, Stephen P./Roberts, Gareth O. 1997: Assessing Convergence of Markov Chain Monte Carlo Algorithms. In: http://www.math.pitt.edu/~swigon/Homework/brooks97assessing.pdf, zugegriffen am 10.10.2018.

Brown, Timothy A. 2015: Confirmatory Factor Analysis for Applied Research. 2nd ed. New York, NY: The Guilford Press.

Browne, Michael W./Cudeck, Robert 1992: Alternative Ways of Assessing Model Fit. In: *Sociological Methods & Research*, Jg. 21, Nr. 2, 230–258.

Carpenter, James R./Kenward, Michael G. 2013: Multiple Imputation and its Application. 1st ed. Chichester: John Wiley & Sons.

Carsey, Thomas M./Harden, Jeffrey J. 2014: Monte Carlo Simulation and Resampling Methods for Social Science. Thousand Oaks, CA: SAGE.

Chen, Feinian/Bollen, Kenneth A./Paxton, Pamela/Curran, Patrick J./Kirby, James B. 2001: Improper Solutions in Structural Equation Models. Causes, Consequences, and Strategies. In: *Sociological Methods & Research*, Jg. 29, Nr. 4, 468–508.

Chen, Hua Yun/Xie, Hui/Qian, Yi 2011: Multiple Imputation for Missing Values through Conditional Semiparametric Odds Ratio Models. In: *Biometrics*, Jg. 67, Nr. 3, 799–809.

Chen, Nan/Li, Meijuan/Liu, Hongyun 2018: Comparison of maximum likelihood approach, Diggle–Kenward selection model, pattern mixture model with MAR and MNAR dropout data. In: *Communications in Statistics - Simulation and Computation*, 1–22.

Cohen, Jacob 1988: Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, Stephen R./Chu, Haitao/Greenland, Sander 2014: Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer. In: *American Journal of Epidemiology*, Jg. 179, Nr. 2, 252–260.

Collins, Linda M./Schafer, Joseph L./Kam, Chi-Ming 2001: A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. In: *Psychological Methods*, Jg. 6, Nr. 4, 330–351.

Curran, Patrick J./West, Stephen G./Finch, John F. 1996: The Robustness of Test Statistics to Nonnormality and Specification Error in Confirmatory Factor Analysis. In: *Psychological Methods*, Jg. 1, Nr. 1, 16–29.

Davey, Adam/Savla, Jyoti/Luo, Zupei 2005: Issues in Evaluating Model Fit with Missing Data. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 12, Nr. 4, 578–597.

Dempster, A. P./Laird, Nan M./Rubin, Donald B. 1977: Maximum Likelihood from Incomplete Data via the EM Algorithm. In: *Journal of the Royal Statistical Society: Series B (Methodological)*, Jg. 39, Nr. 1, 1–38.

Duncan, Otis D. 1975: Introduction to Structural Equation Models. New York, NY: Academic Press.

Eekhout, Iris/de Vet, Henrica C.W./Twisk, Jos W.R./Brand, Jaap P.L./de Boer, Michiel R./Heymans, Martijn W. 2014: Missing Data in a Multi-Item Instrument were best handled by Multiple Imputation at the Item Score Level. In: *Journal of Clinical Epidemiology*, Jg. 67, Nr. 3, 335–342.

Enders, Craig K. 2001a: A Primer on Maximum Likelihood Algorithms Available for Use with Missing Data. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 8, Nr. 1, 128–141.

Enders, Craig K. 2001b: The Impact of Nonnormality on Full Information Maximum-Likelihood Estimation for Structural Equation Models With Missing Data. In: *Psychological Methods*, Jg. 6, Nr. 4, 352–370.

Enders, Craig K. 2001c: The Performance Of The Full Information Maximum Likelihood Estimator In Multiple Regression Models With Missing Data. In: *Educational and Psychological Measurement*, Jg. 61, Nr. 5, 713–740.

Enders, Craig K. 2010: Applied Missing Data Analysis. New York, NY: The Guilford Press.

Enders, Craig K./Bandalos, Deborah L. 2001: The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 8, Nr. 3, 430–457.

Enders, Craig K./Baraldi, Amanda N./Cham, Heining 2014: Estimating Interaction Effects With Incomplete Predictor Variables. In: *Psychological Methods*, Jg. 19, Nr. 1, 39–55.

Enders, Craig K./Gottschall, Amanda C. 2011: Multiple Imputation Strategies for Multiple Group Structural Equation Models. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 18, Nr. 1, 35–54.

Ferro, Mark A. 2014: Missing data in longitudinal studies: cross-sectional multiple imputation provides similar estimates to full-information maximum likelihood. In: *Annals of Epidemiology*, Jg. 24, Nr. 1, 75–77.

Fiebig, Joachim 2012: Viktimisierung und Delinquenz. Die Bedeutung von Motivlagen bei der Erklärung pädosexuell straffälligen Verhaltens. Diss., Universität Stuttgart.

Finch, W. Holmes 2010: Imputation Methods for Missing Categorical Questionnaire Data: A Comparison of Approaches. In: *Journal of Data Science*, Jg. 8, 361–378.

Finney, Sara J./DiStefano, Christine 2013: Nonnormal And Categorical Data In Structural Equation Modeling. In: Hancock, Gregory R./Mueller, Ralph O. (Hrsg.): *Structural Equation Modeling. A Second Course*, 2nd ed., Charlotte, NC: Information Age Publishing, 439–492.

Firth, David 1993: Bias reduction of maximum likelihood estimates. In: *Biometrika*, Jg. 80, Nr. 1, 27–38.

Forero, Carlos G./Maydeu-Olivares, Alberto/Gallardo-Pujol, David 2009: Factor Analysis with Ordinal Indicators: A Monte Carlo Study Comparing DWLS and ULS Estimation. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 16, Nr. 4, 625–641.

Gaffert, Philipp/Meinfelder, Florian/Bosch, Volker 2016: Towards an MI-proper Predictive Mean Matching. Vorabveröffentlichung. In: https://www.uni-bamberg.de/fileadmin/uni/fakultaeten/sowi_lehrstuehle/statistik/Personen/Dateien_Florian/properPMM.pdf, zugegriffen am 06.08.2019.

Gelman, Andrew/Carlin, John B./Stern, Hal S./Dunson, David B./Vehtari, Aki/Rubin, Donald B. 2014: Bayesian Data Analysis. 3rd ed. Boca Raton, FL: CRC Press.

Gelman, Andrew/Rubin, Donald B. 1992: Inference from Iterative Simulation Using Multiple Sequences. In: *Statistical Science*, Jg. 7, Nr. 4, 457–511.

Gerbing, David W./Anderson, James C. 2016: Monte Carlo Evaluations of Goodness of Fit Indices for Structural Equation Models. In: *Sociological Methods & Research*, Jg. 21, Nr. 2, 132–160.

GESIS 2017: Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2016.

Ghani, Rayid/Schierholz, Malte 2017: Machine Learning. In: Foster, Ian/Ghani, Rayid/Jarmin, Ron S./Kreuter, Frauke/Lane, Julia (Hrsg.): *Big Data and Social Science. A Practical Guide to Methods and Tools*, Boca Raton, FL: CRC Press, 147–186.

Gold, Michael S./Bentler, Peter M. 2000: Treatments of Missing Data: A Monte Carlo Comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximization. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 7, Nr. 3, 319–355.

Gold, Michael S./Bentler, Peter M./Kim, Kevin H. 2003: A Comparison of Maximum-Likelihood and Asymptotically Distribution-Free Methods of Treating Incomplete Non-Normal Data. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 10, Nr. 1, 47–79.

Goldstein, Harvey/Carpenter, James R./Browne, William J. 2014: Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Jg. 177, Nr. 2, 553–564.

Goldstein, Harvey/Carpenter, James R./Kenward, Michael G./Levin, Kate A. 2009: Multilevel models with multivariate mixed response types. In: *Statistical Modelling*, Jg. 9, Nr. 3, 173–197.

Gottschall, Amanda C./West, Stephen G./Enders, Craig K. 2012: A Comparison of Item-Level and Scale-Level Multiple Imputation for Questionnaire Batteries. In: *Multivariate Behavioral Research*, Jg. 47, 1–25.

Graham, John W. 2003: Adding Missing-Data-Relevant Variables to FIML-Based Structural Equation Models. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 10, Nr. 1, 80–100.

Graham, John W. 2012: Missing Data. Analysis and Design. New York, NY: Springer.

Graham, John W./Hofer, Scott M./MacKinnon, David P. 1996: Maximizing the Usefulness of Data Obtained with Planned Missing Value Patterns: An Application of Maximum Likelihood Procedures. In: *Multivariate Behavioral Research*, Jg. 31, Nr. 2, 197–218.

Graham, John W./Olchowski, Allison E./Gilreath, Tamika D. 2007: How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. In: *Prevention Science*, Jg. 8, Nr. 3, 206–213.

Graham, John W./Schafer, Joseph L. 1999: On the Performance of Multiple Imputation for Multivariate Data With Small Sample Size. In: Hoyle, Rick H. (Hrsg.): *Statistical Strategies for Small Sample Research*, Thousand Oaks, CA: SAGE, 1–29.

Hallquist, Michael N. 2018: Package 'MplusAutomation'. In: https://cran.r-project.org/web/packages/MplusAutomation/MplusAutomation.pdf, zugegriffen am 24.10.2018.

Hallquist, Michael N./Wiley, Joshua F. 2018: MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 25, Nr. 4, 621–638.

Heinze, Georg/Schemper, Michael 2002: A solution to the problem of separation in logistic regression. In: *Statistics in Medicine*, Jg. 21, Nr. 16, 2409–2419.

Held, Leonhard/Bové, Daniel Sabanés 2014: Applied Statistical Inference. Likelihood and Bayes. Berlin: Springer.

Herrmann, Andrea 2001: Ursachen des Ethnozentrismus in Deutschland. Wiesbaden: VS Verlag für Sozialwissenschaften.

Heyder, Aribert/Gaßner, Anna 2012: Anomia, Deprivation und Werteorientierung zur Vorhersage rechtsextremistischer Einstellungen. Eine empirische Studie mit Repräsentativdaten aus Deutschland,. In: *Österreichische Zeitschrift für Politikwissenschaft*, Jg. 41, Nr. 3, 277–298.

Honaker, James/King, Gary 2010: What to Do about Missing Values in Time-Series Cross-Section Data. In: *American Journal of Political Science*, Jg. 54, Nr. 2, 561–581.

Honaker, James/King, Gary/Blackwell, Matthew 2011: Amelia II: A Program for Missing Data. In: *Journal of Statistical Software*, Jg. 45, Nr. 7, 1–47.

Honaker, James/King, Gary/Blackwell, Matthew 2012: AMELIA II: A Program for Missing Data. In: https://r.iq.harvard.edu/docs/amelia/amelia.pdf, zugegriffen am 11.10.2018.

Horton, Nicholas J./Kleinman, Ken P. 2007: Much Ado about Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. In: *The American Statistician*, Jg. 61, Nr. 1, 79–90.

Horton, Nicholas J./Lipsitz, Stuart R./Parzen, Michael 2003: A Potential for Bias when Rounding in Multiple Imputation. In: *The American Statistician*, Jg. 54, Nr. 4, 229–232.

Howard, Waylon J./Rhemtulla, Mijke/Little, Todd D. 2015: Using Principal Components as Auxiliary Variables in Missing Data Estimation. In: *Multivariate Behavioral Research*, Jg. 50, 285–299.

Hu, Li-tze/Bentler, Peter M. 1999: Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 6, Nr. 1, 1–55.

Iacobucci, Dawn 2010: Structural equations modeling: Fit Indices, sample size, and advanced topics. In: *Journal of Consumer Psychology*, Jg. 20, Nr. 1, 90–98.

Jackman, Simon 2009: Bayesian Analysis for the Social Sciences. Chichester: John Wiley & Sons.

Jia, Fan 2016: Methods for Handling Missing Non-Normal Data in Structural Equation Modeling. Diss., University of Kansas.

Jia, Fan/Wu, Wei 2019: Evaluating methods for handling missing ordinal data in structural equation modeling. In: *Behavior Research Methods*, 1–19.

Joanes, D. N./Gill, C. A. 1998: Comparing measures of sample skewness and kurtosis. In: *The Statistican*, Jg. 47, Nr. 1, 183–189.

Kaplan, David 2000: Structural Equation Modeling. Foundations and Extensions. Thousand Oaks, CA: SAGE.

Kaplan, David/Depaoli, Sarah 2012: Bayesian Structural Equation Modeling. In: Hoyle, Rick H. (Hrsg.): *Handbook of Structural Equation Modeling*, New York, NY: The Guilford Press, 650–673.

Kenny, David A./Kaniskan, Burcu/McCoach, D. Betsy 2014: The Performance of RMSEA in Models With Small Degrees of Freedom. In: *Sociological Methods & Research*, Jg. 44, Nr. 3, 1–22.

King, Gary/Honaker, James/Joseph, Anne/Scheve, Kenneth 2001: Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. In: *American Political Science Review*, Jg. 95, Nr. 1, 49–69.

Kleinke, Kristian 2017: Multiple Imputation Under Violated Distributional Assumptions: A Systematic Evaluation of the Assumed Robustness of Predictive Mean Matching. In: *Journal of Educational and Behavioral Statistics*, Jg. 42, Nr. 4, 371–404.

Kleinke, Kristian 2018: Multiple Imputation by Predictive Mean Matching when Sample Size is Small. In: *Methodology*, Jg. 14, Nr. 1, 3–15.

Krause, Thomas 2019: Multiple Indicator Growth Mixture Models: eine statistische Simulation zur Performanzevaluation für sozialwissenschaftliche Analysen. Diss., Universität Stuttgart.

Kropko, Jonathan/Goodrich, Ben/Gelman, Andrew/Hill, Jennifer 2013: Multiple Imputation for Continuous and Categorical Data: Comparing Joint and Conditional Approaches. In: http://www.stat.columbia.edu/~gelman/research/published/MI_manuscript_RR.pdf, zugegriffen am 03.12.2019.

Lall, Ranjit 2016: How Multiple Imputation Makes a Difference. In: *Political Analysis*, Jg. 24, Nr. 4, 414–433.

Lang, Kyle M./Wu, Wei 2017: A Comparison of Methods for Creating Multiple Imputations of Nominal Variables. In: *Multivariate Behavioral Research*, Jg. 52, Nr. 3, 290–304.

Lee, Katherine J./Carlin, John B. 2010: Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation. In: *American Journal of Epidemiology*, Jg. 171, Nr. 5, 624–632.

Lee, Min Cherng/Mitra, Robin 2016: Multiply imputing missing values in data sets with mixed measurement scales using a sequence of generalised linear models. In: *Computational Statistics & Data Analysis*, Jg. 95, 24–38.

Lei, Pui-Wa 2009: Evaluating estimation methods for ordinal data in structural equation modeling. In: *Quality and Quantity*, Jg. 43, Nr. 3, 495–507.

Leite, Walter/Beretvas, S. Natasha 2010: The Performance of Multiple Imputation for Likert-type Items with Missing Data. In: *Journal of Modern Applied Statistical Methods*, Jg. 9, Nr. 1, 64–74.

Li, Jian 2010: Effects of Full Information Maximum Likelihood, Expectation Maximization, Multiple Imputation, and Similar Response Pattern Imputation on Structural Equation Modeling with Incomplete and Multivariate Nonnormal Data. Diss., The Ohio State University.

Li, Jian/Lomax, Richard G. 2017: Effects of Missing Data Methods in SEM Under Conditions of Incomplete and Nonnormal Data. In: *The Journal of Experimental Education*, Jg. 85, Nr. 2, 231–258.

Lin, Ting Hsiang 2010: A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. In: *Quality and Quantity*, Jg. 44, Nr. 2, 277–287.

Little, Roderick J. A./Rubin, Donald B. 1987: Statistical Analysis with Missing Data. New York, NY: John Wiley & Sons.

Little, Roderick J. A./Rubin, Donald B. 2002: Statistical Analysis with Missing Data. 2nd ed. Hoboken, NJ: John Wiley & Sons.

Loehlin, John C./Beaujean, A. Alexander 2017: Latent Variable Models. An Introduction to Factor, Path, and Structural Equation Analysis. 5th ed. New York, NY: Routledge.

Longford, Nicholas T. 2005: Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician. New York, NY: Springer.

MacCallum, Robert C./Browne, Michael W./Sugawara, Hazuki M. 1996: Power Analysis and Determination of Sample Size for Covariance Structure Modeling. In: *Psychological Methods*, Jg. 1, Nr. 2, 130–149.

Marsh, Herbert W. 1998: Pairwise Deletion for Missing Data in Structural Equation Models: Nonpositive Definite Matrices, Parameter Estimates, Goodness of Fit, and Adjusted Sample Sizes. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 5, Nr. 1, 22–36.

McNeish, Daniel 2017: Missing data methods for arbitrary missingness with small samples. In: *Journal of Applied Statistics*, Jg. 44, Nr. 1, 24–39.

Meng, Xiao-Li 1994: Multiple-Imputation Inferences with Uncongenial Sources of Input. In: *Statistical Science*, Jg. 9, Nr. 4, 538–73.

Mooney, Christopher Z. 1997: Monte Carlo Simulation. Thousand Oaks, CA: SAGE.

Morris, Tim P./White, Ian R./Royston, Patrick 2014: Tuning multiple imputation by predictive mean matching and local residual draws. In: *BMC Medical Research Methodology*, Jg. 14, Nr. 75, 1–13.

Murray, Jared S./Reiter, Jerome P. 2016: Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models With Local Dependence. In: *Journal of the American Statistical Association*, Jg. 111, Nr. 516, 1466–1479.

Muthén, Bengt O./Kaplan, David 1992: A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. In: *British Journal of Mathematical and Statistical Psychology*, Jg. 45, 19–30.

Muthén, Bengt O./Kaplan, David/Hollis, Michael 1987: On Structural Equation Modeling With Data That Are Not Missing Completely At Random. In: *Psychometrika*, Jg. 52, Nr. 3, 431–462.

Muthén, Linda K./Muthén, Bengt O. 2012: Mplus User's Guide. 7th ed. Los Angeles, CA: Muthén & Muthén.

Muthén, Linda K./Muthén, Bengt O. 2017: Mplus User's Guide. 8th ed. Los Angeles, CA: Muthén & Muthén.

Newman, Daniel. A. 2003: Longitudinal Modeling With Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques. In: *Organizational Research Methods*, Jg. 6, Nr. 3, 328–362.

Olinsky, Alan/Chen, Shaw/Harlow, Lisa 2003: The comparative efficacy of imputation methods for missing data in structural equation modeling. In: *European Journal of Operational Research*, Jg. 151, Nr. 1, 53–79.

Orcan, Fatih 2013: Use Of Item Parceling In Structural Equation Modeling With Missing Data. Diss., Florida State University.

Paxton, Pamela/Curran, Patrick J./Bollen, Kenneth A./Kirby, James B./Chen, Feinian 2001: Monte Carlo Experiments: Design and Implementation. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 8, Nr. 2, 287–312.

Peters, Cara Lee Okleshen/Enders, Craig K. 2002: A primer for the estimation of structural equation models in the presence of missing data: Maximum likelihood algorithms. In: *Journal of Targeting, Measurement and Analysis for Marketing*, Jg. 11, Nr. 1, 81–95.

Pigott, Therese D. 2001: A Review of Methods for Missing Data. In: *Educational Research and Evaluation*, Jg. 7, Nr. 4, 353–383.

Plumpton, Catrin O./Morris, Tim P./Hughes, Dyfrig A./White, Ian R. 2016: Multiple imputation of multiple multi-item scales when a full imputation model is infeasible. In: *BMC Research Notes*, Jg. 9, 1–15.

Pritikin, Joshua N./Brick, Timothy R./Neale, Michael C. 2018: Multivariate normal maximum likelihood with both ordinal and continuous variables, and data missing at random. In: *Behavior Research Methods*, Jg. 50, Nr. 2, 490–500.

R Core Team 2015: R: A language and environment for statistical computing. Wien: R Foundation for Statistical Computing.

R Core Team 2017: R: A language and environment for statistical computing. Wien: R Foundation for Statistical Computing.

Raghunathan, Trivellore E./Lepkowski, James M./van Hoewyk, John/Solenberger, Peter 2001: A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. In: *Survey Methodology*, Jg. 27, Nr. 1, 85–95.

Rhemtulla, Mijke/Brosseau-Liard, Patricia É./Savalei, Victoria 2012: When Can Categorical Variables Be Treated as Continuous? A Comparison of Robust Continuous and Categorical SEM Estimation Methods Under Suboptimal Conditions. In: *Psychological Methods*, Jg. 17, Nr. 3, 354–373.

Rippl, Susanne/Baier, Dirk 2005: Das Deprivationskonzept in der Rechtsextremismusforschung: Eine vergleichende Analyse. In: *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, Jg. 57, Nr. 4, 644–666.

RStudio Team 2016: RStudio: Integrated Development for R. Boston, MA: RStudio, Inc.

Rubin, Donald B. 1976: Inference and missing data. In: *Biometrika*, Jg. 63, Nr. 3, 581–592.

Rubin, Donald B. 1987: Multiple Imputation for Nonresponse in Surveys. New York, NY: John Wiley & Sons.

Savalei, Victoria/Bentler, Peter M. 2005: A Statistically Justified Pairwise ML Method for Incomplete Nonnormal Data: A Comparison With Direct ML and Pairwise ADF. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 12, Nr. 2, 183–214.

Savalei, Victoria/Falk, Carl F. 2014: Robust Two-Stage Approach Outperforms Robust Full Information Maximum Likelihood With Incomplete Nonnormal Data. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 21, Nr. 2, 280–302.

Schafer, Joseph L. 1997: Analysis of Incomplete Multivariate Data. 1st ed. Boca Raton, FL: CRC Press.

Schafer, Joseph L. 2016: User's Guide for norm2. In: https://cran.r-project.org/src/contrib/Archive/norm2/, zugegriffen am 10.10.2018.

Schafer, Joseph L./Graham, John W. 2002: Missing Data: Our View of the State of the Art. In: *Psychological Methods*, Jg. 7, Nr. 2, 147–177.

Schafer, Joseph L./Novo, Alvaro A. 2015: Package 'norm'. In: https://cran.r-project.org/web/packages/norm/norm.pdf, zugegriffen am 11.10.2018.

Schenker, Nathaniel/Taylor, Jeremy M. G. 1996: Partially parametric techniques for multiple imputation. In: *Computational Statistics & Data Analysis*, Jg. 22, 425–446.

Schnell, Rainer/Hill, Paul B./Esser, Elke 2013: Methoden der empirischen Sozialforschung. 10., überarb. Aufl. München: Oldenbourg.

Shin, Tacksoo/Davison, Mark L./Long, Jeffrey D. 2017: Maximum Likelihood Versus Multiple Imputation for Missing Data in Small Longitudinal Samples With Nonnormality. In: *Psychological Methods*, Jg. 22, Nr. 3, 426–449.

Si, Yajuan/Reiter, Jerome P. 2013: Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys. In: *Journal of Educational and Behavioral Statistics*, Jg. 38, Nr. 5, 499–521.

Sinharay, Sandip 2003: Assessing Convergence of the Markov Chain Monte Carlo Algorithms: A Review. In: *ETS Research Report Series*, Nr. 1, 1-52.

Skrondal, Anders 2000: Design and Analysis of Monte Carlo Experiments: Attacking the Conventional Wisdom. In: *Multivariate Behavioral Research*, Jg. 35, Nr. 2, 137–167.

Spieß, Martin 2008: Missing-Data Techniken. Analyse von Daten mit fehlenden Werten. Münster: LIT.

Spieß, Martin 2010: Der Umgang mit fehlenden Werten. In: Wolf, Christof/Best, Henning (Hrsg.): *Handbuch der sozialwissenschaftlichen Datenanalyse*, 1. Aufl., Wiesbaden: VS Verlag für Sozialwissenschaften, 117–142.

Steiger, James H. 2007: Understanding the limitations of global fit assessment in structural equation modeling. In: *Personality and Individual Differences*, Jg. 42, Nr. 5, 893–898.

Tanner, Martin A./Wong, Wing Hung 1987: The Calculation of Posterior Distributions by Data Augmentation. In: *Journal of the American Statistical Association*, Jg. 82, Nr. 398, 528–540.

Teman, Eric Douglas 2012: The Performance Of Multiple Imputation And Full Information Maximum Likelihood For Missing Ordinal Data In Structural Equation Models. Diss., University of Northern Colorado.

Urban, Dieter/Mayerl, Jochen 2014: Strukturgleichungsmodellierung. Ein Ratgeber für die Praxis. Wiesbaden: Springer.

Urban, Dieter/Mayerl, Jochen 2018: Angewandte Regressionsanalyse: Theorie, Technik und Praxis. 5., überarb. Aufl. Wiesbaden: Springer.

van Buuren, Stef 2007: Multiple imputation of discrete and continuous data by fully conditional specification. In: *Statistical Methods in Medical Research*, Jg. 16, Nr. 3, 219–242.

van Buuren, Stef 2012: Flexible Imputation of Missing Data. Boca Raton, FL: CRC Press.

van Buuren, Stef 2015: Fully Conditional Specification. In: Molenberghs, Geert (Hrsg.): *Handbook of Missing Data Methodology*, Boca Raton, FL: CRC Press, 267–294.

van Buuren, Stef 2018: Flexible Imputation of Missing Data. 2nd ed. Boca Raton, FL: CRC Press.

van Buuren, Stef/Boshuizen, H. C./Knook, D. L. 1999: Multiple Imputation Of Missing Blood Pressure Covariates In Survival Analysis. In: *Statistics in Medicine*, Nr. 18, 681–694.

van Buuren, Stef/Brand, Jaap P.L./Groothuis-Oudshoorn, C. G. M./Rubin, Donald B. 2006: Fully conditional specification in multivariate imputation. In: *Journal of Statistical Computation and Simulation*, Jg. 76, Nr. 12, 1049–1064.

van Buuren, Stef/Groothuis-Oudshoorn, Karin 2011: mice: Multivariate Imputation by Chained Equations in R. In: *Journal of Statistical Software*, Jg. 45, Nr. 3, 1–67.

van Buuren, Stef/Vink, Gerko 2018: mice: Algorithmic convergence and inference pooling, zugegriffen am 11.10.2018.

van Ginkel, Joost R./Kroonenberg, Pieter M. 2017: Evaluation of multiple-imputation procedures for three-mode component models. In: *Journal of Statistical Computation and Simulation*, Jg. 87, Nr. 16, 3059–3081.

Vink, Gerrit 2015: Restrictive Imputation of Incomplete Survey Data. Diss., Universiteit Utrecht.

von Hippel, Paul T. 2009: How To Impute Interactions, Squares, And Other Transformed Variables. In: *Sociological Methodology*, Jg. 39, Nr. 1, 265–291.

von Hippel, Paul T. 2018: How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. In: *Sociological Methods & Research*, 1-20.

Wang, Huaping 2007: Missing Data Analysis In Structural Equation Modeling: Expectation Maximization And Multiple Imputation Methods. Diss., The University of Alabama.

White, Ian R./Carlin, John B. 2010: Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. In: *Statistics in Medicine*, Jg. 29, 2920–2931.

Wright, Sewall 1934: The Method Of Path Coefficients. In: *The Annals of Mathematical Statistics*, Jg. 5, Nr. 3, 161–215.

Wu, Wei/Jia, Fan/Enders, Craig K. 2015: A Comparison of Imputation Strategies for Ordinal Missing Data on Likert Scale Variables. In: *Multivariate Behavioral Research*, Jg. 50, Nr. 5, 484–503.

Yang-Wallentin, Fan/Jöreskog, Karl G./Luo, Hao 2010: Confirmatory Factor Analysis of Ordinal Variables With Misspecified Models. In: *Structural Equation Modeling: A Multidisciplinary Journal*, Jg. 17, Nr. 3, 392–423.

Yoo, Jin Eun 2009: The Effect of Auxiliary Variables and Multiple Imputation on Parameter Estimation in Confirmatory Factor Analysis. In: *Educational and Psychological Measurement*, Jg. 69, Nr. 6, 929–947.

Yoo, Jin Eun/French, Brian/Maller, Susan 2007: Inclusive Strategy with Confirmatory Factor Analysis, Multiple Imputation, and All Incomplete Variables. In: http://assets.pearsonglobalschools.com/asset_mgr/legacy/200746/CFA%20and%20Multiple%20Imputation_3969_1.pdf, zugegriffen am 20.07.2018.

Yu, L-M/Burton, Andrea/Rivero-Arias, Oliver 2007: Evaluation of software for multiple imputation of semi-continuous data. In: *Statistical Methods in Medical Research*, Jg. 16, Nr. 3, 243–258.

Zhang, Xiao/Li, Quanlin/Cropsey, Karen/Yang, Xiaowei/Zhang, Kui/Belin, Thomas R. 2017: Amultiple imputation method for incomplete correlated ordinal data using multivariate probit models. In: *Communications in Statistics - Simulation and Computation*, Jg. 46, Nr. 3, 2360–2375.

Zhu, Xiaoping 2014: Comparison of Four Methods for Handing Missing Data in Longitudinal Data Analysis through a Simulation Study. In: *Open Journal of Statistics*, Jg. 4, Nr. 11, 933–944.

14 Anhang

Tabellenverzeichnis Anhang

Abbildungsverzeichnis Anhang

A1 Forschungsstand: Simulationskonfigurationen

Tabelle A1: Simulationskonfigurationen der Studien in Kapitel 5.1

Autor(en)	FCS-Spezifikation	Modell[141]	Fallzahl	Skalierung (Kategorien)	Verteilungen	Missings (%)	Ausfallm.
Jia/Wu (2019)	binär bzw. multinomial log. Reg., *random forests*	3-Faktor-SE-Modell	300, 600	binär, ordinal (3, 5)	symmetrisch normal, (stark) asymmetrisch[142]	15, 30	MCAR, MAR
Pritikin u. a. (2018)	PMM, *proportional odds model*	1-Faktor-CFA-Modell	200, 500	metrisch, ordinal (3)	normal, symmetrisch normal	15, 25	--
Kleinke (2017)	PMM	lin. Reg.	50, 100, 500, 1000, 10000	metrisch	Skewness: .60, .68, .81, .99, 1.27, 1.63, 2.06, 2.57, 3.19	10, 20, 30, 40, 50	MAR
Kleinke (2018)	PMM	lin. Reg.	20, 30, 50, 100	metrisch	normal	10, 20, 30, 40, 50	MAR
Kropko u. a. (2013)	multinomial log. Reg., *renormalized* log. Reg., lin. Reg.	lin. Reg.	--	binär, ordinal (3 bis 10), metrisch	symmetrisch normal	25	MAR
Lang/Wu (2017)	multinomial log. Reg., *CART*	lin. Reg., log. Reg.	250, 500, 1000	ordinal (3, 5, 7, 10)	symmetrisch normal	10, 20, 30, 40, 50	MAR
McNeish (2017)	PMM, lin. Reg.	lin. Reg.	20, 50, 100, 250	metrisch	normal	10, 20, 30, 40, 50	MAR, NI
Raghunathan u. a. (2001)	lin. Reg.	lin. Reg.	100	metrisch	normal	30	MAR
White/Carlin (2010)	lin. Reg.	lin. Reg., log. Reg	300	metrisch	normal	30	MCAR, MAR
Wu u. a. (2015)	binär bzw. multinomial log. Reg.	lin. Reg.	125, 500	binär, ordinal (3, 5, 7)	symmetrisch normal, (stark) asymmetrisch	30, 50	MAR
Zhang u. a. (2017)	*proportional odds model*	*GEE method assuming a cumulative link model*[143]	500	ordinal (6, 7)	symmetrisch normal	10, 20, 30, 40, 50	MCAR, MAR

Anmerkungen: log. Reg.: logistische Regression; lin. Reg.: lineare Regression; CART: *classification and regressions trees*; symmetrisch normal: die Verteilung entspricht annähernd einer Normalverteilung; (stark) asymmetrisch: mehr Fälle liegen links/rechts der Mitte der Skala.

[141] Populationsmodell mit dem die Daten generiert werden.
[142] Generierung der Verteilungen durch Zuschneiden der normalverteilten Ausgangsvariablen in Kategorien anhand von z-Werten (siehe auch Kapitel 6).
[143] Interessierende Parameter sind darin die Intercepts, Zeit- und Gruppeneffekte.

Tabelle A2: Simulationskonfigurationen der Studien in Kap. 5.1 (Forts.) und Kap. 5.2

Autor(en)	FCS-Spezifikation	Modell	Fallzahl	Skalierung (Kategorien)	Verteilungen	Missings (%)	Ausfallm.
Lee/Carlin (2010)	log. Reg., *proportional odds model*	lin. Reg	1000	binär, ordinal	--[144]	33	MAR
van Buuren u. a. (2006)	lin. Reg., binär bzw. multinomial log. Reg.	Korrelationen, log. Reg.	400, 412	binär, ordinal, metrisch	--	62,5	MAR
Yu u. a. (2007)	PMM, lin. Reg.	MW, SD, Skewness	1060	metrisch	--	20, 40	MAR

Autor(en)	Modell	Fallzahl	Skalierung (Kategorien)	Verteilungen	Missings (%)	Ausfallm.
Asparouhov/Muthén (2010)	1-Faktor-CFA-Modell	70, 100, 200, 1000	metrisch, diskret	normal, symmetrisch normal	30	MAR
Chen u. a. (2018)	LGM	100, 300, 500, 1000	metrisch	normal	5, 10, 20, 40	MAR, NI
Gold/Bentler (2000)	4-Faktor-SEM	100, 500	metrisch	S = -2 bis S = 2 K = -1 bis K = 8	4, 8, 12, 16	MCAR
Newman (2003)	*Three-Wave Panel Model*	440	metrisch	normal	25, 50 75	MCAR, MAR, NI
Olinsky u. a. (2003)	4-Faktor-SEM	100, 500	metrisch	normal	2, 4, 8, 12, 16, 24, 32	MCAR
Enders (2001b)	lin. Reg.	100, 250, 400	metrisch	normal	5, 15, 25, 35	MCAR, MAR, NI
Graham u. a. (1996)	Kovarianzen	100, 500	metrisch	normal	33	MCAR, MAR
Honaker/King (2010)	lin. Reg.	1000	metrisch	normal	17, 22, 50	MCAR, MAR
King u. a. (2001)	lin. Reg.	1000	metrisch	normal	17, 22, 50	MCAR, MAR
Leite/Beretvas (2010)	Korrelationen	400	quasi-metrisch (3, 5, 7)	symmetrisch normal, asymmetrisch	10, 30 50	MCAR, MAR
Finch (2010)	multinomial log. Reg.	200, 500, 1000, 2000	ordinal (5)	--	25	MCAR, MAR
Lin (2010)	Nähe der Imp.	260, 650, 1300	quasi-metrisch (5)	--	2, 5, 10	--

Anmerkungen: LGM: Latent Growth Model; log. Reg.: logistische Regression; lin. Reg.: lineare Regression; Nähe der Imp.: die Distanz zwischen dem imputierten Wert und dem ursprünglich beobachteten Wert; MW: Mittelwert; SD: Standardabweichung; S: Skewness; K: Kurtosis.

[144] Bei empirischen Studien liegen höchstwahrscheinlich keine Normalverteilungen vor; allerdings können Verteilungen in solchen Studien auch nicht systematisch variiert werden, weshalb hier keine Angabe gemacht wird.

Tabelle A3: Simulationskonfigurationen der Studien in Kapitel 5.3

Autor(en)	Modell	Fallzahl	Skalierung (Kategorien)	Verteilungen	Missings (%)	Ausfallm.
Enders (2001a)	3-Faktoren-SE-Modell	250, 500, 750	metrisch	normal, S = 1.25/K = 3.5, S = 2.25/K = 7, S = 3.25/K = 20, S = 0/K = 3.5, S = 0/K = 7, S = 0/K = 20	0, 5, 10, 15, 25	MCAR, MAR
Enders/Bandalos (2001)	3-Faktoren-SE- und CFA-Modell	100, 250, 500, 750	metrisch	normal	2, 5, 10, 15, 25	MCAR, MAR
Ferro (2014)	LGM	5000	metrisch	normal	5, 10, 20	MAR
Li (2010)	3-Faktoren-CFA-Modell	100, 200, 500, 1000	metrisch	normal, S = 0.35/K = 1, S = 0.5/K = 1.5, S = 1.25/K = 3.5, S = 2.25/K = 7, S = 3.25/K = 20, S = 5/K = 70	5, 15, 30	MCAR, MAR
Li/Lomax (2017)	3-Faktoren-CFA-Modell	100, 200, 500, 1000	metrisch	normal, S = 0.35/K = 1, S = 0.5/K = 1.5, S = 1.25/K = 3.5, S = 2.25/K = 7, S = 3.25/K = 20, S = 5/K = 70	5, 15, 30	MCAR, MAR
Gold u. a. (2003)	4-Faktoren-SE-Modell	500, 5000	metrisch	normal	15, 30	MCAR, MAR
Savalei/Bentler (2005)	4-Faktoren-CFA-Modell	200, 300, 400, 500, 5000	metrisch	S = -3.03 bis 6.67 und K = 19.48 bis 328.81	15, 30	MCAR, MAR
Savalei/Falk (2014)	2-Faktoren-CFA- bzw. 3-Faktoren-SE-Modell	200, 400, 600	metrisch	S = 2/K = 7, S = 2/K = 15	15, 30	MCAR, MAR
Shin u. a. (2017)	LGM	50, 70, 100, 150	metrisch	S = 1/K = 3, S = 2.5/K = 9	0, 10, 16, 24	MAR
Teman (2012)	3-Faktoren-CFA- und 3-Faktoren-SE-Modell	250, 500, 750	ordinal (5)	S = 0/K = -0.08, S = 0/K = -1.305, S = 0/K = 2.810, S = 2.183/K = 3.843	5, 10, 25	MCAR, MAR
Wang (2007)	3-Faktoren-CFA-Modell	100, 300, 500	metrisch	normal	5, 15, 25	MCAR, MAR
Peters/Enders (2002)	3-Faktoren-CFA-Modell	300	metrisch	--	20	MCAR, MAR

Anmerkungen: LGM: Latent Growth Model; S: Skewness; K: Kurtosis.

A2 Durchführung der Simulationen auf dem bwUniCluster

Mithilfe des bwUniClusters war es möglich die Dauer der Simulationsläufe zu reduzieren, denn mittels dieses Parallelrechners konnten gleichzeitig bis zu 50 verschiedene Konfigurationen gerechnet werden. Um das Cluster ausnutzen zu können, war es allerdings notwendig für jegliche Konfiguration ein eigenes R-Skript zu schreiben[145], das dann an das Cluster weitergegeben werden konnte. Für alle MI-Techniken (mit Ausnahme von EMB) waren jeweils 54 Skripte notwendig.[146] Für EMB musste nicht jede Konfiguration einzeln gerechnet werden. Stattdessen konnten alle Konfigurationen eines Modells innerhalb eines Simulationslaufes durchgeführt werden. Damit waren für EMB lediglich drei Skripte notwendig – für jedes Modell eines. Schlussendlich wurden 219 Skripte verfasst. Die Simulationen für Direct-ML und EM (sowie für die vollständigen Daten; Referenzmodell) wurden auf einem handelsüblichen Desktoprechner durchgeführt.[147]

Der Ablauf der Simulationen ging dann wie folgt vonstatten: In einem ersten Schritt wurde die notwendige Software auf dem Cluster installiert. Das ist in diesem Fall M*plus* und R in den angesprochenen Versionen. Nachdem die Software installiert wurde, wurden zunächst die generierten Daten (die Datensätze der einzelnen Konfigurationen mit fehlenden Werten) auf das Cluster hochgeladen. Danach wurden die notwendigen R-Skripte vom lokalen Desktop-Rechner auf das Cluster kopiert. Um die zur Verfügung stehenden 50 möglichen, parallelen Berechnungen nutzen zu können, benötigte das Cluster nun sogenannte Batch-Jobs. Das sind Anweisungen, die dem Cluster mitteilen, welche Berechnungen es durchführen soll. Sie enthalten neben der Information welches R-Skript anzusteuern ist (also welche Berechnungen durchzuführen sind), auch dessen Speicherort, Angaben darüber welche Software für die Berechnungen notwendig ist (M*plus* und verschiedene R-Pakete), was die maximale Laufzeit der Berechnungen ist (hier: 24 Stunden) und die Anzahl der genutzten Knoten.[148] Demnach beinhalten die Batch-Jobs alle notwendigen Informationen, die das Cluster benötigt, um die Berechnungen durchführen zu können.[149] Der Batch-Job selbst ist allerdings lediglich als Anweisung an das

---

[145] Hierbei handelt es sich um automatische Prozess-Skripte: Zunächst werden die unvollständigen Daten einer Konfiguration eingelesen, dann werden die Datensätze vervielfacht (*m*) und die fehlenden Werte imputiert, um im Anschluss die aufgefüllten Datensätze abzuspeichern, damit sie analysiert werden können.

[146] Eines dieser Skripte – MNV, Modell 1, N = 250, symmetrisch, 5 % Missings („MNV_Model1_250_skew1_mar2“) – findet sich im Anhang O1.5. Die vorangestellten Ziffern verdeutlichen den Ablauf des Prozesses auf dem Cluster.

[147] Windows 7 (64-Bit), Intel Core i5-750, 2.76 GHz, 16 GB RAM.

[148] Eine Konfiguration wird auf einem, von den insgesamt 50 zur Verfügung stehenden Knoten gerechnet.

[149] Der Batch-Job für das R-Skript „MNV, Modell 1, N = 250, symmetrisch, 5 % Missings“ findet sich im Anhang O1.5 mit dem Namen „script1“.

Cluster zu verstehen. Um diese Anweisungen dem Cluster übermitteln zu können, müssen diese in sogenannten Shellskripten verfasst sein. Shellskripte sind dabei nichts Anderes als Textdateien, in der die gewünschten Anweisungen gespeichert sind. Im Endeffekt beinhalten also die Shellskripte, die Batch-Jobs, welche wiederum als Anweisungen an das Cluster verstanden werden können.[150] Da nun für die 219 R-Skripte je ein eigener Batch-Job und damit insgesamt 219 Shellskripte notwendig wurden, wurde dieser Prozess mit einem Generierungs-Shellskript automatisiert.[151] Dieses erstellte automatisch die gewünschten Batch-Jobs. Mithilfe der Übermittlungs-Shellskripte[152] konnten dann jeweils 50 Batch-Jobs an das Uni-Cluster mit dessen MOAB-Software übermittelt werden. Das Cluster stellte dann Warteschlangen auf und führte Benutzeraufträge auf der Grundlage von Fair-Sharing-Richtlinien aus.

Sowohl nach Beginn als auch nach Ende der Berechnungen wurde für jeden der 50 Batch-Jobs eine Nachricht versandt. Damit lag für jeden Batch-Job eine Nachricht für den Beginn der Berechnungen, sowie eine Nachricht für den Abschluss der Berechnungen vor. In letzterer wurde mittgeteilt, ob der jeweilige Batch-Job erfolgreich war oder nicht. Konnten die Berechnungen wie gewünscht durchgeführt werden (gab es keine Probleme bei dem Durchlauf der R-Skripte) wurde der Rechenvorgang erfolgreich beendet. An diesem Punkt lagen dann imputierte Datensätze für 50 Konfigurationen vor (bzw. für alle EMB-Konfigurationen). Im Folgenden konnten dann weitere Batch-Jobs mit den nächsten Übermittlungs-Skripten an das Cluster gesendet werden. Nachdem alle Imputationen vorlagen, wurden auf dem Cluster die Analysen für EMB, FCS, H0, MNV und PMM gerechnet (die Analysen für das Referenzmodell, Direct-ML und EM erfolgten auf dem Desktoprechner). Insgesamt lagen dann innerhalb von drei Wochen die Analyseergebnisse vor, die dann entsprechend der Bewertungskriterien ausgewertet werden konnten (das wiederum erfolgte auf dem Desktoprechner).

[150] Auf dem Cluster ist mit MOAB ein Cluster-Workload-Management-Paket (Adaptive Computing, Inc; https://www.bwhpc-c5.de/wiki/index.php/Batch_Jobs) installiert. Die Shellskripte enthalten die gewünschten MOAB-Befehle (den Batch-Job).
[151] „generate_SKRIPTE_MNV“ im Anhang O1.5.
[152] „submitMNV_1-50“ im Anhang O1.5.

A3 Konvergenz der Modellschätzungen

Tabelle A4: Konvergenzraten mit den MDTs

			Konvergenzrate																																			
			Modell 1														Modell 2												Modell 3											
			EM		EMB		FCS		FIML		H0		MNV		PMM		EM		EMB		FCS		FIML		H0		MNV		PMM		EM		EMB		FCS		FIML		H0		MNV		PMM	
Modellkonfiguration			N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
750	skew1	mar2	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
		mar3	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
		mar4	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
	skew2	mar2	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
		mar3	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
		mar4	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
	skew3	mar2	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
		mar3	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
		mar4	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
250	skew1	mar2	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
		mar3	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
		mar4	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
	skew2	mar2	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
		mar3	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
		mar4	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
	skew3	mar2	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
		mar3	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100
		mar4	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	**499**	100	500	100	500	100	500	100	500	100

Anmerkungen: Modellkonfiguration (von links nach rechts): Samplegröße, Verteilungen (skew1 = symmetrisch; skew2 = asymmetrisch; skew3 = stark asymmetrisch), Anteil an Missing Values (mar2 = 5 %; mar3 = 20 %, mar4 = 35 %). FIML = Direct-ML.[153]

[153] Diese Anmerkungen gelten, sofern nichts anderes angegeben wird, für alle Ergebnisdarstellungen (Tabellen und Grafiken).

Tabelle A5: Konvergenzraten des Referenzmodells

Modellkonfiguration			Konvergenzrate Modell 1 N	Modell 1 %	Modell 2 N	Modell 2 %	Modell 3 N	Modell 3 %
750	skew1	mar1	500	100	500	100	500	100
	skew2		500	100	500	100	500	100
	skew3		500	100	500	100	500	100
250	skew1		500	100	500	100	500	100
	skew2		500	100	500	100	500	100
	skew3		500	100	500	100	500	100

A4 Modellablehnungsraten

A4.1 Deskriptive Auswertung

Abbildung A1: Ablehnungsraten: SRMR und CFI

rejection rates in %

Modellkonfiguration			EM SRMR	EM CFI	EMB SRMR	EMB CFI	FCS SRMR	FCS CFI	FIML SRMR	FIML CFI	H0 SRMR	H0 CFI	MNV SRMR	MNV CFI	PMM SRMR	PMM CFI
750	skew1	mar2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		mar3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		mar4	0.0	0.4	0.0	1.2	0.0	0.8	0.0	0.0	0.0	0.0	0.0	1.4	0.0	1.2
	skew2	mar2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		mar3	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		mar4	0.0	**5.2**	0.0	**16.6**	0.0	**14.6**	0.0	0.0	0.0	0.0	0.0	**12.8**	0.0	**10.0**
	skew3	mar2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		mar3	0.0	1.0	0.0	1.0	0.0	1.6	0.0	0.0	0.0	0.0	0.0	1.4	0.0	0.8
		mar4	0.0	**23.6**	0.0	**67.6**	0.0	**63.8**	0.0	0.0	0.0	0.0	0.0	**55.8**	0.0	**49.0**
250	skew1	mar2	0.0	0.4	0.0	0.2	0.0	0.4	0.0	0.2	0.0	0.0	0.0	0.4	0.0	0.4
		mar3	0.0	**20.8**	0.0	**28.0**	0.0	**26.6**	0.0	0.0	0.0	0.0	0.0	**26.4**	0.0	**23.2**
		mar4	0.0	**70.0**	0.0	**97.2**	0.0	**95.4**	0.0	2.6	0.0	0.0	0.0	**95.8**	0.0	**93.6**
	skew2	mar2	0.0	1.0	0.0	1.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.8
		mar3	0.0	**36.6**	0.0	**52.6**	0.0	**50.8**	0.0	1.4	0.0	0.0	0.0	**49.2**	0.0	**45.8**
		mar4	0.0	**85.2**	0.0	**100.0**	0.0	**99.2**	0.0	**6.2**	0.0	0.0	0.0	**99.6**	0.0	**99.2**
	skew3	mar2	0.0	**6.2**	0.0	**6.0**	0.0	**6.6**	0.0	1.4	0.0	0.0	0.0	**6.2**	0.0	**6.2**
		mar3	0.0	**64.2**	0.0	**86.2**	0.0	**82.8**	0.0	**5.4**	0.0	0.0	0.0	**80.8**	0.0	**75.6**
		mar4	0.0	**94.8**	1.4	**100.0**	0.6	**100.0**	0.6	**18.2**	0.0	0.0	0.8	**100.0**	0.0	**99.8**

Anmerkungen: weiß: zufriedenstellende Ablehnungsraten (≤ 5 %); grün: eher zufriedenstellende Ablehnungsraten, die zwar über dem nominellen Level von 5 % liegen, in Relation zu den Ablehnungsraten des Referenzmodells aber als akzeptabel zu betrachten sind (die MDTs fügen den Modellablehnungen des Referenzmodells nicht nochmals mehr als 5 % an zusätzlichen Modellablehnungen hinzu); rot: inakzeptable Ablehnungsraten.

Abbildung A2: Ablehnungsraten: p-Wert, RMSEA und dessen 90 %iges KI

rejection rates in %																					
Modellkonfiguration			EM			EMB			FCS			FIML			H0			MNV			PMM		
			p-Wert	RMSEA	R. KI (90 %)	p-Wert	RMSEA	R. KI (90 %)	p-Wert	RMSEA	R. KI (90 %)	p-Wert	RMSEA	R. KI (90 %)	p-Wert	RMSEA	R. KI (90 %)	p-Wert	RMSEA	R. KI (90 %)	p-Wert	RMSEA	R. KI (90 %)
750	skew1	mar2	14.2	0.0	0.0	10.0	0.0	0.0	9.4	0.0	0.0	3.8	0.0	0.0	1.6	0.0	0.0	10.2	0.0	0.0	9.6	0.0	0.0
		mar3	68.2	4.4	0.0	69.2	4.4	0.0	67.8	4.2	0.0	3.8	0.0	0.0	0.0	0.0	0.0	67.6	4.8	0.0	65.8	4.6	0.0
		mar4	95.6	42.2	8.8	99.8	79.8	20.4	99.8	78.6	17.4	4.8	0.0	0.0	0.0	0.0	0.0	100.0	79.4	19.0	99.6	75.6	15.8
	skew2	mar2	19.2	0.0	0.0	13.0	0.0	0.0	13.0	0.0	0.0	4.2	0.0	0.0	1.4	0.0	0.0	13.2	0.0	0.0	11.6	0.0	0.0
		mar3	75.6	9.8	0.6	82.6	13.2	0.6	80.6	12.4	0.6	5.6	0.0	0.0	0.0	0.0	0.0	81.8	13.0	0.8	76.4	9.0	0.6
		mar4	97.4	62.4	19.6	100.0	96.8	54.8	100.0	94.4	48.6	6.6	0.0	0.0	0.0	0.0	0.0	100.0	94.0	45.0	100.0	88.4	30.4
	skew3	mar2	19.8	0.0	0.0	16.6	0.0	0.0	15.4	0.0	0.0	4.6	0.0	0.0	2.8	0.0	0.0	16.4	0.0	0.0	14.2	0.0	0.0
		mar3	83.2	19.4	1.8	90.4	33.8	2.2	89.2	31.4	2.4	5.0	0.0	0.0	0.4	0.0	0.0	88.6	30.0	2.2	83.4	16.0	0.6
		mar4	99.2	77.6	38.8	100.0	100.0	83.4	100.0	99.2	76.8	6.6	0.0	0.0	0.0	0.0	0.0	100.0	98.8	72.8	100.0	93.4	43.4
250	skew1	mar2	19.8	16.4	0.8	14.6	14.2	0.4	14.6	14.0	0.4	6.2	4.6	0.2	3.2	3.0	0.0	15.2	14.8	0.4	14.8	14.6	0.4
		mar3	72.2	69.0	27.2	76.4	85.8	38.0	73.4	82.0	29.2	6.8	6.2	0.0	0.2	0.4	0.0	73.6	82.0	32.4	70.0	80.4	26.6
		mar4	95.8	95.2	76.4	99.8	100.0	98.4	99.2	100.0	96.8	6.8	4.6	0.0	0.0	0.0	0.0	99.6	100.0	97.8	99.0	100.0	95.0
	skew2	mar2	27.6	23.2	1.4	18.6	18.2	0.4	18.6	18.0	0.4	7.4	5.8	0.0	2.4	2.2	0.0	18.8	18.6	0.2	17.0	17.2	0.2
		mar3	81.8	78.4	41.0	85.8	92.4	55.8	83.6	90.8	50.0	8.8	6.4	0.2	0.2	0.4	0.0	85.6	91.2	49.4	76.6	87.0	40.6
		mar4	98.6	98.6	88.6	100.0	100.0	100.0	99.8	100.0	99.8	11.2	8.8	0.0	0.0	0.0	0.0	100.0	100.0	100.0	100.0	100.0	99.6
	skew3	mar2	30.0	25.0	3.8	25.8	25.8	2.6	25.2	26.0	2.4	7.8	6.6	0.0	5.4	4.8	0.0	25.8	25.8	2.2	22.6	23.0	2.2
		mar3	88.2	86.4	54.0	93.2	97.2	75.8	92.0	95.4	69.8	9.2	7.6	0.2	0.0	0.2	0.0	91.0	95.0	72.0	82.8	92.4	53.4
		mar4	99.6	99.2	95.2	100.0	100.0	100.0	100.0	100.0	100.0	15.4	11.6	0.8	0.0	0.0	0.0	100.0	100.0	100.0	100.0	100.0	99.2

Anmerkungen: R. KI (90 %): 90 %iges Konfidenzintervall von RMSEA. Siehe außerdem die Anmerkungen zu Abbildung A1.

A4.2 Modellbasierte Analyse

Tabelle A6: Direct-ML – p-Wert

	p-Wert		
	LOGIT (95 % KI)	Odds Ratio (95 % KI)	AME in % (95 % KI)
Constant	-3.43***	0.03***	
	(-3.67, -3.20)	(0.03, 0.04)	
250	0.61***	1.85***	3.76***
	(0.44, 0.78)	(1.56, 2.19)	(2.74, 4.79)
skew2	0.33**	1.39**	2.11**
	(0.12, 0.54)	(1.13, 1.72)	(0.70, 3.52)
skew3	0.44***	1.56***	2.89***
	(0.24, 0.65)	(1.27, 1.92)	(1.46, 4.31)
mar3	0.15	1.16	0.95
	(-0.06, 0.37)	(0.94, 1.44)	(-0.40, 2.30)
mar4	0.45***	1.57***	2.91***
	(0.25, 0.65)	(1.28, 1.92)	(1.52, 4.31)
Fallzahl	9000		
Devianz (-2LL)	4437		
McFadden-R^2	2		
Nagelkerke-R^2	2.6		
Note:			*p<0.05; **p<0.01; ***p<0.001

Tabelle A7: Ergebnistabelle p-Wert

	p-Wert					
	Haupteffekte LOGIT (95 % KI)	Interaktionsmodell LOGIT (95 % KI)	Haupteffekte Odds Ratio (95 % KI)	Interaktionsmodell Odds Ratio (95 % KI)	Haupteffekte AME in % (95 % KI)	Interaktionsmodell AME in % (95 % KI)
Constant	-2.27*** (-2.34, -2.20)	-2.14*** (-2.25, -2.03)	0.10*** (0.10, 0.11)	0.12*** (0.11, 0.13)		
250	0.35*** (0.29, 0.41)	0.48*** (0.35, 0.60)	1.42*** (1.34, 1.51)	1.61*** (1.42, 1.82)	4.67*** (3.86, 5.48)	5.83*** (4.21, 7.45)
skew2	0.50*** (0.43, 0.58)	0.31*** (0.18, 0.45)	1.65*** (1.54, 1.78)	1.37*** (1.19, 1.57)	6.31*** (5.41, 7.21)	3.69*** (2.10, 5.29)
skew3	0.89*** (0.82, 0.96)	0.53*** (0.40, 0.67)	2.44*** (2.27, 2.62)	1.71*** (1.49, 1.96)	10.74*** (9.78, 11.69)	6.15*** (4.57, 7.73)
mar3	3.06*** (2.99, 3.12)	2.89*** (2.77, 3.01)	21.22*** (19.98, 22.56)	18.03*** (15.96, 20.39)	31.68*** (30.11, 33.26)	28.03*** (25.72, 30.34)
mar4	6.86*** (6.65, 7.08)	6.57*** (6.22, 6.91)	957.42*** (775.73, 1,198.51)	710.31*** (507.94, 1,021.27)	64.58*** (63.85, 65.31)	59.56*** (56.09, 63.04)
250:skew2		-0.02 (-0.16, 0.12)		0.98 (0.85, 1.13)		-0.26 (-2.01, 1.49)
250:skew3		0.06 (-0.09, 0.21)		1.07 (0.92, 1.24)		0.76 (-1.00, 2.53)
250:mar3		-0.24*** (-0.36, -0.12)		0.78*** (0.69, 0.89)		-3.14*** (-4.83, -1.45)
250:mar4		-0.47* (-0.91, -0.04)		0.62* (0.40, 0.96)		-6.49* (-13.15, 0.16)
skew2:mar3		0.28*** (0.14, 0.43)		1.33*** (1.15, 1.54)		3.21*** (1.68, 4.74)
skew3:mar3		0.58*** (0.43, 0.74)		1.79*** (1.54, 2.09)		6.07*** (4.69, 7.46)
skew2:mar4		0.74** (0.23, 1.25)		2.09** (1.27, 3.57)		7.36** (3.39, 11.32)
skew3:mar4		1.73*** (0.88, 2.58)		5.65*** (2.62, 14.73)		13.35*** (10.22, 16.47)
Fallzahl	44999	44999				
Devianz (-2LL)	29140	29044				
McFadden-R^2	49.8	49.9				
Nagelkerke-R^2	65.3	65.5				

Note: *p<0.05; **p<0.01; ***p<0.001

Tabelle A8: Ergebnistabelle RMSEA

	RMSEA					
	Haupteffekte LOGIT (95 % KI)	Interaktionsmodell LOGIT (95 % KI)	Haupteffekte Odds Ratio (95 % KI)	Interaktionsmodell Odds Ratio (95 % KI)	Haupteffekte AME in % (95 % KI)	Interaktionsmodell AME in % (95 % KI)
Constant	-6.25*** (-6.38, -6.11)	-6.14*** (-6.36, -5.92)	0.002*** (0.002, 0.002)	0.002*** (0.002, 0.003)		
250	4.01*** (3.91, 4.10)	4.37*** (4.18, 4.56)	55.04*** (50.16, 60.47)	79.14*** (65.66, 96.15)	76.18*** (75.20, 77.16)	79.68*** (77.93, 81.43)
skew2	0.73*** (0.65, 0.81)	0.57*** (0.28, 0.86)	2.08*** (1.92, 2.25)	1.77*** (1.32, 2.36)	17.86*** (15.97, 19.76)	13.95*** (7.03, 20.86)
skew3	1.38*** (1.30, 1.47)	1.19*** (0.90, 1.48)	3.99*** (3.68, 4.34)	3.28*** (2.46, 4.38)	32.54*** (30.79, 34.29)	28.17*** (21.86, 34.47)
mar3	3.59*** (3.50, 3.68)	3.13*** (2.99, 3.27)	36.10*** (33.00, 39.53)	22.88*** (19.81, 26.47)	67.39*** (66.29, 68.49)	61.60*** (59.66, 63.55)
mar4	7.30*** (7.17, 7.43)	7.02*** (6.79, 7.25)	1,484.06*** (1,305.32, 1,690.04)	1,117.38*** (890.96, 1,410.61)	90.54*** (90.08, 90.99)	89.34*** (88.44, 90.25)
250:skew2		-0.27* (-0.52, -0.02)		0.76* (0.59, 0.98)		-6.81* (-13.09, -0.53)
250:skew3		-0.55*** (-0.81, -0.29)		0.58*** (0.45, 0.75)		-13.65*** (-19.91, -7.39)
skew2:mar3		0.36*** (0.15, 0.57)		1.44*** (1.17, 1.78)		8.91*** (3.87, 13.94)
skew3:mar3		0.75*** (0.52, 0.97)		2.11*** (1.68, 2.65)		17.68*** (12.75, 22.62)
skew2:mar4		0.47** (0.15, 0.78)		1.60** (1.16, 2.19)		11.36** (3.93, 18.80)
skew3:mar4		0.65*** (0.31, 0.98)		1.91*** (1.36, 2.67)		15.49*** (7.97, 23.00)
Fallzahl	44999	44999				
Devianz (-2LL)	25257	25084				
McFadden-R^2	59.5	59.8				
Nagelkerke-R^2	74.9	75.1				

Note: *p<0.05; **p<0.01; ***p<0.001

Tabelle A9: Ergebnistabelle 90 %iges Konfidenzintervall des RMSEA

	RMSEA KI (90 %)					
	Haupteffekte LOGIT (95 % KI)	Interaktionsmodell LOGIT (95 % KI)	Haupteffekte Odds Ratio (95 % KI)	Interaktionsmodell Odds Ratio (95 % KI)	Haupteffekte AME in % (95 % KI)	Interaktionsmodell AME in % (95 % KI)
Constant	-9.87*** (-10.13, -9.62)	-9.68*** (-10.28, -9.09)	0.0001*** (0.0000, 0.0001)	0.0001*** (0.0000, 0.0001)		
250	4.29*** (4.17, 4.41)	4.34*** (4.16, 4.52)	72.77*** (64.75, 82.07)	76.49*** (64.06, 91.85)	40.58*** (38.63, 42.53)	40.21*** (37.28, 43.14)
skew2	0.96*** (0.88, 1.04)	0.05 (-0.79, 0.89)	2.61*** (2.40, 2.83)	1.05 (0.45, 2.46)	7.93*** (7.04, 8.82)	0.34 (-5.48, 6.16)
skew3	1.81*** (1.73, 1.90)	1.67*** (0.98, 2.35)	6.12*** (5.63, 6.66)	5.29*** (2.75, 10.92)	17.45*** (16.12, 18.77)	15.13*** (7.44, 22.82)
mar3	4.52*** (4.31, 4.74)	4.51*** (3.93, 5.08)	92.07*** (74.84, 114.76)	90.54*** (53.51, 170.15)	61.13*** (58.92, 63.35)	59.93*** (51.75, 68.12)
mar4	8.44*** (8.20, 8.68)	8.00*** (7.42, 8.59)	4,619.92*** (3,659.30, 5,899.32)	2,995.43*** (1,743.14, 5,692.53)	95.36*** (94.90, 95.81)	93.76*** (91.83, 95.70)
250:skew2		0.03 (-0.26, 0.33)		1.03 (0.77, 1.39)		0.22 (-1.84, 2.27)
250:skew3		0.06 (-0.24, 0.36)		1.06 (0.79, 1.44)		0.42 (-1.69, 2.54)
skew2:mar3		0.64 (-0.16, 1.43)		1.89 (0.84, 4.21)		5.40 (-2.82, 13.62)
skew3:mar3		-0.28 (-0.91, 0.35)		0.75 (0.38, 1.36)		-1.76 (-5.26, 1.74)
skew2:mar4		1.19** (0.36, 2.02)		3.29** (1.42, 7.61)		12.18** (0.32, 24.04)
skew3:mar4		0.53 (-0.15, 1.21)		1.70 (0.83, 3.26)		4.35 (-2.37, 11.07)
Fallzahl	44999	44999				
Devianz (-2LL)	22526	22434				
McFadden-R^2	59.6	59.7				
Nagelkerke-R^2	73.5	73.6				

Note: *p<0.05; **p<0.01; ***p<0.001

Tabelle A10: Ergebnistabelle CFI

	CFI					
	Haupteffekte LOGIT (95 % KI)	Interaktionsmodell LOGIT (95 % KI)	Haupteffekte Odds Ratio (95 % KI)	Interaktionsmodell Odds Ratio (95 % KI)	Haupteffekte AME in % (95 % KI)	Interaktionsmodell AME in % (95 % KI)
Constant	-11.02*** (-11.26, -10.78)	-11.41*** (-12.09, -10.73)	0.0000*** (0.0000, 0.0000)	0.0000*** (0.0000, 0.0000)		
250	5.38*** (5.25, 5.50)	5.78*** (5.62, 5.95)	215.94*** (190.18, 245.89)	324.87*** (275.81, 385.26)	44.39*** (42.82, 45.96)	45.77*** (43.30, 48.24)
skew2	1.14*** (1.04, 1.24)	0.90* (0.12, 1.68)	3.14*** (2.84, 3.46)	2.46* (1.17, 5.64)	6.94*** (6.18, 7.70)	4.54* (0.04, 9.05)
skew3	2.90*** (2.79, 3.00)	2.91*** (2.24, 3.59)	18.10*** (16.31, 20.12)	18.41*** (9.96, 38.96)	25.49*** (24.05, 26.93)	23.05*** (15.28, 30.81)
mar3	4.31*** (4.15, 4.47)	4.52*** (3.86, 5.18)	74.34*** (63.38, 87.68)	92.04*** (50.60, 192.69)	48.07*** (46.07, 50.08)	48.06*** (38.36, 57.76)
mar4	7.97*** (7.77, 8.16)	7.71*** (7.05, 8.38)	2,879.51*** (2,379.33, 3,502.53)	2,240.80*** (1,225.90, 4,706.81)	91.39*** (90.73, 92.06)	88.65*** (84.85, 92.46)
skew2:mar3		0.08 (-0.71, 0.86)		1.08 (0.47, 2.30)		0.34 (-3.26, 3.93)
skew3:mar3		-0.54 (-1.23, 0.14)		0.58 (0.27, 1.09)		-1.95 (-3.87, -0.03)
skew2:mar4		0.70 (-0.10, 1.50)		2.01 (0.86, 4.34)		3.89 (-1.86, 9.65)
skew3:mar4		0.83* (0.13, 1.53)		2.29* (1.06, 4.37)		4.86* (-0.79, 10.51)
Fallzahl	44999	44999				
Devianz (-2LL)	19210	19045				
McFadden-R^2	64.2	64.5				
Nagelkerke-R^2	76.8	77				
Note:						*p<0.05; **p<0.01; ***p<0.001

A4.2.1 Weitere Analysen und Validierungen

Tabelle A11: RMSEA: Ergebnisse getrennt nach Fallzahl

	RMSEA					
	250 LOGIT (95 % KI)	250 Odds Ratio (95 % KI)	250 AME in % (95 % KI)	750 LOGIT (95 % KI)	750 Odds Ratio (95 % KI)	750 AME in % (95 % KI)
Constant	-0.11***	0.89***		-3.77***	0.02***	
	(-0.17, -0.06)	(0.84, 0.94)		(-3.89, -3.65)	(0.02, 0.03)	
skew2	0.26***	1.30***	2.87***	1.00***	2.71***	18.61***
	(0.18, 0.34)	(1.20, 1.40)	(2.00, 3.74)	(0.88, 1.12)	(2.41, 3.05)	(16.23, 20.99)
skew3	0.49***	1.63***	5.31***	1.86***	6.45***	36.33***
	(0.41, 0.57)	(1.51, 1.77)	(4.40, 6.22)	(1.74, 1.99)	(5.69, 7.33)	(33.81, 38.85)
mar#	5.25***	190.06***	46.17***	4.67***	106.90***	80.82***
	(4.91, 5.58)	(138.48, 270.47)	(45.35, 46.99)	(4.57, 4.78)	(96.25, 118.97)	(79.88, 81.76)
Fallzahl	22499			22500		
Devianz (-2LL)	21022			13226		
McFadden-R^2	24.8			53.5		
Nagelkerke-R^2	37.3			68.5		

Note: *p<0.05; **p<0.01; ***p<0.001

Anmerkungen: Da die PML-Schätzung keine Generierung der AMEs zulässt, wird die mar-Variable rekodiert (mar#). Anstatt der 5 %-Missingquote besteht die Referenzkategorie des Dummies nun aus einer Mischung der 5 % und 20 %-Missingquote. Die Interpretation des Dummies ändert sich damit etwas, lässt aber weiterhin die Interpretation des Einflusses des größten Missinganteils zu. Zusätzlich wird eine Analyse der Ausgangskodierung der mar-Variablen mit der PML-Schätzung durchgeführt. Hierbei zeigen sich für das kleine Sample weniger ausgeprägte Effekte für die Missingkategorien als für das große Sample. Auch dabei zeigt sich aber, dass vor allem die höchste Missingkategorie problematisch ist – weniger dagegen die 20 %-Kategorie.

Tabelle A12: RMSEA KI (90 %): Ergebnisse getrennt nach Fallzahl

	RMSEA KI (90 %)					
	250 LOGIT (95 % KI)	250 Odds Ratio (95 % KI)	250 AME in % (95 % KI)	750 LOGIT (95 % KI)	750 Odds Ratio (95 % KI)	750 AME in % (95 % KI)
Constant	-1.75***	0.17***		-6.98***	0.001***	
	(-1.82, -1.67)	(0.16, 0.19)		(-7.25, -6.70)	(0.001, 0.001)	
skew2	0.61***	1.85***	14.48***	1.23***	3.43***	2.64***
	(0.52, 0.71)	(1.68, 2.03)	(12.30, 16.65)	(1.10, 1.36)	(3.01, 3.92)	(2.11, 3.18)
skew3	1.10***	2.99***	25.07***	2.19***	8.96***	6.10***
	(1.00, 1.19)	(2.73, 3.28)	(23.10, 27.05)	(2.06, 2.33)	(7.86, 10.24)	(5.06, 7.14)
mar#	4.58***	97.09***	73.25***	5.33***	205.83***	37.24***
	(4.44, 4.71)	(85.36, 110.91)	(72.44, 74.06)	(5.07, 5.58)	(160.77, 268.73)	(36.02, 38.46)
Fallzahl	22499			22500		
Devianz (-2LL)	18414			9622		
McFadden-R^2	40.9			46		
Nagelkerke-R^2	57.7			55.8		

Note: *p<0.05; **p<0.01; ***p<0.001

Anmerkungen: Wie schon für den RMSEA, wird auch hier für die normale ML-Schätzung die mar-Variable rekodiert (mar#). Gleichzeitig wird eine PML-Schätzung für das Konfidenzintervall mit der Ausgangskodierung durchgeführt. Darin zeigt sich, dass für das große Sample vor allem die Missingkategorien problematisch sind und wie in der vorliegenden Tabelle auch, weniger die anderen unabhängigen Variablen (vor allen Dingen im Vergleich der Koeffizienten). Etwaiges kann bereits in den Ergebnissen der deskriptiven Analyse beobachtet werden. Dort zeigt sich, dass das Konfidenzintervall bei großer Fallzahl eher robust ist und nur Probleme bekundet, wenn sehr viele Missing Values vorliegen.

Tabelle A13: CFI: Ergebnisse getrennt nach Fallzahl

	CFI					
	250 LOGIT (95 % KI)	250 Odds Ratio (95 % KI)	250 AME in % (95 % KI)	750 LOGIT (95 % KI)	750 Odds Ratio (95 % KI)	750 AME in % (95 % KI)
Constant	-1.98***	0.14***		-9.92***	0.0000***	
	(-2.06, -1.90)	(0.13, 0.15)		(-10.46, -9.39)	(0.0000, 0.0001)	
skew2	0.84***	2.31***	19.69***	2.59***	13.31***	1.66***
	(0.74, 0.94)	(2.10, 2.55)	(17.52, 21.87)	(2.18, 3.00)	(9.01, 20.61)	(1.15, 2.18)
skew3	1.67***	5.34***	36.81***	4.69***	108.59***	6.85***
	(1.58, 1.77)	(4.85, 5.87)	(34.98, 38.63)	(4.29, 5.09)	(74.35, 166.55)	(5.07, 8.63)
mar#	4.36***	78.24***	71.98***	5.32***	203.88***	10.10***
	(4.24, 4.48)	(69.50, 88.35)	(71.11, 72.84)	(4.95, 5.68)	(144.20, 300.33)	(8.83, 11.37)
Fallzahl	22499			22500		
Devianz (-2LL)	18741			5939		
McFadden-R^2	39.9			49.7		
Nagelkerke-R^2	56.7			56.2		

Note: *p<0.05; **p<0.01; ***p<0.001

Anmerkungen: siehe die Anmerkungen zu Tabelle A12.

A5 Parameterbias und Effizienz

A5.1 Deskriptive Auswertung: Relativer Parameterbias

Abbildung A3: Bereinigter relativer Parameterbias der Faktorladungen

bereinigter relativer Bias in %: Faktorladungen																													
			EM				EMB				FCS				FIML				H0				MNV				PMM			
			ind2	ind3	ind4	ind5	ind2	ind3	ind4	ind5	ind2	ind3	ind4	ind5	ind2	ind3	ind4	ind5	ind2	ind3	ind4	ind5	ind2	ind3	ind4	ind5	ind2	ind3	ind4	ind5
Modellkonfiguration			.8	.7	.6	.5	.8	.7	.6	.5	.8	.7	.6	.5	.8	.7	.6	.5	.8	.7	.6	.5	.8	.7	.6	.5	.8	.7	.6	.5
750	skew1	mar2	0.0	0.0	0.1	-0.2	0.0	0.0	0.2	-0.1	0.0	-0.1	0.1	-0.2	0.0	0.0	0.1	-0.1	0.0	-0.1	0.1	-0.2	0.0	0.0	0.1	-0.2	0.0	-0.1	0.1	-0.2
		mar3	-0.2	-0.2	0.0	-0.3	0.1	0.1	0.4	0.1	-0.2	-0.3	0.0	-0.4	0.0	0.0	0.3	-0.1	-0.1	-0.1	0.2	-0.1	-0.1	-0.2	0.2	-0.3	-0.2	-0.2	-0.1	-0.4
		mar4	-0.1	0.0	0.7	0.6	0.0	0.1	0.8	0.6	-0.6	-0.6	-0.3	-0.6	-0.1	0.0	0.5	0.3	-0.3	-0.2	0.3	0.0	-0.3	-0.3	0.3	-0.1	-0.5	-0.5	-0.3	-0.7
	skew2	mar2	0.1	0.2	0.4	0.3	0.1	0.2	0.5	0.4	0.1	0.1	0.4	0.2	0.1	0.2	0.5	0.3	0.1	0.2	0.5	0.3	0.1	0.2	0.5	0.3	-0.1	-0.1	0.0	-0.1
		mar3	0.4	0.7	1.7	1.6	0.5	0.8	1.9	1.9	0.3	0.4	1.2	1.2	0.5	0.8	1.8	1.9	0.5	0.6	1.7	1.7	0.4	0.6	1.7	1.6	-0.2	-0.2	-0.1	-0.1
		mar4	0.9	1.1	3.1	3.4	0.8	1.0	3.1	3.3	0.2	0.2	1.7	1.8	0.8	0.9	3.0	3.1	0.6	0.7	2.7	2.8	0.5	0.6	2.6	2.6	-0.4	-0.8	-0.4	-0.7
	skew3	mar2	0.3	0.3	0.7	0.7	0.3	0.4	0.9	0.7	0.3	0.3	0.8	0.6	0.3	0.4	0.9	0.7	0.3	0.3	0.8	0.6	0.3	0.4	0.8	0.6	0.0	-0.1	0.1	-0.1
		mar3	0.8	0.8	2.9	2.2	0.9	1.1	3.1	2.7	0.7	0.7	2.3	1.8	1.0	1.1	3.0	2.6	0.9	1.0	2.8	2.4	0.9	0.9	3.0	2.4	-0.1	-0.5	0.0	-0.8
		mar4	1.3	1.5	5.3	4.6	1.2	1.3	4.9	4.6	0.6	0.6	3.1	2.6	1.3	1.5	4.7	4.4	1.1	1.2	4.4	4.1	0.9	0.9	4.3	3.8	-0.4	-1.2	-0.8	-1.7
250	skew1	mar2	0.1	-0.1	0.1	-0.1	0.1	-0.1	0.2	0.0	-0.1	-0.3	-0.1	-0.3	0.0	-0.1	0.1	-0.1	0.0	-0.2	0.1	-0.2	0.0	-0.3	0.0	-0.3	-0.1	-0.3	-0.1	-0.4
		mar3	-0.2	0.0	0.2	-0.3	0.0	0.4	0.7	0.2	-0.8	-0.7	-0.6	-1.3	-0.2	0.0	0.4	-0.2	-0.4	-0.3	0.1	-0.5	-0.5	-0.2	-0.2	-0.6	-0.8	-0.7	-0.8	-1.5
		mar4	-0.1	0.3	0.8	-0.2	0.0	0.6	1.2	0.7	-1.8	-1.6	-1.7	-2.5	-0.3	0.1	0.7	0.2	-0.8	-0.4	-0.1	-0.6	-1.0	-0.7	-0.4	-1.1	-1.7	-1.7	-2.1	-3.0
	skew2	mar2	0.1	0.3	0.6	0.3	0.1	0.2	0.6	0.4	-0.1	0.0	0.2	0.0	0.1	0.2	0.5	0.4	0.0	0.1	0.5	0.3	0.0	0.0	0.4	0.2	-0.3	-0.2	-0.2	-0.4
		mar3	0.3	0.8	1.7	1.5	0.5	1.0	1.9	1.8	-0.4	-0.1	0.3	0.0	0.5	0.9	1.9	1.8	0.2	0.7	1.5	1.4	0.1	0.5	1.0	1.2	-0.9	-0.8	-1.2	-1.5
		mar4	0.3	1.2	2.6	2.4	0.6	1.2	2.9	3.2	-1.1	-1.3	-0.7	-0.7	0.7	1.2	3.0	3.0	0.1	0.6	2.2	2.1	-0.2	0.1	1.3	1.5	-1.7	-2.0	-2.4	-2.5
	skew3	mar2	0.2	0.3	0.9	0.1	0.3	0.3	0.9	0.6	0.1	0.0	0.6	0.2	0.3	0.4	0.9	0.6	0.2	0.3	0.8	0.5	0.2	0.1	0.8	0.4	-0.2	-0.3	-0.1	-0.5
		mar3	0.6	0.5	2.9	1.5	0.7	1.0	3.1	2.2	-0.2	-0.2	1.2	0.1	0.9	1.2	3.3	2.4	0.7	0.8	2.8	1.9	0.4	0.6	2.2	1.5	-0.9	-1.3	-1.2	-2.3
		mar4	0.5	1.1	5.6	3.9	0.2	0.8	5.1	4.0	-1.7	-1.7	0.4	-0.4	1.1	1.7	5.4	4.4	0.5	0.9	4.3	3.3	-0.6	-0.2	3.4	2.3	-2.4	-2.8	-2.2	-3.1

Abbildung A4: Bereinigter relativer Parameterbias der Strukturpfade

bereinigter relativer Bias in %: Strukturpfade																															
			EM				EMB				FCS				FIML				H0				MNV				PMM				
			x1>f1	x2>f1	x3>f1	x4>f1	x1>f1	x2>f1	x3>f1	x4>f1	x1>f1	x2>f1	x3>f1	x4>f1	x1>f1	x2>f1	x3>f1	x4>f1	x1>f1	x2>f1	x3>f1	x4>f1	x1>f1	x2>f1	x3>f1	x4>f1	x1>f1	x2>f1	x3>f1	x4>f1	
Modellkonfiguration			.1	.3	.3	.5	.1	.3	.3	.5	.1	.3	.3	.5	.1	.3	.3	.5	.1	.3	.3	.5	.1	.3	.3	.5	.1	.3	.3	.5	
		mar2	-0.3	0.1	0.2	-0.1	0.4	0.0	0.2	-0.2	0.6	0.1	0.2	-0.2	0.4	0.0	0.2	-0.2	0.4	0.0	0.1	-0.2	0.8	0.1	0.1	-0.2	0.6	0.0	0.1	-0.2	
	skew1	mar3	-1.0	-0.3	0.4	-0.1	-0.5	-0.3	0.2	0.2	0.4	0.2	0.0	-0.2	-0.1	0.0	0.2	0.0	0.3	0.1	-0.3	-0.3	0.7	-0.1	0.1	0.0	0.6	0.1	-0.1	-0.1	
		mar4	-1.1	0.2	0.4	0.2	-2.3	-0.8	0.7	0.5	0.0	0.3	0.0	-0.4	-0.9	-0.1	0.2	0.1	1.4	0.5	-0.9	-0.8	0.4	-0.2	0.2	0.0	0.1	0.2	-0.3	-0.2	
		mar2	-1.2	-0.3	0.4	0.2	-0.7	-0.3	0.3	0.2	0.0	0.1	0.3	-0.2	-0.8	-0.2	0.2	0.2	-0.6	0.0	0.0	-0.1	-0.2	-0.2	0.2	0.1	0.6	0.3	0.0	-0.3	
750	skew2	mar3	-5.2	-0.1	0.8	0.7	-4.5	-1.0	0.8	1.1	-2.3	-0.1	0.8	0.1	-4.2	-0.8	0.7	0.9	-3.0	-0.2	-0.4	0.0	-2.9	-0.9	0.7	0.8	0.3	0.6	-0.5	-0.6	
		mar4	-3.5	-1.1	0.7	1.3	-6.3	-2.2	1.7	1.8	-3.7	-0.8	1.3	0.6	-6.4	-1.6	1.1	1.5	-2.7	-0.2	-1.1	-0.3	-3.7	-1.6	1.1	1.4	2.0	0.2	-1.0	-0.9	
		mar2	-2.8	-0.3	0.2	0.4	-1.7	-0.4	0.4	0.3	-1.3	-0.1	0.5	-0.1	-1.7	-0.4	0.4	0.3	-1.4	-0.1	-0.1	-0.2	-0.9	-0.3	0.3	0.3	0.0	0.2	-0.2	-0.3	
	skew3	mar3	-5.5	-1.1	1.6	1.1	-5.0	-1.0	1.5	1.2	-5.5	-0.6	1.7	0.5	-5.1	-1.1	1.4	1.1	-2.4	0.1	-1.1	-1.0	-3.3	-0.9	1.4	0.9	0.8	0.5	-1.0	-1.0	
		mar4	-2.7	-1.8	1.4	1.3	-6.9	-2.6	2.0	2.1	-8.8	-2.5	2.3	1.5	-8.6	-2.2	1.7	1.8	-1.6	0.2	-3.2	-2.1	-4.0	-2.1	1.7	1.6	4.0	-0.4	-3.1	-1.5	
		mar2	-0.1	-0.1	0.4	0.1	-0.3	-0.2	0.6	0.1	0.6	0.1	0.2	-0.2	0.1	0.0	0.4	0.0	0.4	0.0	0.2	-0.2	0.3	-0.1	0.4	0.0	0.4	0.0	0.2	-0.1	
	skew1	mar3	-4.9	-0.5	1.4	1.1	-2.9	-0.8	1.4	1.0	0.3	0.8	0.1	-0.3	-1.4	0.1	0.5	0.4	-0.5	0.4	-0.3	-0.3	-0.4	0.0	0.8	0.3	0.7	0.6	-0.2	-0.2	
		mar4	-6.4	-0.8	0.9	1.0	-7.6	-2.2	2.9	1.8	1.1	1.3	-0.5	-1.2	-2.5	0.1	0.7	0.4	1.5	1.1	-1.7	-1.6	0.1	0.0	0.5	0.0	0.9	1.0	-0.6	-0.9	
		mar2	-3.9	-0.6	0.5	0.8	-1.9	-0.5	0.4	0.5	-0.7	0.0	0.1	-0.1	-1.9	-0.3	0.3	0.4	-1.4	-0.1	0.0	0.0	-1.4	-0.3	0.3	0.4	0.1	0.1	-0.1	-0.2	
250	skew2	mar3	**-10.1**	-0.7	1.7	1.7	-8.7	-1.4	2.0	2.0	-4.8	0.4	0.9	0.1	-8.4	-0.6	1.2	1.5	-6.4	0.3	-0.7	0.1	-5.5	-0.2	1.3	1.1	-2.7	0.7	-0.4	-0.2	
		mar4	**-12.9**	-2.4	2.2	2.8	**-12.8**	-3.3	3.1	3.2	-6.9	-0.1	-0.2	-0.1	**-11.0**	-1.4	1.7	2.0	-5.2	0.5	-2.2	-1.0	-5.8	-1.0	0.8	1.4	-2.0	0.5	-1.4	-0.8	
		mar2	-5.7	-0.4	1.2	0.8	-2.3	-0.5	0.7	0.6	-2.0	-0.2	0.5	-0.1	-2.5	-0.5	0.7	0.5	-2.0	-0.1	-0.1	-0.1	-1.7	-0.4	0.6	0.5	-0.5	0.1	-0.1	-0.3	
	skew3	mar3	-9.4	-1.7	1.6	2.2	-7.0	-2.2	2.6	2.2	-8.5	-1.4	1.2	0.1	-8.6	-1.6	2.0	1.7	-5.1	-0.3	-1.6	-1.1	-4.4	-1.2	2.0	1.3	-0.4	-0.1	-1.3	-0.8	
		mar4	**-20.0**	-3.4	3.6	4.1	**-15.2**	-4.5	4.4	4.1	**-23.3**	-3.4	-0.3	0.8	**-16.1**	-2.9	2.6	3.1	-7.6	0.0	-5.1	-2.2	-9.0	-2.0	1.7	2.3	1.5	-1.4	-3.2	-1.7	

Anmerkungen: Fettdruck: Überschreitung des gesetzten Grenzwertes von ± 10 %. Rot und Fettdruck: Überschreitung des gesetzten Grenzwertes von ± 15 %.

Abbildung A5: Bereinigter relativer Parameterbias der Kovarianzen I

bereinigter relativer Bias in %: Kovarianzen																									
			EM						EMB						FCS						FIML					
			x1-x2	x1-x3	x2-x3	x1-x4	x2-x4	x3-x4	x1-x2	x1-x3	x2-x3	x1-x4	x2-x4	x3-x4	x1-x2	x1-x3	x2-x3	x1-x4	x2-x4	x3-x4	x1-x2	x1-x3	x2-x3	x1-x4	x2-x4	x3-x4
Modellkonfiguration			.2	.2	.2	.4	.4	.2	.2	.2	.2	.4	.4	.2	.2	.2	.2	.4	.4	.2	.2	.2	.2	.4	.4	.2
		mar2	0	0.0	-0.5	-0.2	0.0	-0.1	0	0.1	0.0	-0.1	0.0	0.3	0	-0.1	-0.2	-0.1	0.0	0.1	0	0.0	-0.1	-0.1	0.0	0.2
	skew1	mar3	0	-0.3	-1.2	-1.0	0.0	-0.5	0	0.1	0.2	-0.2	0.1	0.6	0	-0.5	-0.3	-0.5	-0.3	-0.4	0	-0.1	0.0	-0.3	0.0	0.1
		mar4	0	0.9	0.5	-0.1	0.0	-0.4	0	0.1	0.6	-0.3	-0.1	0.3	0	-1.0	-0.4	-0.6	-0.6	-1.3	0	-0.2	0.1	-0.4	-0.3	-0.4
		mar2	0	0.3	-0.1	0.3	0.4	1.4	0	0.3	0.3	0.4	0.3	1.2	0	0.1	0.3	0.3	0.0	0.6	0	0.2	0.3	0.4	0.2	1.1
750	skew2	mar3	0	0.7	-0.1	1.6	0.6	2.5	0	0.9	0.8	1.8	1.0	3.8	0	0.5	1.5	1.5	-0.2	1.2	0	0.7	0.6	1.8	0.9	3.5
		mar4	0	2.7	1.2	4.0	1.7	6.8	0	1.6	1.9	3.8	1.6	6.7	0	0.6	3.0	3.8	-0.4	1.7	0	1.3	1.5	3.7	1.5	6.0
		mar2	0	0.6	-0.6	0.4	0.4	1.4	0	0.5	0.4	0.8	0.4	1.8	0	0.6	1.0	0.9	0.0	1.1	0	0.5	0.4	0.8	0.3	1.7
	skew3	mar3	0	1.5	0.7	2.6	1.1	3.9	0	1.0	0.2	2.6	1.0	4.4	0	1.0	4.5	3.8	-0.3	2.1	0	0.9	0.3	2.7	1.0	4.1
		mar4	0	2.7	4.8	5.8	2.5	8.2	0	1.7	3.6	5.1	1.7	7.7	0	1.6	**12.5**	8.8	-0.6	4.1	0	1.6	3.4	5.4	1.7	7.4
		mar2	0	-0.6	-1.1	-0.6	0.1	-0.2	0	-0.1	0.1	-0.2	0.0	0.2	0	-0.3	-0.2	-0.4	-0.2	-0.1	0	-0.1	0.0	-0.3	0.0	0.1
	skew1	mar3	0	-1.1	-3.1	-1.2	0.1	-1.2	0	0.1	0.9	-0.2	-0.1	0.0	0	-0.9	-0.5	-0.7	-1.1	-1.8	0	-0.2	0.3	-0.4	-0.4	-0.6
		mar4	0	-0.5	-0.5	-0.6	0.0	-1.7	0	-0.4	1.4	0.0	-0.1	-0.3	0	-2.7	-1.7	-0.6	-1.8	-3.4	0	-0.9	0.4	-0.2	-0.6	-1.0
		mar2	0	0.3	-0.2	0.0	0.3	0.7	0	0.3	1.4	0.7	0.3	1.3	0	0.0	1.7	0.5	-0.2	0.4	0	0.2	1.1	0.7	0.2	1.1
250	skew2	mar3	0	1.3	-4.9	0.6	1.0	1.3	0	1.9	-0.2	2.5	0.8	3.0	0	0.8	2.0	2.6	-1.1	-0.3	0	1.6	-1.0	2.3	0.7	2.3
		mar4	0	3.3	-1.6	3.6	1.6	6.5	0	2.6	1.2	4.1	1.2	7.1	0	1.1	6.8	6.7	-2.1	2.5	0	2.6	0.9	4.4	1.1	6.6
		mar2	0	0.8	-1.3	-0.1	0.3	1.3	0	0.8	1.3	0.7	0.2	1.6	0	0.7	3.6	1.1	-0.3	1.3	0	0.8	1.3	0.7	0.3	1.4
	skew3	mar3	0	0.8	-1.6	1.6	0.7	2.6	0	1.7	2.4	2.5	1.0	4.6	0	1.9	**15.0**	6.5	-1.5	4.2	0	1.6	2.0	2.9	0.9	4.5
		mar4	0	1.3	0.3	5.1	1.4	3.8	0	1.4	2.4	4.4	0.9	6.1	0	3.0	**32.3**	**17.1**	-3.5	9.4	0	2.0	3.4	5.3	1.1	6.9

Anmerkungen: Fettdruck: Überschreitung des gesetzten Grenzwertes von ± 10 %. Rot und Fettdruck: Überschreitung des gesetzten Grenzwertes von ± 15 %.

Abbildung A6: Bereinigter relativer Parameterbias der Kovarianzen II

bereinigter relativer Bias in %: Kovarianzen																				
			H0						MNV						PMM					
			x1-x2	x1-x3	x2-x3	x1-x4	x2-x4	x3-x4	x1-x2	x1-x3	x2-x3	x1-x4	x2-x4	x3-x4	x1-x2	x1-x3	x2-x3	x1-x4	x2-x4	x3-x4
Modellkonfiguration			.2	.2	.2	.4	.4	.2	.2	.2	.2	.4	.4	.2	.2	.2	.2	.4	.4	.2
750	skew1	mar2	0	0.0	-0.1	-0.1	0.0	0.2	0	0.0	-0.4	-0.2	0.0	0.1	0	-0.1	-0.1	-0.2	0.0	0.2
		mar3	0	-0.1	-0.1	-0.3	-0.1	0.2	0	-0.3	0.3	-0.4	-0.2	0.0	0	-0.6	-0.4	-0.6	-0.2	-0.3
		mar4	0	-0.4	-0.2	-0.6	-0.4	0.1	0	-0.8	-0.4	-0.5	-0.6	-1.1	0	-1.0	-0.9	-0.9	-0.8	-1.2
	skew2	mar2	0	0.3	0.4	0.4	0.2	1.2	0	0.1	-0.2	0.3	0.2	1.0	0	-0.1	-0.1	0.0	-0.1	0.5
		mar3	0	0.7	0.9	1.8	0.7	4.2	0	0.6	1.0	1.7	0.7	3.3	0	-0.5	-0.6	0.4	-0.5	1.2
		mar4	0	1.2	1.7	3.6	1.1	7.1	0	0.7	1.0	3.4	1.1	4.9	0	-0.7	-0.1	1.1	-0.9	2.6
	skew3	mar2	0	0.6	0.6	0.9	0.4	2.2	0	0.4	-0.2	0.6	0.3	1.4	0	0.1	0.3	0.3	-0.2	1.1
		mar3	0	1.0	1.3	3.2	1.1	6.1	0	0.6	0.8	2.6	0.8	3.9	0	-0.4	0.7	0.6	-1.0	2.5
		mar4	0	1.5	4.7	5.9	1.7	10.0	0	0.7	2.9	5.2	1.3	6.0	0	-0.8	4.0	1.8	-1.5	5.8
250	skew1	mar2	0	-0.1	0.1	-0.3	0.0	0.1	0	-0.1	-0.1	-0.4	0.0	0.1	0	-0.3	-0.2	-0.4	-0.1	-0.1
		mar3	0	-0.3	0.4	-0.5	-0.4	-0.3	0	0.2	0.3	-0.9	-0.6	-1.3	0	-1.3	-0.9	-1.4	-1.1	-2.2
		mar4	0	-1.1	0.2	-0.6	-0.8	-0.1	0	-2.4	0.3	-1.1	-1.3	-3.1	0	-3.1	-2.5	-2.2	-2.0	-4.5
	skew2	mar2	0	0.2	1.3	0.7	0.1	1.2	0	0.3	1.1	0.5	0.2	0.9	0	-0.2	0.8	0.0	-0.2	0.3
		mar3	0	1.5	-0.5	2.4	0.3	3.2	0	1.9	-1.2	1.5	0.4	1.5	0	-0.3	-2.6	0.0	-1.1	-0.4
		mar4	0	2.3	1.7	4.0	0.4	7.5	0	0.7	0.9	3.1	0.3	4.2	0	-0.7	-1.0	0.2	-2.1	2.1
	skew3	mar2	0	0.7	1.4	0.8	0.2	1.9	0	0.9	1.2	0.5	0.2	1.4	0	0.4	1.3	-0.1	-0.4	1.2
		mar3	0	1.6	4.0	3.5	0.8	6.4	0	1.9	1.9	2.0	0.6	3.7	0	0.0	2.7	-0.3	-1.6	3.4
		mar4	0	1.2	6.3	5.8	0.9	9.7	0	-0.2	2.9	3.9	0.2	4.1	0	-0.4	6.1	-0.2	-3.2	7.2

A5.2 Modellbasierte Analyse: Absoluter Parameterbias[154]

Tabelle A14: Modellbasierte Analyse zum Parameterbias: Strukturpfade

	abh. Variable: durchschnittlicher absoluter Bias in den Strukturpfaden						
	b (95 % KI)						
	EM	EMB	FCS	FIML	H0	MNV	PMM
Constant	0.84***	3.21***	3.22***	0.63***	3.22***	3.21***	3.20***
	(0.79, 0.88)	(3.09, 3.33)	(3.10, 3.34)	(0.59, 0.67)	(3.10, 3.34)	(3.09, 3.33)	(3.08, 3.32)
250	0.42***	2.49***	2.49***	0.31***	2.38***	2.51***	2.52***
	(0.34, 0.50)	(2.30, 2.69)	(2.30, 2.68)	(0.24, 0.37)	(2.19, 2.57)	(2.32, 2.70)	(2.33, 2.71)
skew2	0.10**	0.17	0.17	0.08**	0.32***	0.18*	0.19*
	(0.03, 0.17)	(-0.01, 0.35)	(-0.01, 0.35)	(0.02, 0.14)	(0.14, 0.50)	(0.001, 0.36)	(0.01, 0.36)
skew3	0.20***	0.58***	0.57***	0.19***	0.61***	0.58***	0.62***
	(0.12, 0.29)	(0.39, 0.77)	(0.39, 0.76)	(0.12, 0.25)	(0.42, 0.79)	(0.39, 0.77)	(0.44, 0.81)
mar3	0.89***	0.24**	0.25**	0.67***	0.22*	0.24**	0.25**
	(0.81, 0.96)	(0.06, 0.42)	(0.07, 0.43)	(0.61, 0.73)	(0.05, 0.40)	(0.06, 0.42)	(0.08, 0.43)
mar4	1.38***	0.48***	0.47***	1.26***	0.55***	0.47***	0.53***
	(1.28, 1.49)	(0.29, 0.67)	(0.28, 0.66)	(1.17, 1.35)	(0.37, 0.74)	(0.28, 0.66)	(0.34, 0.72)
250:skew2	0.22***	0.33**	0.34**	0.20***	0.14	0.29**	0.29**
	(0.10, 0.34)	(0.10, 0.55)	(0.11, 0.56)	(0.10, 0.30)	(-0.08, 0.36)	(0.07, 0.52)	(0.07, 0.52)
250:skew3	0.61***	0.62***	0.61***	0.57***	0.69***	0.61***	0.52***
	(0.47, 0.75)	(0.38, 0.86)	(0.37, 0.85)	(0.45, 0.69)	(0.47, 0.92)	(0.38, 0.85)	(0.29, 0.76)
250:mar3	0.91***	0.26*	0.22	0.68***	0.15	0.24*	0.23*
	(0.81, 1.01)	(0.04, 0.48)	(-0.01, 0.44)	(0.60, 0.76)	(-0.07, 0.37)	(0.02, 0.46)	(0.01, 0.45)
250:mar4	1.70***	0.78***	0.69***	1.38***	0.42***	0.72***	0.66***
	(1.56, 1.85)	(0.54, 1.02)	(0.45, 0.93)	(1.26, 1.50)	(0.19, 0.65)	(0.48, 0.96)	(0.42, 0.90)
skew2:mar3	0.17**	0.10	0.09	0.28***	0.06	0.08	0.08
	(0.06, 0.27)	(-0.16, 0.36)	(-0.16, 0.35)	(0.20, 0.37)	(-0.20, 0.32)	(-0.17, 0.34)	(-0.18, 0.34)
skew3:mar3	0.44***	0.26	0.29*	0.57***	0.11	0.26	0.23
	(0.32, 0.57)	(-0.02, 0.53)	(0.02, 0.56)	(0.47, 0.67)	(-0.16, 0.37)	(-0.01, 0.53)	(-0.04, 0.50)
skew2:mar4	0.49***	0.23	0.25	0.42***	0.14	0.21	0.17
	(0.34, 0.64)	(-0.04, 0.51)	(-0.03, 0.52)	(0.30, 0.55)	(-0.14, 0.41)	(-0.07, 0.48)	(-0.11, 0.44)
skew3:mar4	1.17***	0.78***	0.83***	1.12***	0.43**	0.80***	0.64***
	(0.99, 1.35)	(0.49, 1.08)	(0.54, 1.13)	(0.97, 1.28)	(0.15, 0.71)	(0.50, 1.09)	(0.35, 0.93)
Fallzahl	9000	8999	9000	9000	9000	9000	9000
F-Statistik	993.774	378.795	371.12	1072.546	334.504	377.342	372.073
R^2	0.557	0.358	0.352	0.572	0.324	0.356	0.349
korr. R^2	0.556	0.357	0.351	0.572	0.323	0.356	0.349
SEE	1.349	2.344	2.342	1.133	2.258	2.328	2.318
Note:							*p<0.05; **p<0.01; ***p<0.001

[154] Anmerkung: F-Statistik in allen Modellen in A5.2: *** (p < 0.001).

Tabelle A15: Modellbasierte Analyse zum Parameterbias: Kovarianzen

	abh. Variable: durchschnittlicher absoluter Bias in den Kovarianzen						
	b (95 % KI)						
	EM	EMB	FCS	FIML	H0	MNV	PMM
Constant	0.94***	3.80***	3.79***	0.70***	3.76***	3.79***	3.76***
	(0.90, 0.98)	(3.68, 3.92)	(3.67, 3.92)	(0.66, 0.73)	(3.63, 3.88)	(3.66, 3.91)	(3.64, 3.89)
250	0.54***	2.80***	2.81***	0.40***	3.13***	2.81***	2.85***
	(0.46, 0.61)	(2.61, 2.99)	(2.62, 3.00)	(0.34, 0.47)	(2.94, 3.33)	(2.63, 3.00)	(2.67, 3.04)
skew2	0.16***	0.20*	0.22*	0.14***	0.25**	0.21*	0.23*
	(0.09, 0.23)	(0.02, 0.38)	(0.05, 0.40)	(0.09, 0.20)	(0.07, 0.43)	(0.04, 0.39)	(0.05, 0.40)
skew3	0.27***	0.58***	0.59***	0.26***	0.65***	0.59***	0.62***
	(0.19, 0.35)	(0.39, 0.77)	(0.40, 0.78)	(0.19, 0.32)	(0.46, 0.85)	(0.40, 0.78)	(0.43, 0.81)
mar3	0.94***	0.18	0.18*	0.73***	0.24**	0.18*	0.20*
	(0.87, 1.01)	(-0.0005, 0.36)	(0.01, 0.36)	(0.67, 0.79)	(0.06, 0.43)	(0.005, 0.36)	(0.02, 0.38)
mar4	1.74***	0.53***	0.54***	1.47***	0.60***	0.55***	0.59***
	(1.64, 1.84)	(0.34, 0.72)	(0.35, 0.73)	(1.38, 1.55)	(0.41, 0.79)	(0.36, 0.74)	(0.40, 0.77)
250:skew2	0.26***	0.25*	0.19	0.21***	-0.04	0.22*	0.18
	(0.14, 0.37)	(0.03, 0.46)	(-0.03, 0.40)	(0.12, 0.31)	(-0.26, 0.18)	(0.01, 0.44)	(-0.04, 0.39)
250:skew3	0.61***	0.60***	0.58***	0.57***	0.31*	0.58***	0.46***
	(0.48, 0.74)	(0.37, 0.84)	(0.34, 0.81)	(0.45, 0.68)	(0.07, 0.55)	(0.34, 0.81)	(0.23, 0.69)
250:mar3	0.87***	0.25*	0.23*	0.75***	0.22	0.22*	0.17
	(0.77, 0.97)	(0.03, 0.47)	(0.01, 0.46)	(0.66, 0.83)	(-0.004, 0.44)	(0.002, 0.44)	(-0.05, 0.39)
250:mar4	1.44***	0.54***	0.47***	1.32***	0.47***	0.44***	0.34**
	(1.31, 1.57)	(0.31, 0.77)	(0.24, 0.70)	(1.21, 1.43)	(0.23, 0.71)	(0.21, 0.67)	(0.11, 0.57)
skew2:mar3	0.25***	0.15	0.14	0.23***	0.08	0.13	0.06
	(0.14, 0.35)	(-0.11, 0.40)	(-0.11, 0.40)	(0.15, 0.32)	(-0.18, 0.35)	(-0.12, 0.38)	(-0.19, 0.32)
skew3:mar3	0.64***	0.35*	0.38**	0.68***	0.29*	0.35*	0.15
	(0.52, 0.77)	(0.07, 0.62)	(0.11, 0.66)	(0.58, 0.78)	(0.01, 0.56)	(0.08, 0.62)	(-0.12, 0.42)
skew2:mar4	0.40***	0.30*	0.28*	0.50***	0.29*	0.25	0.08
	(0.26, 0.54)	(0.03, 0.57)	(0.02, 0.55)	(0.38, 0.62)	(0.01, 0.56)	(-0.02, 0.51)	(-0.18, 0.35)
skew3:mar4	1.15***	0.69***	0.85***	1.22***	0.58***	0.68***	0.26
	(0.99, 1.32)	(0.40, 0.98)	(0.56, 1.13)	(1.07, 1.36)	(0.29, 0.88)	(0.40, 0.97)	(-0.02, 0.54)
Fallzahl	9000	8999	9000	9000	9000	9000	9000
F-Statistik	1293.506	435.356	434.362	1384.276	445.47	429.674	411.861
R^2	0.611	0.386	0.385	0.63	0.387	0.382	0.37
korr. R^2	0.611	0.385	0.384	0.629	0.386	0.381	0.369
SEE	1.259	2.295	2.291	1.094	2.331	2.284	2.234
Note:							*p<0.05; **p<0.01; ***p<0.001

Tabelle A16: Test auf Unterschiede in den b-Koeffizienten: Strukturpfade

Differenzen der b-Koeffizienten zwischen den MDTs mit 5 %iger Irrtumswahrscheinlichkeit signifikant?

Effekt	MNV vs. EMB	MNV vs. FCS	MNV vs. H0	MNV vs. PMM	MI (MNV) vs. EM	MI (MNV) vs. FIML	FIML vs. EM
250	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
skew2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
skew3	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
mar3	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
mar4	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
250:skew2	--	--	--	--	--	--	--
250:skew3	--	--	--	--	--	--	--
250:mar3	--	--	--	--	--	--	TRUE
250:mar4	--	--	--	--	--	--	TRUE
skew2:mar3	--	--	--	--	--	--	--
skew3:mar3	--	--	--	--	--	--	--
skew2:mar4	--	--	--	--	--	--	--
skew3:mar4	--	--	--	--	--	--	--

Anmerkungen: FALSE: Unterschied nicht signifikant; TRUE: signifikanter Unterschied; -- kein als substantiell einzuordnender Effekt (ΔR^2) und deshalb auch nicht geprüft bzw. irrelevant.

Tabelle A17: Test auf Unterschiede in den b-Koeffizienten: Kovarianzen

Differenzen der b-Koeffizienten zwischen den MDTs mit 5 %iger Irrtumswahrscheinlichkeit signifikant?

Effekt	MNV vs. EMB	MNV vs. FCS	MNV vs. H0	MNV vs. PMM	MI (MNV) vs. EM	MI (MNV) vs. FIML	FIML vs. EM
250	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
skew2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
skew3	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
mar3	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
mar4	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
250:skew2	--	--	--	--	--	--	--
250:skew3	--	--	--	--	--	--	--
250:mar3	--	--	--	--	--	--	FALSE
250:mar4	--	--	--	--	--	--	FALSE
skew2:mar3	--	--	--	--	--	--	--
skew3:mar3	--	--	--	--	--	--	--
skew2:mar4	--	--	--	--	--	--	--
skew3:mar4	--	--	--	--	--	--	--

Anmerkungen: FALSE: Unterschied nicht signifikant; TRUE: signifikanter Unterschied; -- kein als substantiell einzuordnender Effekt (ΔR^2) und deshalb auch nicht geprüft bzw. irrelevant.

A6 Standardfehlerbias

A6.1 Deskriptive Auswertung: Relativer Standardfehlerbias

Abbildung A7: Relativer Standardfehlerbias der Faktorladungen

relativer SE-Bias in %: Faktorladungen																															
			EM				EMB				FCS				FIML				H0				MNV				PMM			
			ind2	ind3	ind4	ind5	ind2	ind3	ind4	ind5	ind2	ind3	ind4	ind5	ind2	ind3	ind4	ind5	ind2	ind3	ind4	ind5	ind2	ind3	ind4	ind5	ind2	ind3	ind4	ind5
Modellkonfiguration			.8	.7	.6	.5	.8	.7	.6	.5	.8	.7	.6	.5	.8	.7	.6	.5	.8	.7	.6	.5	.8	.7	.6	.5	.8	.7	.6	.5
750	skew1	mar2	-4.6	-3.2	-5.3	-3.2	0.9	1.3	1.1	2.5	1.0	1.4	0.7	3.5	1.0	1.2	0.9	2.8	0.9	1.1	1.1	3.0	1.1	1.1	1.2	3.0	0.6	1.2	0.6	2.9
		mar3	**-22.8**	**-19.7**	**-20.0**	**-18.2**	-0.5	3.6	-0.3	1.3	0.6	4.3	1.1	2.7	0.0	4.1	0.8	2.3	0.0	4.1	0.6	2.0	0.2	4.4	2.0	2.7	-0.5	5.3	1.2	2.6
		mar4	**-33.1**	**-30.1**	**-32.9**	**-29.7**	-1.5	2.3	-4.2	3.2	-0.5	4.8	-2.7	5.0	-0.3	2.0	-3.4	4.3	-0.3	2.7	-3.4	4.8	-0.4	2.7	-2.5	4.0	-0.7	3.7	-2.1	4.4
	skew2	mar2	-6.8	-7.1	-6.4	-6.0	-2.4	0.8	1.8	1.7	-2.7	0.8	1.5	1.9	-1.6	1.6	3.0	3.0	-2.6	0.6	1.8	2.2	-2.3	0.5	1.9	2.0	-2.1	1.5	2.8	3.2
		mar3	**-27.4**	**-19.2**	**-25.9**	**-23.0**	-7.2	3.7	-2.9	2.2	-7.5	3.2	-4.4	0.9	-3.7	5.9	-0.5	4.2	-7.0	2.3	-3.9	1.4	-6.9	3.3	-3.4	1.7	-5.5	5.2	-2.5	3.2
		mar4	**-37.8**	**-32.7**	**-38.2**	**-36.8**	-6.5	3.8	-2.4	-1.1	-8.5	2.3	-6.1	-2.5	-2.5	5.9	0.4	3.8	-7.3	0.5	-5.0	-0.1	-8.2	0.9	-3.7	-1.1	-5.4	5.7	-3.4	3.4
	skew3	mar2	-8.0	-4.1	-7.2	-9.0	-1.6	3.6	-1.5	-1.4	-1.5	3.9	-1.6	-0.6	1.2	6.4	1.3	1.0	-1.0	3.8	-1.4	-1.7	-1.2	3.7	-1.5	-1.3	0.3	5.8	1.0	0.8
		mar3	**-27.8**	**-25.2**	**-31.2**	**-30.1**	-8.3	-1.7	-7.4	-4.2	-9.2	-2.1	-9.9	-5.7	-1.8	4.6	-0.2	1.7	**-10.3**	-3.6	-9.4	-5.7	**-10.4**	-3.1	-8.5	-5.7	-3.0	2.7	-3.0	1.3
		mar4	**-44.2**	**-38.0**	**-45.9**	**-44.0**	**-12.1**	-4.1	-7.7	-2.3	**-17.3**	-5.1	**-15.2**	-8.1	-6.7	2.2	-0.6	4.4	**-18.6**	**-11.3**	**-14.4**	-7.8	**-17.7**	-8.1	**-14.3**	-7.8	-8.5	0.5	-4.7	4.1
250	skew1	mar2	-4.1	-5.2	-2.0	-1.6	1.6	0.1	3.7	2.6	2.4	1.6	4.4	4.1	2.1	0.8	4.4	3.3	2.2	1.1	4.4	3.5	2.1	0.9	4.3	3.2	1.8	0.5	3.8	3.0
		mar3	**-18.4**	**-22.6**	**-19.8**	**-15.4**	-0.3	-1.7	3.0	4.0	3.6	1.4	6.1	7.4	0.6	-0.9	3.5	4.9	0.8	-0.3	4.9	4.8	2.1	1.0	5.8	6.3	-0.9	-0.5	3.2	5.4
		mar4	**-33.9**	**-32.2**	**-29.9**	**-33.7**	0.0	-2.6	2.1	1.5	5.4	2.9	6.3	5.0	0.3	-2.6	3.4	1.7	1.8	-1.3	4.8	3.3	3.3	-0.1	5.6	3.7	1.3	-1.8	4.4	2.8
	skew2	mar2	**-11.6**	-7.6	-7.4	-5.2	-5.2	-3.0	-0.6	-0.2	-5.0	-2.1	-0.1	0.8	-4.2	-2.0	1.0	1.0	-4.7	-2.8	0.0	-0.1	-5.2	-2.2	0.3	0.2	-4.7	-1.5	-0.2	0.7
		mar3	**-27.3**	**-23.1**	**-23.4**	**-22.9**	-7.4	-1.7	0.9	0.5	-7.2	1.2	0.7	3.3	-5.8	1.2	2.3	3.4	-7.4	-1.2	-0.2	0.7	-6.6	-0.4	2.1	1.9	-6.7	3.8	-0.1	5.3
		mar4	**-41.1**	**-35.4**	**-37.2**	**-40.7**	-8.0	-2.4	2.4	-4.8	-6.1	-1.1	1.5	-2.5	-6.6	-0.3	2.3	-2.8	-8.8	-3.4	-0.9	-6.3	-8.7	-2.3	1.0	-4.8	-6.4	0.0	3.5	-0.2
	skew3	mar2	**-12.9**	-6.1	**-10.8**	-8.1	-4.9	0.4	-3.5	0.5	-5.4	1.6	-3.1	1.9	-2.5	2.7	-1.0	3.4	-5.1	0.4	-3.3	0.9	-5.3	0.4	-3.1	1.2	-2.9	2.4	-1.3	3.4
		mar3	**-28.8**	**-22.3**	**-32.2**	**-27.1**	-7.9	-1.1	-6.3	1.4	-8.2	-1.0	-8.4	2.2	-2.8	5.0	-1.4	6.2	**-10.6**	-4.1	-9.5	-2.1	-8.8	-1.9	-7.4	-0.4	-3.8	3.7	-2.1	8.7
		mar4	**-43.9**	**-41.9**	**-46.3**	**-43.6**	-5.4	-4.5	-4.3	1.3	-7.5	-8.0	-8.3	-4.3	-2.5	-1.3	-1.7	0.9	**-13.3**	**-13.2**	**-12.9**	-9.6	**-11.3**	**-10.3**	**-10.8**	-7.7	-3.2	-3.5	-1.6	3.5

Anmerkungen: Fettdruck: Überschreitung des gesetzten Grenzwertes von ± 10 %. Rot und Fettdruck: Überschreitung des gesetzten Grenzwertes von ± 15 %.

Abbildung A8: Relativer Standardfehlerbias der Strukturpfade

relativer SE-Bias in %: Strukturpfade																															
			EM				EMB				FCS				FIML				H0				MNV				PMM				
			x1>f1	x2>f1	x3>f1	x4>f1	x1>f1	x2>f1	x3>f1	x4>f1	x1>f1	x2>f1	x3>f1	x4>f1	x1>f1	x2>f1	x3>f1	x4>f1	x1>f1	x2>f1	x3>f1	x4>f1	x1>f1	x2>f1	x3>f1	x4>f1	x1>f1	x2>f1	x3>f1	x4>f1	
Modellkonfiguration			.1	.3	.3	.5	.1	.3	.3	.5	.1	.3	.3	.5	.1	.3	.3	.5	.1	.3	.3	.5	.1	.3	.3	.5	.1	.3	.3	.5	
750		mar2	-2.1	0.0	**-10.3**	2.3	3.4	2.1	-3.4	7.5	3.2	1.9	-3.3	7.3	3.7	1.9	-3.3	7.5	3.5	2.0	-3.2	7.6	3.0	2.0	-3.1	7.6	3.4	1.4	-3.4	6.7	
	skew1	mar3	**-18.5**	**-16.3**	**-24.5**	**-14.0**	-2.5	-1.3	-2.3	6.6	-3.1	-1.1	-1.7	5.6	-1.7	-0.4	-1.9	6.7	-2.0	-0.6	-1.3	6.7	-1.2	-0.6	-1.0	7.8	-1.8	-0.8	-2.2	5.7	
		mar4	**-29.1**	**-23.4**	**-36.8**	**-28.0**	-4.8	4.7	0.4	5.8	-5.5	2.8	-0.7	2.7	-2.9	4.7	0.1	5.1	-3.1	5.4	0.2	6.4	-3.3	4.1	0.9	6.9	-4.0	4.4	-0.6	4.3	
		mar2	-6.3	-4.4	**-11.0**	-1.2	1.2	-1.8	-2.8	4.1	1.4	-1.7	-2.6	4.3	1.9	-1.0	-2.2	4.9	1.7	-1.1	-2.2	4.8	1.1	-1.4	-2.9	4.0	2.7	-0.7	-2.9	4.8	
	skew2	mar3	**-27.6**	**-18.2**	**-26.1**	**-22.2**	-6.9	-1.8	-2.4	-0.8	-8.3	-2.3	-2.8	-2.6	-5.0	-0.1	-0.8	1.0	-5.8	-1.7	-0.6	1.1	-4.6	-1.8	-2.7	-0.5	-3.9	0.2	-1.0	-0.3	
		mar4	**-41.1**	**-26.4**	**-44.2**	**-32.8**	-8.5	1.9	-7.4	-1.2	**-12.2**	0.1	-8.4	-7.6	-6.8	4.0	-5.9	0.9	-8.1	1.5	-4.7	0.7	-9.2	2.0	-8.7	-1.6	-6.2	4.5	-3.9	-1.6	
		mar2	-7.6	-9.3	-9.2	-3.2	1.2	-5.4	-1.9	1.4	1.5	-4.8	-0.2	2.0	3.3	-3.9	0.6	3.6	2.4	-4.8	1.0	3.3	0.7	-5.6	-1.9	1.2	4.3	-3.8	0.8	2.8	
	skew3	mar3	**-32.2**	**-24.3**	**-28.2**	**-25.5**	-5.7	-8.2	-7.2	-4.8	**-10.5**	-7.4	-3.7	-7.3	-1.5	-3.8	-0.4	1.5	-4.3	-6.8	0.9	0.6	-6.8	-8.8	-8.9	-5.8	-2.0	-3.0	-0.3	-1.6	
		mar4	**-48.0**	**-34.7**	**-48.4**	**-38.4**	-7.3	-7.7	-9.7	-6.2	**-15.4**	-8.8	**-10.6**	**-12.1**	-4.3	-2.2	-3.0	0.3	-7.5	-6.6	-2.1	-0.5	**-12.7**	**-10.6**	**-15.3**	**-10.3**	-4.5	-1.2	-3.4	-2.4	
250		mar2	-6.7	-3.7	-9.9	-6.2	-2.7	1.1	-3.0	-0.5	-2.3	1.3	-2.2	0.0	-2.3	1.2	-2.5	-0.2	-2.1	1.5	-2.5	-0.1	-2.8	1.2	-2.6	-0.4	-2.4	1.3	-2.4	-0.2	
	skew1	mar3	**-23.2**	**-19.2**	**-26.5**	**-26.5**	-7.0	-1.6	-4.5	-5.0	-5.5	-0.9	-2.0	-3.1	-5.8	-1.2	-3.2	-4.4	-4.9	-0.7	-2.1	-2.9	-5.9	-0.7	-2.1	-3.1	-4.8	-0.7	-2.3	-4.0	
		mar4	**-33.0**	**-28.2**	**-39.8**	**-38.6**	-4.1	-4.0	-5.0	-6.7	-3.3	-2.5	-3.9	-5.4	-4.5	-2.8	-4.5	-6.1	-3.5	-1.2	-2.0	-3.5	-2.7	-1.5	-1.9	-4.6	-1.8	-1.9	-3.2	-4.8	
		mar2	**-11.5**	-9.3	-9.9	-6.8	-3.1	-4.0	-3.2	-0.9	-2.6	-3.8	-1.6	-0.8	-2.0	-3.5	-2.2	0.2	-2.2	-3.6	-1.4	0.2	-3.0	-3.8	-2.4	-0.2	-2.1	-3.2	-1.5	0.4	
	skew2	mar3	**-27.2**	**-20.5**	**-25.6**	**-27.2**	-5.3	-4.8	-1.9	-5.3	-5.3	-3.4	0.3	-5.7	-4.3	-3.2	0.8	-3.4	-4.3	-4.0	2.0	-3.1	-4.4	-3.9	-0.3	-4.4	-1.6	-1.5	2.4	-2.7	
		mar4	**-40.0**	**-33.5**	**-44.7**	**-38.7**	-1.4	-5.4	-6.0	-3.6	-5.9	-4.4	-5.6	-7.3	-3.0	-3.5	-4.6	-3.5	-3.5	-4.6	-2.0	-1.0	-3.6	-5.0	-6.2	-3.6	-0.7	-1.3	-1.3	-2.2	
		mar2	**-12.1**	-8.1	**-13.5**	-7.8	-2.6	-2.9	-7.4	-1.8	-2.0	-1.6	-5.1	-0.6	-1.1	-1.3	-4.9	0.3	-1.6	-1.7	-4.7	0.8	-2.8	-2.8	-6.9	-1.8	0.4	-0.5	-4.5	0.7	
	skew3	mar3	**-33.1**	**-21.6**	**-33.7**	**-28.0**	-4.1	-3.5	**-10.4**	-7.2	-6.3	-1.9	-5.0	-6.1	-0.8	0.1	-3.9	-1.7	-2.0	-3.0	-1.8	-1.1	-6.0	-4.9	**-11.4**	-8.3	1.4	1.7	-2.5	-1.2	
		mar4	**-47.1**	**-34.1**	**-51.2**	**-42.2**	-2.8	-1.5	**-12.3**	-6.6	-9.8	-1.2	-8.9	**-10.5**	-3.2	0.1	-9.7	-4.5	-5.5	-2.9	-7.0	-3.6	-9.9	-5.5	**-18.1**	**-12.4**	-0.9	4.1	-5.8	-3.9	

Anmerkungen: Fettdruck: Überschreitung des gesetzten Grenzwertes von ± 10 %. Rot und Fettdruck: Überschreitung des gesetzten Grenzwertes von ± 15 %.

Abbildung A9: Relativer Standardfehlerbias der Kovarianzen I

relativer SE-Bias in %: Kovarianzen																								
			EM						EMB						FCS						FIML					
Modellkonfiguration			x1-x2 .2	x1-x3 .2	x2-x3 .2	x1-x4 .4	x2-x4 .4	x3-x4 .2	x1-x2 .2	x1-x3 .2	x2-x3 .2	x1-x4 .4	x2-x4 .4	x3-x4 .2	x1-x2 .2	x1-x3 .2	x2-x3 .2	x1-x4 .4	x2-x4 .4	x3-x4 .2	x1-x2 .2	x1-x3 .2	x2-x3 .2	x1-x4 .4	x2-x4 .4	x3-x4 .2
750	skew1	mar2	4.6	0.5	-3.2	-5.6	0.1	-3.8	4.6	4.4	1.1	0.1	3.1	3.5	4.6	4.6	1.4	-0.1	3.2	4.1	4.6	4.6	1.2	0.5	3.3	4.2
		mar3	4.6	**-14.5**	**-21.6**	**-17.2**	**-11.1**	**-20.9**	4.6	1.3	-0.6	-0.1	2.8	1.4	4.6	2.5	-0.3	-3.6	1.4	0.7	4.6	2.3	-1.1	0.2	3.0	2.2
		mar4	4.6	**-28.8**	**-29.8**	**-27.8**	**-19.9**	**-39.6**	4.6	-0.4	0.7	-3.6	-0.1	-5.1	4.6	-0.4	0.7	**-11.9**	-4.3	-8.3	4.6	0.0	0.1	-3.1	0.9	-4.1
	skew2	mar2	1.7	-4.3	-9.4	-7.9	-6.5	1.6	1.7	1.2	1.3	-0.4	-2.3	8.1	1.7	1.7	1.8	-1.5	-2.1	8.3	1.7	1.6	1.7	-0.3	-2.1	8.7
		mar3	1.7	**-18.3**	**-28.8**	**-24.7**	**-14.5**	**-18.8**	1.7	-0.8	-2.0	-2.6	-1.7	5.2	1.7	0.2	-0.7	**-10.0**	-2.4	3.7	1.7	-0.2	-0.7	-1.3	-0.8	6.0
		mar4	1.7	**-31.9**	**-40.3**	**-35.6**	**-26.3**	**-38.5**	1.7	-4.5	-1.8	-3.7	-5.2	1.2	1.7	-2.9	-2.7	**-17.6**	-9.5	-3.0	1.7	-2.4	-1.6	-2.3	-3.6	1.4
	skew3	mar2	4.8	-7.6	**-16.9**	**-10.1**	-3.5	-9.8	4.8	-3.4	-6.0	-2.4	-0.3	-1.0	4.8	-2.7	-3.5	-2.0	0.1	-0.1	4.8	-2.5	-3.5	-0.2	0.6	0.4
		mar3	4.8	**-19.5**	**-39.5**	**-33.7**	**-10.8**	**-29.0**	4.8	-4.9	-8.1	-5.9	-1.3	-3.2	4.8	-1.8	-4.0	**-16.1**	-1.4	-0.7	4.8	-1.9	-5.3	-1.8	1.6	1.5
		mar4	4.8	**-30.7**	**-52.5**	**-47.0**	**-22.7**	**-45.2**	4.8	-4.6	-8.0	-7.9	-1.7	-4.3	4.8	0.2	-3.4	**-28.7**	-5.7	-7.3	4.8	-0.4	-5.6	-4.0	2.7	-1.3
250	skew1	mar2	1.9	-6.8	**-10.6**	-9.7	-5.1	-4.9	1.9	-2.6	-3.3	-4.1	-1.4	3.3	1.9	-2.3	-2.2	-4.5	-1.0	4.1	1.9	-2.1	-2.4	-3.9	-1.1	4.1
		mar3	1.9	**-19.6**	**-23.9**	**-19.2**	**-13.6**	**-25.4**	1.9	-3.9	-5.2	-3.2	-2.0	-3.2	1.9	-3.1	-3.5	-4.5	-2.0	-3.3	1.9	-2.9	-3.6	-2.9	-1.6	-1.3
		mar4	1.9	**-28.2**	**-32.2**	**-26.0**	**-23.9**	**-39.5**	1.9	-4.0	-2.8	-2.5	-4.6	-5.2	1.9	-2.9	-2.5	-9.0	-5.9	-5.8	1.9	-2.6	-3.0	-2.0	-3.6	-3.2
	skew2	mar2	-2.7	-5.6	**-12.6**	**-10.9**	-5.4	-8.2	-2.7	-2.5	-3.1	-5.7	-0.7	-0.3	-2.7	-1.9	-1.4	-5.7	-0.5	-0.1	-2.7	-1.8	-2.3	-5.4	-0.6	0.1
		mar3	-2.7	**-19.5**	**-31.5**	**-27.7**	**-14.5**	**-30.3**	-2.7	-2.6	-4.0	-4.7	-1.5	-3.7	-2.7	0.1	-1.6	**-11.1**	-2.7	-4.6	-2.7	-0.9	-3.1	-3.8	-0.6	-1.7
		mar4	-2.7	**-28.0**	**-43.9**	**-35.6**	**-24.2**	**-39.7**	-2.7	-4.7	-7.6	-1.0	-2.1	-4.3	-2.7	0.0	-4.9	**-13.7**	-6.3	-7.2	-2.7	-1.8	-7.5	-0.9	0.0	-3.3
	skew3	mar2	-1.4	-6.6	**-18.6**	**-16.1**	-4.7	**-13.7**	-1.4	-2.5	-6.5	-7.5	-1.1	-6.3	-1.4	-0.9	-2.9	-7.4	-0.6	-4.1	-1.4	-1.6	-4.2	-6.0	-0.2	-5.1
		mar3	-1.4	**-23.0**	**-40.2**	**-35.7**	**-13.5**	**-31.1**	-1.4	-6.6	-7.3	-8.5	-2.5	-7.3	-1.4	-0.6	-2.9	**-17.2**	-2.7	-3.5	-1.4	-3.6	-5.4	-5.0	0.1	-4.5
		mar4	-1.4	**-30.8**	**-53.2**	**-46.3**	**-24.3**	**-46.2**	-1.4	-4.4	-6.6	-4.3	-3.3	-5.3	-1.4	4.9	-1.2	**-23.8**	-4.5	0.2	-1.4	-1.1	-7.5	-3.6	1.3	-5.8

Anmerkungen: Fettdruck: Überschreitung des gesetzten Grenzwertes von ± 10 %. Rot und Fettdruck: Überschreitung des gesetzten Grenzwertes von ± 15 %.

Abbildung A10: Relativer Standardfehlerbias der Kovarianzen II

relativer SE-Bias in %: Kovarianzen																				
			H0						MNV						PMM					
			x1-x2	x1-x3	x2-x3	x1-x4	x2-x4	x3-x4	x1-x2	x1-x3	x2-x3	x1-x4	x2-x4	x3-x4	x1-x2	x1-x3	x2-x3	x1-x4	x2-x4	x3-x4
Modellkonfiguration			.2	.2	.2	.4	.4	.2	.2	.2	.2	.4	.4	.2	.2	.2	.2	.4	.4	.2
750	skew1	mar2	4.6	4.4	1.1	0.4	3.5	4.3	4.6	4.7	1.4	0.4	3.3	4.1	4.6	4.6	1.7	0.8	3.1	3.5
		mar3	4.6	2.1	-1.0	0.3	2.3	1.4	4.6	2.7	-0.2	0.5	3.1	2.8	4.6	2.7	-0.2	0.5	3.1	3.0
		mar4	4.6	0.4	0.5	-2.9	0.8	-3.4	4.6	-0.2	0.5	-2.0	1.4	-4.4	4.6	-0.5	1.0	-4.0	1.1	-4.6
	skew2	mar2	1.7	1.6	2.0	0.0	-2.2	9.2	1.7	1.2	1.7	-0.6	-2.2	8.4	1.7	2.0	3.1	0.2	-1.5	9.6
		mar3	1.7	-0.1	-0.4	-1.0	-0.8	6.0	1.7	-0.5	-1.4	-2.3	-1.3	6.2	1.7	1.1	1.8	-0.7	0.4	8.6
		mar4	1.7	-2.2	-2.7	-3.1	-4.0	1.6	1.7	-3.1	-3.4	-2.5	-4.5	1.3	1.7	-0.4	1.1	-2.0	-3.5	5.8
	skew3	mar2	4.8	-2.3	-3.0	0.7	0.9	2.2	4.8	-3.3	-5.1	-2.7	-0.1	-1.0	4.8	-1.8	-2.0	1.1	0.9	1.6
		mar3	4.8	-1.1	-5.5	-1.1	2.1	4.4	4.8	-5.1	**-11.1**	-7.7	-1.1	-3.1	4.8	1.3	-2.0	-1.8	3.5	9.7
		mar4	4.8	-0.1	-4.5	-4.0	4.0	1.8	4.8	-5.3	**-12.1**	**-12.0**	-2.5	-6.8	4.8	4.2	-2.4	-5.8	4.0	8.3
250	skew1	mar2	1.9	-2.1	-2.5	-3.7	-0.9	3.9	1.9	-2.1	-2.0	-3.9	-1.2	3.9	1.9	-2.1	-2.0	-3.6	-1.0	4.0
		mar3	1.9	-3.1	-4.1	-2.9	-1.6	-1.9	1.9	-1.9	-2.6	-1.8	-1.8	-1.0	1.9	-3.0	-3.2	-2.3	-1.1	-1.9
		mar4	1.9	-2.9	-2.8	-2.1	-3.1	-4.4	1.9	-1.9	-1.6	-0.7	-3.2	-2.3	1.9	-1.5	-2.0	-0.5	-2.7	-2.2
	skew2	mar2	-2.7	-1.9	-2.5	-5.0	-0.4	0.2	-2.7	-2.0	-2.5	-5.9	-0.6	-0.1	-2.7	-1.8	-0.9	-5.1	0.1	0.9
		mar3	-2.7	-0.4	-3.3	-3.4	-0.4	-0.9	-2.7	-0.5	-2.8	-3.9	-0.7	-2.1	-2.7	0.0	-0.2	-2.3	0.4	0.6
		mar4	-2.7	-2.3	-8.2	0.0	0.2	-2.5	-2.7	-2.5	-7.2	0.3	0.1	-3.8	-2.7	1.4	-3.4	1.5	1.9	0.6
	skew3	mar2	-1.4	-1.3	-3.7	-5.7	0.2	-3.3	-1.4	-2.2	-5.5	-8.1	-1.0	-6.1	-1.4	-0.5	-1.5	-5.4	0.3	-3.0
		mar3	-1.4	-3.2	-5.8	-3.3	0.7	-0.4	-1.4	-5.1	-8.5	**-10.1**	-1.9	-6.6	-1.4	0.2	-0.6	-3.8	1.8	2.7
		mar4	-1.4	-0.3	-8.4	-2.1	2.8	-3.1	-1.4	-2.8	**-10.5**	-9.2	-3.3	-7.3	-1.4	5.2	-1.6	-2.0	3.5	6.3

Anmerkungen: Fettdruck: Überschreitung des gesetzten Grenzwertes von ± 10 %.

A6.2 Modellbasierte Analyse: Absoluter Standardfehlerbias[155]

Tabelle A18: Modellbasierte Analyse zum Standardfehlerbias: Strukturpfade

	abh. Variable: durchschnittlicher absoluter SE-Bias in den Strukturpfaden						
	b (95 % KI)						
	EM	EMB	FCS	FIML	H0	MNV	PMM
Constant	0.17***	0.14***	0.13***	0.15***	0.14***	0.14***	0.14***
	(0.16, 0.18)	(0.13, 0.15)	(0.12, 0.13)	(0.14, 0.15)	(0.13, 0.14)	(0.14, 0.15)	(0.13, 0.14)
250	0.25***	0.15***	0.17***	0.15***	0.15***	0.14***	0.15***
	(0.23, 0.27)	(0.14, 0.17)	(0.16, 0.18)	(0.14, 0.16)	(0.14, 0.17)	(0.13, 0.15)	(0.14, 0.16)
skew2	0.08***	0.003	0.01	0.004	-0.01	-0.01	0.01
	(0.07, 0.10)	(-0.01, 0.01)	(-0.005, 0.02)	(-0.01, 0.01)	(-0.02, 0.004)	(-0.02, 0.003)	(-0.004, 0.02)
skew3	0.06***	-0.005	0.02***	-0.01	0.01**	-0.02*	0.01
	(0.04, 0.08)	(-0.02, 0.01)	(0.01, 0.04)	(-0.02, 0.003)	(0.004, 0.03)	(-0.03, -0.004)	(-0.01, 0.02)
mar3	0.50***	0.03***	0.03***	0.01	0.02***	0.02**	0.02***
	(0.49, 0.52)	(0.02, 0.04)	(0.02, 0.04)	(-0.001, 0.02)	(0.01, 0.03)	(0.01, 0.03)	(0.01, 0.03)
mar4	0.96***	0.07***	0.08***	0.02**	0.07***	0.05***	0.07***
	(0.94, 0.97)	(0.06, 0.08)	(0.07, 0.09)	(0.01, 0.03)	(0.06, 0.08)	(0.04, 0.07)	(0.06, 0.08)
250:skew2	0.04**	0.03***	0.03**	0.03***	0.04***	0.05***	0.03***
	(0.01, 0.06)	(0.01, 0.05)	(0.01, 0.04)	(0.02, 0.05)	(0.02, 0.05)	(0.03, 0.06)	(0.01, 0.04)
250:skew3	0.30***	0.18***	0.10***	0.19***	0.11***	0.20***	0.14***
	(0.27, 0.33)	(0.16, 0.20)	(0.08, 0.12)	(0.17, 0.21)	(0.09, 0.13)	(0.18, 0.22)	(0.12, 0.16)
250:mar3	0.56***	0.08***	0.03***	0.09***	0.03***	0.06***	0.05***
	(0.54, 0.59)	(0.06, 0.10)	(0.02, 0.05)	(0.08, 0.11)	(0.02, 0.05)	(0.04, 0.08)	(0.04, 0.07)
250:mar4	1.31***	0.24***	0.15***	0.25***	0.12***	0.19***	0.17***
	(1.29, 1.34)	(0.22, 0.26)	(0.13, 0.17)	(0.23, 0.27)	(0.10, 0.14)	(0.17, 0.21)	(0.15, 0.19)
skew2:mar3	0.12***	0.02**	0.05***	0.01	0.03***	0.02*	0.01
	(0.10, 0.15)	(0.01, 0.04)	(0.03, 0.06)	(-0.002, 0.03)	(0.01, 0.04)	(0.003, 0.03)	(-0.01, 0.02)
skew3:mar3	0.48***	0.12***	0.15***	0.03***	0.04***	0.17***	0.03***
	(0.45, 0.51)	(0.10, 0.14)	(0.13, 0.16)	(0.02, 0.05)	(0.02, 0.06)	(0.15, 0.19)	(0.01, 0.05)
skew2:mar4	0.45***	0.08***	0.15***	0.06***	0.05***	0.10***	0.06***
	(0.43, 0.48)	(0.05, 0.10)	(0.13, 0.17)	(0.04, 0.08)	(0.04, 0.07)	(0.08, 0.12)	(0.04, 0.08)
skew3:mar4	1.23***	0.31***	0.43***	0.19***	0.15***	0.51***	0.16***
	(1.20, 1.27)	(0.28, 0.34)	(0.40, 0.46)	(0.16, 0.21)	(0.12, 0.17)	(0.48, 0.54)	(0.14, 0.19)
Fallzahl	9000	8999	9000	9000	9000	9000	9000
F-Statistik	17411.928	952.883	1046.671	840.272	779.862	1080.366	743.061
R^2	0.944	0.601	0.597	0.593	0.527	0.653	0.532
korr. R^2	0.944	0.6	0.596	0.593	0.526	0.652	0.532
SEE	0.273	0.202	0.199	0.181	0.159	0.2	0.177
Note:							*p<0.05; **p<0.01; ***p<0.001

[155] Anmerkung: F-Statistik in allen Modellen in A6.2: *** (p < 0.001).

Tabelle A19: Modellbasierte Analyse zum Standardfehlerbias: Kovarianzen

	abh. Variable: durchschnittlicher absoluter SE-Bias in den Kovarianzen						
	b (95 % KI)						
	EM	EMB	FCS	FIML	H0	MNV	PMM
Constant	0.20***	0.12***	0.13***	0.15***	0.13***	0.13***	0.12***
	(0.20, 0.21)	(0.12, 0.13)	(0.13, 0.14)	(0.14, 0.15)	(0.13, 0.14)	(0.13, 0.14)	(0.11, 0.12)
250	0.21***	0.13***	0.12***	0.10***	0.12***	0.11***	0.14***
	(0.20, 0.22)	(0.12, 0.14)	(0.11, 0.13)	(0.09, 0.10)	(0.11, 0.13)	(0.10, 0.12)	(0.13, 0.15)
skew2	0.02**	-0.002	-0.01*	-0.01***	0.002	-0.01**	0.02***
	(0.01, 0.03)	(-0.01, 0.01)	(-0.02, -0.0003)	(-0.02, -0.01)	(-0.01, 0.01)	(-0.02, -0.004)	(0.01, 0.03)
skew3	0.13***	0.02***	0.001	-0.04***	-0.01	0.01	0.01
	(0.12, 0.14)	(0.01, 0.03)	(-0.01, 0.01)	(-0.05, -0.03)	(-0.02, 0.001)	(-0.001, 0.02)	(-0.001, 0.01)
mar3	0.49***	0.004	-0.01*	-0.02***	0.002	0.002	0.03***
	(0.48, 0.50)	(-0.004, 0.01)	(-0.02, -0.002)	(-0.03, -0.01)	(-0.01, 0.01)	(-0.01, 0.01)	(0.02, 0.03)
mar4	1.18***	0.07***	0.16***	-0.02***	0.05***	0.05***	0.08***
	(1.17, 1.19)	(0.06, 0.08)	(0.15, 0.17)	(-0.03, -0.01)	(0.04, 0.06)	(0.04, 0.06)	(0.07, 0.09)
250:skew2	0.16***	0.02***	0.03***	0.04***	0.01*	0.04***	-0.02**
	(0.15, 0.17)	(0.01, 0.04)	(0.02, 0.04)	(0.03, 0.05)	(0.003, 0.03)	(0.03, 0.05)	(-0.03, -0.004)
250:skew3	0.46***	0.16***	0.13***	0.22***	0.13***	0.17***	0.07***
	(0.44, 0.47)	(0.15, 0.18)	(0.12, 0.15)	(0.20, 0.23)	(0.11, 0.14)	(0.15, 0.18)	(0.06, 0.08)
250:mar3	0.63***	0.11***	0.11***	0.10***	0.08***	0.06***	0.04***
	(0.62, 0.65)	(0.10, 0.12)	(0.10, 0.12)	(0.09, 0.11)	(0.07, 0.09)	(0.05, 0.08)	(0.03, 0.05)
250:mar4	1.04***	0.22***	0.13***	0.26***	0.21***	0.17***	0.12***
	(1.03, 1.06)	(0.20, 0.24)	(0.11, 0.14)	(0.25, 0.28)	(0.19, 0.22)	(0.15, 0.18)	(0.11, 0.14)
skew2:mar3	0.27***	0.04***	0.09***	0.04***	0.04***	0.05***	0.02***
	(0.25, 0.28)	(0.03, 0.05)	(0.08, 0.10)	(0.03, 0.05)	(0.03, 0.05)	(0.04, 0.06)	(0.01, 0.03)
skew3:mar3	0.57***	0.14***	0.22***	0.12***	0.11***	0.18***	0.09***
	(0.56, 0.59)	(0.13, 0.16)	(0.21, 0.24)	(0.10, 0.13)	(0.10, 0.12)	(0.17, 0.20)	(0.08, 0.10)
skew2:mar4	0.42***	0.11***	0.15***	0.10***	0.09***	0.12***	0.06***
	(0.41, 0.44)	(0.09, 0.12)	(0.13, 0.16)	(0.08, 0.11)	(0.08, 0.11)	(0.10, 0.13)	(0.04, 0.07)
skew3:mar4	1.16***	0.25***	0.40***	0.27***	0.22***	0.35***	0.18***
	(1.14, 1.18)	(0.23, 0.27)	(0.38, 0.42)	(0.26, 0.29)	(0.20, 0.24)	(0.33, 0.37)	(0.16, 0.20)
Fallzahl	9000	8999	9000	9000	9000	9000	9000
F-Statistik	48464.658	1526.433	2213.527	1175.998	1184.36	1477.728	1109.734
R^2	0.986	0.692	0.756	0.696	0.638	0.702	0.604
korr. R^2	0.986	0.692	0.755	0.696	0.637	0.701	0.604
SEE	0.139	0.154	0.143	0.143	0.146	0.146	0.126
Note:							*p<0.05; **p<0.01; ***p<0.001

Tabelle A20: Test auf Unterschiede in den b-Koeffizienten: Faktorladungen (SE-Bias)

Differenzen der b- Koeffizienten zwischen den MDTs mit 5 %iger Irrtumswahrscheinlichkeit signifikant?

Effekt	EM vs. EMB	EM vs. FCS	EM vs. FIML	EM vs. H0	EM vs. MNV	EM vs. PMM	EMB, vs. FCS	EMB vs. FIML	EMB vs. H0	EMB vs. MNV	EMB vs. PMM	FCS vs. FIML	FCS vs. H0	FCS vs. MNV	FCS vs. PMM	FIML vs. H0	FIML vs. MNV	FIML vs. PMM	H0 vs. MNV	H0 vs. PMM	PMM vs. MNV
250	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
skew2	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
skew3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
mar3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
mar4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE
250:skew2	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
250:skew3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
250:mar3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE
250:mar4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE
skew2:mar3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
skew3:mar3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE
skew2:mar4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
skew3:mar4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE

Anmerkungen: FALSE: Unterschied nicht signifikant; TRUE: signifikanter Unterschied.

Tabelle A21: Test auf Unterschiede in den b-Koeffizienten: Strukturpfade (SE-Bias)

Differenzen der b- Koeffizienten zwischen den MDTs mit 5 %iger Irrtumswahrscheinlichkeit signifikant?

Effekt	EM vs. EMB	EM vs. FCS	EM vs. FIML	EM vs. H0	EM vs. MNV	EM vs. PMM	EMB, vs. FCS	EMB vs. FIML	EMB vs. H0	EMB vs. MNV	EMB vs. PMM	FCS vs. FIML	FCS vs. H0	FCS vs. MNV	FCS vs. PMM	FIML vs. H0	FIML vs. MNV	FIML vs. PMM	H0 vs. MNV	H0 vs. PMM	PMM vs. MNV
250	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
skew2	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
skew3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE
mar3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
mar4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
250:skew2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
250:skew3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
250:mar3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
250:mar4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
skew2:mar3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
skew3:mar3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
skew2:mar4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
skew3:mar4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE

Anmerkungen: FALSE: Unterschied nicht signifikant; TRUE: signifikanter Unterschied.

Tabelle A22: Test auf Unterschiede in den b-Koeffizienten: Kovarianzen (SE-Bias)

Differenzen der b- Koeffizienten zwischen den MDTs mit 5 %iger Irrtumswahrscheinlichkeit signifikant?

Effekt	EM vs. EMB	EM vs. FCS	EM vs. FIML	EM vs. H0	EM vs. MNV	EM vs. PMM	EMB, vs. FCS	EMB vs. FIML	EMB vs. H0	EMB vs. MNV	EMB vs. PMM	FCS vs. FIML	FCS vs. H0	FCS vs. MNV	FCS vs. PMM	FIML vs. H0	FIML vs. MNV	FIML vs. PMM	H0 vs. MNV	H0 vs. PMM	PMM vs. MNV
250	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE
skew2	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
skew3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
mar3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE
mar4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE
250:skew2	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
250:skew3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
250:mar3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
250:mar4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
skew2:mar3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE
skew3:mar3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
skew2:mar4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
skew3:mar4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Anmerkungen: FALSE: Unterschied nicht signifikant; TRUE: signifikanter Unterschied.

A7 Empirische Exemplifikation

Tabelle A23: Deskriptive Statistiken für das empirische Beispielmodell

	Indikatoren der latenten Konstrukte							Prädiktoren					
	AUSL1	AUSL2	AUSL3	ANTI1	ANTI2	ANTI3	STOLZ	ANOMIA	LAGE	DEPRIV	SCHICHT	BILD	EK
Datensatz ohne Missing Values													
Fallzahl	250	250	250	250	250	250	250	250	250	250	250	250	250
Mittelwert	2,19	2,21	1,58	2,40	2,73	1,84	2,99	2,39	2,33	2,31	3,10	2,55	0,42
SD	1,184	1,260	1,062	1,209	1,204	1,072	0,799	1,266	0,770	0,651	0,724	1,196	0,495
Skewness	0,692	0,707	1,897	0,452	0,223	1,001	-0,645	-0,332	1,111	0,461	0,230	0,106	0,309
Kurtosis	-0,212	-0,449	2,831	-0,504	-0,522	0,072	0,224	-0,937	1,944	0,313	0,146	-1,552	-1,920
Datensatz mit Missing Values													
Fallzahl	164	163	155	159	162	151	250	179	250	170	250	160	164
Fehlende Werte (in %)	86 (34,4)	87 (34,8)	95 (38,0)	91 (36,4)	88 (35,2)	99 (39,6)	0	71 (28,4)	0	80 (32,0)	0	90 (36,0)	86 (34,4)
Mittelwert	2,24	2,19	1,59	2,31	2,83	1,86	2,99	2,34	2,33	2,34	3,10	2,48	0,41
SD	1,194	1,220	1,073	1,174	1,170	1,007	0,799	1,307	0,770	0,679	0,724	1,187	0,493
Skewness	0,656	0,683	1,861	0,471	0,141	0,681	-0,645	-0,306	1,111	0,617	0,230	0,182	0,376
Kurtosis	-0,249	-0,397	2,698	-0,456	-0,430	-0,776	0,224	-1,013	1,944	0,383	0,146	-1,500	-1,882

Anmerkungen: SD: Standardabweichung.

O1 Templates und Programmcodes

Folgender Link führt zu allen notwendigen Templates und Programmcodes, um die Ergebnisse zu replizieren: https://osf.io/bywpd/?view_only=0a48ee8223cf49e4949270093c58d999.

In dem zur Verfügung stehenden Ordner finden sich die Templates, die zur Datengenerierung benutzt wurden (Anhang O1.1). Mithilfe dieser wird es möglich, dieselben Daten zu generieren, die auch dieser MC-Studie vorlagen. Im Anhang O1.2 findet sich die Mplus-Syntax für die Implementation von H0. Anhang O1.3 beinhaltet die Konfiguration der MDTs, für welche R benutzt wurde: das ist neben EM und EMB auch FCS, MNV und PMM. Mithilfe der zur Verfügung gestellten Programmcodes lässt sich die Implementation dieser Techniken und deren Konfiguration nachvollziehen. Im Anhang O1.4 findet sich die Analysesyntax für die aufbereiteten Daten. Da sich die Analyse der Daten für Direct-ML, für EM oder für die MI-Varianten voneinander unterscheiden (für Direct-ML beruht die Analyse auf einem Datensatz, während für die MI-Techniken eine Liste als Grundlage dient, mit welcher die einzelnen *m* Datensätze angesteuert werden), muss für die jeweilige Technik die Analysesyntax der Modellschätzung bisweilen angepasst werden. Die Anpassung betrifft aber weder den Modellschätzer noch das Modell selbst, sondern bezieht sich auf die Art der vorliegenden Daten (ob Liste oder einfacher Datensatz) oder auf die Reihenfolge der Variablen in den Datensätzen. Im Anhang O1.5 findet sich ein Beispiel für den Imputationsprozess auf dem bwUniCluster. Mit dem ersten File wird das zweite File generiert, das die Informationen enthält, welche Software das Cluster benötigt und wo das eigentliche Skript liegt, das im Cluster durchlaufen werden soll. Das dritte File stellt eines dieser Prozess-Skripte dar. Das vierte File beinhaltet den Auftrag das dritte File an das Cluster zu senden und den Prozess zu starten. Das Skript läuft dann automatisiert durch und speichert die Ergebnisse an dem gewünschten Ort ab.

O2 Weitere Grafiken und Tabellen

Derselbe Link (https://osf.io/bywpd/?view_only=0a48ee8223cf49e4949270093c58d999) führt zu den Grafiken und Tabellen, auf die der Text verweist.

- O2.1: Enthält die Analysen aus Kapitel 7.1 für Modell 1
    - Tabellen: O2.1.1
    - Grafik: O2.1.2
- O2.2: Enthält die Analysen aus Kapitel 7.2 für Modell 1
    - Grafik der AMEs: O2.2.1
    - Tabellen der log. Regressionen: O2.2.2

- O2.3: Enthält zusätzliche Ergebnistabellen für die Analysen aus Kapitel 8.1
    - Populationsbias Modell 3: O2.3.1
    - Ergebnisse Parameterbias Modell 1: O2.3.2
    - Ergebnisse Parameterbias Modell 2: O2.3.3
    - Grafik für Modell 1 und Modell 2: O2.3.4
- O2.4: Enthält die Analysen für Modell 1 und Modell 2 aus Kapitel 8.2
    - Effektstärken der unabhängigen Variablen: O2.4.1
    - Regressionsmodelle: O2.4.2
    - Test auf Unterschiede in den b-Koeffizienten: O2.4.3
- O2.5: Analysen zur relativen Effizienz für das erste und zweite Modell (Kapitel 8.3)
- O2.6: Enthält die Analysen für Modell 1 und Modell 2 aus Kapitel 9.1
    - Ergebnisse Standardfehlerbias Modell 1: O2.6.1
    - Ergebnisse Standardfehlerbias Modell 2: O2.6.2
    - Grafik für Modell 1 und Modell 2: O2.6.3
- O2.7: Enthält die Analysen für Modell 1 und Modell 2 aus Kapitel 9.2
    - Effektstärken der unabhängigen Variablen: O2.7.1
    - Regressionsmodelle: O2.7.2
    - Test auf Unterschiede in den b-Koeffizienten: O2.7.3
- O2.8: Enthält alle Files der empirischen Exemplifizierung in Kapitel 11