Die Auswirkungen nicht-spezifizierter Mehrdimensionalität bei der linearen Strukturgleichungsmodellierung

Esther Theresa Beierl

Die Auswirkungen nicht-spezifizierter Mehrdimensionalität bei der linearen Strukturgleichungsmodellierung

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Philosophie an der Ludwig-Maximilians-Universität
München

vorgelegt von
Esther Theresa Beierl
aus Augsburg

München, 2016

Referent: Prof. Dr. Moritz Heene
Korreferent: Prof. Dr. Markus Bühner
Tag der mündlichen Prüfung: 01. Juli 2016

Danksagung

Ich danke einer Vielzahl an Menschen, die mich in dieser akademischen Lebensphase unterstützt haben und ohne die ich diese Dissertation sicherlich nicht zu einem so guten Ende gebracht hätte.

An erster Stelle möchte ich Prof. Dr. Moritz Heene danken, dem besten Doktorvater, den man sich wünschen kann. Er hat diese Arbeit immer höchst kompetent begleitet und mich gleichzeitig in seiner bescheidenen Art auch die Demut vor dem Wissen gelehrt, welches noch nicht bekannt ist. Mit seinen Werten und Prinzipien hinsichtlich unserer Forschungstätigkeit war und ist er mir immer ein Vorbild. Er hat mir jeglichen Freiraum gegeben, welchen ich brauchte, um mit der Arbeit voran zu kommen. Dabei hat er mir stets sehr wertvolles Feedback gegeben sowie immer an mich und meine Arbeit geglaubt.

Ich danke Herrn Prof. Dr. Markus Bühner für seine kreative Idee für die zweite Studie, seinen Blick für das „große Ganze“ und seine kritischen Rückfragen zu jedem Schritt, welchen ich auf dem Weg zur Promotion gegangen bin.

Ich danke Herrn Prof. Dr. André Beauducel, durch dessen Vorschlag hinsichtlich der Art der Simulation und Starthilfe bei deren Umsetzung die zweite Studie stark an Qualität zugenommen hat.

Ich danke Herrn Prof. Dr. Thomas Eckert für die freundliche und unkomplizierte Übernahme der Drittbegutachtung.

Ich danke zwei super Teams, dem Team der Methodenlehre (insbesondere Florian Pargent, Felix Naumann, Dr. Sven Hilbert, Dr. Philipp Sckopke, Dr. Stella Bollmann und Cora Laugs) und dem Team der Learning Sciences (vor allem Wendy Symes, Julia Hölzl und Jessica Thomson; thank you so much for the ’Cheer-up-Esther-day’, Ladies!), die mich sowohl fachlich, als auch moralisch auf dem Weg zur Promotion immer unterstützt haben.

Des Weiteren danke ich meinen „frühen Förderern“ an der Universität Salzburg, Dr. Paul Lengenfelder, Ass.-Prof. Dr. Anton Kühberger und Dr. Thomas Scherndl, die mir schon im „Toddler-Stadium“ des Psychologiestudiums die Begeisterung für die Methodenlehre und die Psychometrie vermittelt haben. Ohne deren Zuspruch bereits zu Beginn des Studiums hätte ich wahrscheinlich nicht das nötige Selbstvertrauen

aufgebracht, mich fachlich in Richtung der Methodenlehre und Psychometrie zu entwickeln, geschweige denn eine Doktorarbeit in diesem Bereich zu beginnen.

Ich danke meinen Eltern Marianne und Dr. Otto Beierl, die mich in meinen akademischen Vorhaben immer unterstützt haben und für mich da waren.

Zudem danke ich meinem Freundeskreis (insbesondere Tina, Wendy, Ellen, Nils und Vera) für den moralischen Beistand, wenn mir mal wieder die Nerven durchgegangen sind, aber natürlich auch für diverse gemeinsame Erlebnisse, wenn ich gerade nicht an der Arbeit saß.

Ferner danke ich Kristina Maier für die großartige moralische Unterstützung.

Ich danke der *Studienstiftung des deutschen Volkes* für die Unterstützung meiner Promotion und den anderen Stipendiatinnen und Stipendiaten für regen Austausch, kritische Fragen und vor allem eine tolle gemeinsame Zeit.

Abschließend danke ich Marianna Rusche für das engagierte Korrekturlesen der Arbeit.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

Zusammenfassung

Analysen zur Entdeckung von Modellmissspezifikationen bei Strukturgleichungsmodellen anhand der Fit-Indizes nehmen in der Fachliteratur der letzten Jahrzehnte einen großen Raum ein. Dennoch wurden Missspezifikationen im Strukturmodell in Form einer nicht-spezifizierten Zweidimensionalität beispielsweise noch nicht untersucht. Daher wurde im Rahmen einer ersten Studie untersucht, inwieweit die am meisten verwendeten Fit-Indizes diese Art und unterschiedliche Grade dieser Missspezifikation (operationalisiert durch die Höhe der Faktorkorrelation und (un-) ausgewogene Indikatorenaufteilung im Populationsmodell) zuverlässig erkennen würden. Es wurden ferner realistisch hohe und heterogene Faktorladungen für die Populationsmodelle verwendet, aus denen die Stichproben erzeugt wurden. Der CFI führte bei schwerwiegender und mittelschwerer Missspezifikation zur Modellablehnung anhand des Cut-Offs nach Hu und Bentler (1998, 1999), RMSEA und SRMR erwiesen sich als ungeeignet, diese Form der Missspezifikation anzuzeigen. Insbesondere wurde jedoch die Frage nach den Konsequenzen missspezifizierter Modelle auf individualdiagnostische Entscheidungen basierend auf den Faktorwerten bisher nicht gestellt. Im Rahmen der zweiten Studie wurde daher mit einer populationsbasierten Simulation auf Basis vorab definierter Faktorwerte der Frage nachgegangen, inwieweit die Diagnosegenauigkeit leiden würde, sofern dichotome Diagnosen auf Basis der Bartlett-Faktorwerte missspezifizierter Modelle anstatt auf Basis der Bartlett-Faktorwerte korrekter Modelle vergeben wurden. Des Weiteren wurde die Güte der Diagnostik auf Basis der üblicherweise verwendeten Gesamtsummenwerte untersucht. Es wurde dasselbe Design wie für die erste Studie verwendet, wobei zusätzlich die Basisraten für die Diagnosegebung variiert wurden. Vor allem die unterschiedlichen Basisraten und die Höhe der Faktorladungen hatten bereits einen entscheidenden Einfluss auf die Güte der Diagnostik auf Basis korrekter Modelle; ebenso Basisraten, Faktorladungen und der Grad der Missspezifikation/die unterschiedlichen Populationsmodelle auf die Güte der Diagnostik auf Basis missspezifizierter Modelle und der Gesamtsummenwerte. Die Konsequenzen der Befunde für Wissenschaft und Praxis werden diskutiert.

Schlüsselbegriffe: Strukturgleichungsmodellierung, Missspezifikation, Mehrdimensionalität, Fit-Indizes, Faktorwerte, Psychometrie, Diagnose

I EINLEITUNG

> „Das Bestehen der experimentellen Methode lä[ss]t uns glauben, wir hätten das Mittel, die Probleme, die uns beunruhigen, loszuwerden; obgleich Problem und Methode windschief aneinander vorbeilaufen."
>
> (Wittgenstein, 1953, S. 232)

In diesem Zitat, das Wittgenstein bereits 1953 an die Psychologie richtete, findet sich ein Grund wieder, der sicherlich zu einem großen Teil zur aktuellen Replikationskrise der Psychologie beitrug: der Irrglaube, durch die Anwendung einer Methode sei Wissenschaftlichkeit bereits erreicht. Oder noch drastischer ausgedrückt: Der Zweck heilige die Mittel.

In Gang gesetzt wurde die Replikationskrise mit einem kritischen Kommentar Wagenmakers', Wetzels', Borsbooms, und van der Maas' (2011) auf einen Artikel Bems (2011) im *Journal of Personality and Social Psychology*. Bems Befunde konnten nicht repliziert werden (Schimmack, 2012) – wie viele andere psychologische Befunde auch: In einem groß angelegten „Reproducibility Project: Psychology" (Open Science Collaboration, 2015, S. 1) wurden beispielsweise nur 36% von 100 als signifikant publizierter psychologischer Befunde signifikant, wobei 83% der Effektstärken unter den im Original berichteten Effektstärken lagen. Vor allem Ursachen wie der Publikationsbias und fragwürdige Forschungspraktiken (für eine Auflistung vgl. Schimmack [2012]) führten zu diesen und weiteren falsch positiven Befunden psychologischer Studien (Schimmack, 2012, 2016).

Auch diese Dissertation beschäftigte sich im weitesten Sinne mit der von Wittgenstein (1953) beschriebenen Diskrepanz zwischen dem Status-Quo in der Forschung und dem wissenschaftlichen Gold-Standard. Als Beispiel wurde die lineare Strukturgleichungsmodellierung herangezogen. Es wurden die Auswirkungen einer realistisch niedrigen Reliabilität (Faktorladungen) sowie mangelnder faktorieller Validität/einer Missspezifikation im Strukturmodell untersucht, und zwar einerseits hinsichtlich der Sensitivität der Fit-Indizes für diese Modellmissspezifikation,

andererseits hinsichtlich der Validität von Diagnosen, die vor dem Hintergrund von Test- und Fragebogenvalidierungen mittels Strukturgleichungsmodellen aus den Faktorwerten dieser Modelle vergeben wurden.

Mit der linearen Strukturgleichungsmodellierung (Jöreskog, 1969, 1973) wurde ein Datenanalyseverfahren für diese Arbeit ausgewählt, welches innerhalb der letzten vier Jahrzehnte in Grundlagen-, wie auch in angewandten Forschungsbereichen der Psychologie und verwandten Disziplinen zunehmend an Bedeutung gewann (Hershberger, 2003; MacCallum & Austin, 2000; Reinecke, 2014; Tremblay & Gardner, 1996). Wird beispielsweise in Google Scholar der einigermaßen spezifische Suchbegriff „Structural Equation Modeling“ eingegeben, erscheinen 2,47 Mio. Einträge. Im Vergleich dazu erreicht der sehr allgemeine Suchbegriff „Psychological Research“, der letztendlich alle Forschungsbereiche innerhalb der Psychologie abdeckt, mit 4,34 Mio. Einträgen nur fast doppelt so viele Verweise. Für die renommierte Fachzeitschrift *Structural Equation Modeling* lässt zudem sich ein kontinuierlicher Anstieg der Zitationen aus dieser Zeitschrift verzeichnen (SCImago, 2007). Außerdem rangiert diese Fachzeitschrift an erster Stelle der 45 Zeitschriften aus dem Bereich der Mathematischen Methoden (“Structural Equation Modeling: A Multidisciplinary Journal,” 2015).

Die kritische Frage in Bezug auf die Modelltestung im Rahmen der Strukturgleichungsmodellierung bezieht sich auf die Modellpassung, also darauf, wie zuverlässig Missspezifikationen als solche erkannt werden (Fan & Sivo, 2005; Fan, Thompson, & Wang, 2009). Dieser Frage wurde im Rahmen einer ersten Simulationsstudie nachgegangen. Anhand von zahlreichen Studien (siehe Kapitel III. 1 und 2) wurde bereits gezeigt, dass die Höhe der Faktorladungen (Reliabilität) einen entscheidenden Einfluss darauf hat, inwieweit die Fit-Indizes eine Modellmissspezifikation als solche erkennen. Daher wurden realistisch hohe und zugleich realistisch heterogene Faktorladungen (Buzick, 2010; Peterson, 2000) für die Generierung der Populationsmodelle verwendet, aus denen die Stichproben gezogen wurden. Die untersuchte Art der Missspezifikation stellte einen typischen Fall einer Missspezifikation in der Psychologie und verwandter Disziplinen dar (Little, Cunningham, Shahar, & Widaman, 2002b): Es wurde fälschlicherweise ein einfaktorielles Modell spezifiziert, wohingegen im Populationsmodell zwei korrelierte latente Variablen definiert wurden. Diese Art der Missspezifikation stellte eine Verletzung der faktoriellen Validität eines Modells/Testverfahrens dar und wurde bisher

noch kaum untersucht (siehe III. 2.3). Der Grad dieser Missspezifikation wurde einerseits variiert durch unterschiedlich hohe, für die Psychologie typische Korrelationen (Rost, 2009; Steel, Schmidt, & Shultz, 2008) zwischen den latenten Variablen, andererseits durch eine (un-)ausgewogene Indikatorenaufteilung auf die beiden Faktoren im Populationsmodell, wobei insgesamt eine für psychologische Testverfahren/Fragebögen typische Indikatorenanzahl verwendet wurde (Peterson, 2000; Shrout & Yager, 1989). Konkret wurde im Rahmen der ersten Simulationsstudie untersucht, ob die am häufigsten verwendeten (Beauducel & Wittmann, 2005; Marsh, Hau, & Grayson, 2013; McDonald & Ho, 2002; Savalei, 2012) Fit-Indizes CFI (Bentler, 1990), RMSEA (Steiger & Lind, 1980) und SRMR (Bentler, 1995) eine Missspezifikation der beschriebenen Form anhand der Cut-Off-Kriterien nach Hu und Bentler (1998, 1999) anzeigen würden. Diese Fragestellung betraf die Güte des Modells.

Im Rahmen der zweiten Studie dieser Dissertation wurde noch ein Schritt weiter in Richtung der angewandten Forschung und der psychologischen Praxis gegangen und die Forschungsfrage gestellt, welche psychometrischen Konsequenzen Missspezifikationen im Gegensatz zu korrekten Modellen auf die psychologische Individualdiagnostik aus den Faktorwerten nach sich ziehen würden. Letztere Forschungsfrage ist insbesondere aufgrund der Tatsache von Interesse, dass mehr und mehr Testverfahren und Fragebögen anhand konfirmatorischer Faktorenanalysen und anhand von Strukturgleichungsmodellen konstruktvalidiert werden ("Datenbanksegment PSYNDEX Tests," 2013). Es wurde dasselbe Forschungsdesign wie für die erste Studie verwendet, das ein wenig reliables (realistisch niedrige Faktorladungen) und nicht konstruktvalides (im Strukturmodell missspezifiziertes) Testverfahren darstellen sollte. Die Verwendung unterschiedlich hoher Basisraten in klinischen Größenordnungen (Wittchen et al., 2011) sowie zum Vergleich auch größerer Basisraten (vgl. die Eignungsdiagnostik; Schuler, 2014) für die Vergabe der Diagnosen im Rahmen der zweiten Studie komplettierte das realitätsnahe Forschungsdesign. Zudem wurde die diagnostische Präzision auf Basis der üblicherweise verwendeten Gesamtsummenwerte (Estabrook & Neale, 2013) untersucht.

Zunächst soll im folgenden Kapitel ein Überblick über die lineare Strukturgleichungsmodellierung selbst gegeben werden, bevor anschließend die

Forschungsfragen konkretisiert, die Studien beschrieben und deren Relevanz für Wissenschaft und Praxis diskutiert wird.

II THEORETISCHER HINTERGRUND

1 Grundlegende Konzeption von Faktorenanalysen und linearen Strukturgleichungsmodellen

Ausgangspunkt für Datenanalysemethoden wie exploratorische und konfirmatorische Faktorenanalysen sowie lineare Strukturgleichungsmodelle ist die Definition des Messwerts eines Indikators. Die Fundamentalgleichung der Faktorenanalyse stellt dar, dass sich die beobachteten Variablen aus gewichteten gemeinsamen und uniquen Faktorwerten zusammensetzen (Mulaik, 2009, S. 136):

$$\mathbf{Y} = \mathbf{\Lambda}\,\mathbf{X} + \mathbf{\Psi}\,\mathbf{E} \qquad (1)$$

Y stellt eine n × 1 Matrix dar, wobei n die Anzahl der beobachteten Variablen ist. **X** stellt eine r × 1 Matrix dar, wobei r die Anzahl der gemeinsamen latenten Faktoren darstellt. **Λ** ist die n × r Mustermatrix, **Ψ** ist die n × n Mustermatrix der Koeffizienten[1] der uniquen Faktorwerte[2] und **E** die n × 1 Matrix der uniquen Faktorwerte (Mulaik, 2009). Voraussetzungen sind, dass die Erwartungswerte der beobachteten Variablen **Y**, der Faktorwerte **X** sowie der Fehler **E** Null sind.

Im faktorenanalytischen Modell wird davon ausgegangen, dass die uniquen Faktorwerte sowohl Messfehleranteile enthalten, als auch einen Anteil, der spezifisch für die entsprechende beobachtete Variable ist und mit keiner anderen beobachteten Variablen geteilt wird (Eid, Gollwitzer, & Schmitt, 2013). Im Gegensatz dazu repräsentiert der Varianzanteil einer beobachteten Variablen, welcher nicht durch die latenten Variablen erklärt werden kann, im Rahmen eines True-Score-Modells (Eid et al., 2013, S. 856) nur Messfehler. Bei den unter IV und V beschriebenen Studien handelt es sich ausschließlich um True-Score-Modelle.

Unter der Voraussetzung, dass die gemeinsamen und die uniquen Faktorwerte unkorreliert sind, d.h. die Matrix der uniquen Faktorwerte **E** multipliziert mit ihrer Transponierten eine Identitätsmatrix darstellt, gibt das Fundamentaltheorem der

[1] Residual-Standardabweichungen (Janssen & Laatz, 2013, S. 555)

[2] Residuen (Janssen & Laatz, 2013, S. 548)

Faktorenanalyse an, wie die Korrelationsmatrix der beobachteten Variablen errechnet wird (Mulaik, 2009, S. 136):

$$\mathbf{R_{YY}} = \mathbf{\Lambda}\, \mathbf{\Phi_{XX}}\, \mathbf{\Lambda'} + \mathbf{\Psi^2} \qquad (2)^3$$

Dabei gibt $\mathbf{\Phi_{XX}}$ die r × r Korrelationsmatrix der Faktoren an. Die Differenz aus $\mathbf{R_{YY}}$ und $\mathbf{\Psi^2}$ ergibt die sog. reduzierte Korrelationsmatrix (Mulaik, 2009, S. 136), wobei $\mathbf{\Psi^2}$ positiv definit sein muss, d.h., die Eigenwerte dieser symmetrischen Matrix sind größer als Null (Fahrmeier, Hamerle, & Tutz, 1996). In der Diagonalen letzterer Matrix stehen die Kommunalitäten, die Varianzen der Items, welche nur durch die gemeinsamen Faktoren erklärt werden (Bühner, 2011). Die Kommunalität eines Items entspricht dem Determinationskoeffizienten in der multiplen Regressionsanalyse (Eid et al., 2013). Sie gibt die quadrierte multiple Korrelation zwischen der beobachteten Variablen und der latenten Variablen an.

Sofern in den Off-Diagonal-Elementen von $\mathbf{\Phi_{XX}}$ Nullen stehen, handelt es sich um ein orthogonales Faktorenmodell, andernfalls um ein obliques (Mulaik, 2009). Im Rahmen eines orthogonalen Faktorenmodells entspricht die Mustermatrix, die die Faktorladungen enthält, der Strukturmatrix, letztere enthält die Korrelationen der beobachteten Variablen mit den Faktoren. Andernfalls stellen die Einträge der Mustermatrix semipartielle Regressionsgewichte dar und werden aus der Multiplikation der Inversen von $\mathbf{\Phi_{XX}}$ mit der Strukturmatrix berechnet (Mulaik, 2009, S. 138):

$$\mathbf{\Lambda} = \mathbf{R_{YX}}\, \mathbf{\Phi_{XX}^{-1}} \qquad (3)^4$$

Die reproduzierte Kovarianzmatrix $\mathbf{\Sigma}$ des implizierten Modells kann im Rahmen der konfirmatorischen Faktorenanalyse konzeptuell genauso wie in Formel (2) dargestellt werden (Mulaik, 2009, S. 440), $\mathbf{\Phi_{XX}}$ stellt die r × r Kovarianzmatrix der latenten Faktoren und $\mathbf{\Theta}$ die n × n Varianz-Kovarianzmatrix der uniquen Faktorwerte dar:

$$\mathbf{\Sigma} = \mathbf{\Lambda}\, \mathbf{\Phi_{XX}}\, \mathbf{\Lambda'} + \mathbf{\Theta} \qquad (4)$$

[3] $\mathbf{\Lambda'}$ stellt die Transponierte der Mustermatrix dar; der Apostroph kennzeichnet hier und im Folgenden die Transponierte eines Vektors oder einer Matrix.

[4] Der Exponent -1 stellt hier und im Folgenden die Inverse einer Matrix dar.

Bei der konfirmatorischen Faktorenanalyse müssen die uniquen Faktorwerte – im Vergleich zur exploratischen Faktorenanalyse – nicht unkorreliert sein (Mulaik, 2009). Sollten sie unkorreliert sein, entspricht das $\mathbf{\Psi^2}$ aus Formel (2) im nicht-standardisierten Fall der Varianz-Kovarianz-Matrix $\mathbf{\Theta}$.

Das Modell in Gleichung (4) besteht aus zwei oder mehreren Messmodellen bzw. konfirmatorischen Faktorenanalysen sowie einem Strukturmodell in Form von korrelierten latenten Variablen bzw. Faktoren. Gleichung (4) stellt insofern den einfachsten Fall eines linearen Strukturgleichungsmodells dar. Diese Arbeit fokussiert sich auf diesen Fall, weshalb an dieser Stelle auf die technische Einführung von Strukturmodellen in Form von gerichteten Pfaden verzichtet wird. Für eine Darstellung dieser Form von Strukturgleichungsmodellen werden die Lehrbücher von Byrne (1998), Byrne (2009), Kline (2001), Reinecke (2014) sowie Weiber und Mühlhaus (2010) empfohlen.

2 Beurteilung der Modellpassung

2.1 χ^2- Anpassungstest

Die Schwierigkeit, die mit Datenanalyseverfahren wie konfirmatorischen Faktorenanalysen und linearen Strukturgleichungsmodellen einhergeht, liegt in der Beurteilung der Modellpassung. Der Fokus dieser Arbeit liegt auf einer Form der Missspezifikation: nicht-spezifizierter Zweidimensionalität im Strukturmodell. Daher werden im Folgenden die Kriterien zur Beurteilung der Modellpassung näher erläutert, die im Rahmen der ersten Studie zur Modellevaluation verwendet wurden. Als erstes wird der χ^2-Anpassungstest beschrieben, bevor die Fit-Indizes, deren Sensitivität in Studie 1 untersucht wurde, näher erläutert werden.

Mittels eines χ^2-Anpassungstests wird approximativ der globale Modellfit überprüft, d.h., ob die aus den empirischen Daten geschätzte Populations-Kovarianzmatrix mit der aus dem theoretischen Modell implizierten Kovarianzmatrix übereinstimmt bzw. die Elemente in den Off-Diagonalen der Uniqueness-Matrix Null sind oder nahe an Null liegen (Schermelleh-Engel, Moosbrugger, & Müller, 2003). Dabei gilt die Entscheidung über die Modellpassung als dichotom, das theoretische Modell wird an der aus den Daten geschätzten Populations-Kovarianzmatrix entweder abgelehnt (bei einem Signifikanzniveau von 5%: $p \leq .05$) oder angenommen ($p > .05$). Wenn die Nullhypothese korrekt ist (theoretisches Modell passt zu den Daten) erreicht die Diskrepanzfunktion F ein Minimum und der χ^2-Wert berechnet sich folgendermaßen (S. 32):

$$\chi^2 \text{ (df)} = (N-1)\, F\, [\mathbf{S}, \sum(\widehat{\boldsymbol{\theta}})] \qquad (5)$$

wobei df: Anzahl der Freiheitsgrade

N: Stichprobengröße

F: Diskrepanzfunktion

S: empirische Kovarianzmatrix

$\sum(\widehat{\boldsymbol{\theta}})$: implizierte Kovarianzmatrix

Der χ^2-Test stellt einen statistischen Test zur Prüfung der Modellpassung dar, die Likelihood-Ratio-Statistik, dessen Verteilung F(N-1) eine χ^2-Verteilung darstellt (Schermelleh-Engel et al., 2003). Diese lässt sich aus der multivariaten Normalverteilung herleiten (Mulaik, 2009). Voraussetzungen für die Analyse von Strukturgleichungsmodellen stellen daher multivariate Normalverteilung und hinsichtlich der Modellkomplexität hinreichend große Stichprobengrößen dar (Bentler & Yuan, 1999; Bollen, 1989; Boomsma & Hoogland, 2001; Boomsma, 1983; Mulaik, 2009; Reinecke, 2014; Yang-Wallentin & Jöreskog, 2001).

Einerseits fordern die Schätzalgorithmen bei der Modelltestung eine große Stichprobengröße, andererseits steigt jedoch der χ^2-Wert bei sehr großen Stichproben stark an, sofern eine Modellabweichung vorliegt (Schermelleh-Engel et al., 2003): Bei sehr großen Stichproben reichen nur sehr geringe Abweichungen zwischen der modellimplizierten und der beobachteten Kovarianzmatrix aus, um eine Ablehnung des theoretischen Modells zu bewirken (Jöreskog, 1969; Reinecke, 2014; Steiger, 2007; Thompson & Daniel, 1996). Insofern liegt die Problematik bei der Beurteilung der Modellpassung anhand des χ^2-Tests darin, dass der Schweregrad der Modellabweichung anhand des χ^2-Tests nicht bestimmt werden kann (Saris, Satorra, & Sorbom, 1987). Eine Orientierung stellt lediglich der χ^2-Wert in Relation zu den Freiheitsgraden des spezifizierten Modells dar. Bei einem korrekten Modell entspricht der Erwartungswert des χ^2-Werts der Anzahl der Freiheitsgrade des spezifizierten Modells (Schermelleh-Engel et al., 2003).

2.2 Fit-Indizes

Wie bereits im vorherigen Abschnitt erläutert, werden bereits geringfügig von den Daten abweichende Modelle von der χ^2-Statistik abgelehnt (Fan & Sivo, 2005; Fan et al., 2009). Gleichzeitig kann aber aus der Ablehnung eines implizierten Modells nicht geschlussfolgert werden, dass die Abweichung von der beobachteten Kovarianzmatrix unbedingt klein sei (Saris, Satorra, & Sorbom, 1987). Bei Ablehnung des Modells durch den χ^2-Test können genauso große Missspezifikationen die Ursache sein. Aus diesem Grund wurden alternative Maße für die Überprüfung der Modellpassung entwickelt, die

Fit-Indizes (Schermelleh-Engel et al., 2003). Diese geben – im Gegensatz zum χ^2-Test, anhand dessen die Beurteilung der Modellpassung nach einer dichotomen Entscheidung erfolgt – die Güte der Modellpassung an. Ebenso handelt es sich bei den Fit-Indizes – im Gegensatz zur Inferenzstatistik des χ^2-Tests – um deskriptive Maße zur Evaluation der Modelpassung (Hu & Bentler, 1998; Hu & Bentler, 1999; Steiger, 1990, 2007). In diesem Kontext wurde von der Nullhypothese der exakten Passung zwischen theoretischen Modellmatrix und der empirisch geschätzten Populationsmatrix Abstand genommen und stattdessen der Begriff der annähernden Passung („close fit“) für die Fit-Indizes eingeführt (Browne & Cudeck, 1993, S. 146; Schermelleh-Engel et al., 2003, S. 36).

Die Fit-Indizes sollen folgende – im Gesamten schwer erfüllbare – Kriterien erfüllen: die Spannweite soll zwischen 0 und 1 liegen, die Fit-Indizes sollen unabhängig von der Stichprobengröße sein, und ihre Verteilung sollte bekannt sein (Gerbing & Anderson, 1993). Weitere wünschenswerte Eigenschaften für Fit-Indizes führen Fan et al. (2009) an. So sollten die Fit-Indizes idealerweise auch invariant hinsichtlich der verwendeten Schätzmethode sein; zudem sollten die Fit-Indizes erwartungstreu sein und wenig zufällige Abweichungen anzeigen. Missspezifikationen hingegen sollten sich stark auf die Varianz eines Fit-Index auswirken.

Die am häufigsten verwendeten Fit-Indizes stellen der *Comparative fit index* (CFI; Bentler, 1990), der *Root-mean-square error of approximation* (RMSEA; Steiger & Lind, 1980) und das *Standardized root-mean-square residual* (SRMR; Jöreskog & Sörbom, 1981) dar (Beauducel & Wittmann, 2005; Marsh, Hau, & Grayson, 2013; McDonald & Ho, 2002; Savalei, 2012). Alle drei Fit-Indizes hängen vom χ^2-Wert ab (Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011). Aufgrund ihrer häufigen Verwendung wird in Studie 1 die Sensitivität dieser drei Fit-Indizes ausgetestet. Daher werden CFI, RMSEA und SRMR im Folgenden näher erläutert.

Der RMSEA ist ein absoluter Fit-Index (Steiger & Lind, 1980). Absolute Fit-Indizes indizieren, wie gut ein theoretisches Modell die aus den Daten geschätzten Werte reproduziert (Beauducel & Wittmann, 2005). Der RMSEA gibt die Quadratwurzel der durchschnittlichen Modellabweichung der geschätzten Populations-Kovarianzmatrix von der anhand des Modells spezifizierten Kovarianzmatrix, relativiert an der möglichen Modellabweichung pro Freiheitsgrad, an (Browne & Cudeck, 1993; Heene, Hilbert, Freudenthaler, & Bühner, 2012; Steiger & Lind, 1980; Steiger, 1990). Ein Wert von 0

stellt beim RMSEA eine perfekte Übereinstimmung und ein Wert von 1 eine maximale Abweichung zwischen theoretischer Kovarianzmatrix und der aus den empirischen Daten geschätzten Kovarianzmatrix dar. Die Formel für den RMSEA lautet (Heene et al., 2012, S. 38):

$$\text{RMSEA} = \sqrt{\max\left[\frac{\chi_1^2 - \text{df}_1}{\text{df}_1\,(\text{N}-1)}, 0\right]} \qquad (6)$$

wobei χ^2_1 : χ^2-Statistik des theoretischen Modells, welche eine gewichtete Funktion des Minimums der Maximum-Likelihood-Diskrepanzfunktion für die implizierte Kovarianzmatrix darstellt

df_1 : Freiheitsgrade des theoretischen Modells

N: Stichprobengröße

Somit stellt der RMSEA die Quadratwurzel des normalisierten durchschnittlichen Nonzentralitätsparameters pro Freiheitsgrad dar (Heene et al., 2012). Der Nonzentralitätsparameter gilt als unverzerrter Schätzer des Populationsparameters für die quadrierte Abweichung zwischen theoretischem Modell und empirischen Daten (Schermelleh-Engel et al., 2003) und wird unter III. 1 näher beschrieben.

Das SRMR, entwickelt von Jöreskog und Sörbom (1981), stellt ebenfalls einen absoluten Fit-Index dar. Es entspricht dem durchschnittlichen Residuum der Residualkorrelationen und stellt demnach, wie der RMSEA, einen absoluten Fit-Index dar, welcher zwischen Null (perfekte Modellpassung) und Eins (sehr schlechte Modellpassung) liegen kann (Chen, 2007, S. 466). Es wird die durchschnittliche Abweichung zwischen den Korrelationen der empirischen und der implizierten Korrelationsmatrix berechnet (Heene et al., 2012, S. 39):

$$\text{SRMR} = \sqrt{\frac{2\sum_{i=1}^{p}\sum_{j=1}^{i}\left(\frac{s_{ij} - \widehat{\sigma_{ij}}}{s_{ii}s_{jj}}\right)^2}{p(p+1)}} \qquad (7)$$

wobei p: Anzahl an manifesten Variablen

s_{ij}: empirische Kovarianzen

$\widehat{\sigma}_{ij}$: implizierte Kovarianzen

s_{ii}, s_{jj}: empirische Standardabweichungen

Die gefitteten Residuen, welche aus der Differenz aus empirischen Kovarianzen und modellimplizierten Kovarianzen resultieren, werden anhand von Division durch die empirischen Standardabweichungen standardisiert. Dies kompensiert die Abhängigkeit von der Skalierung der Indikatoren (Jöreskog & Sörbom, 1981).

Der CFI stellt einen komparativen Fit-Index dar und ist der am häufigsten berichtete Fit-Index aus dieser Kategorie (Mahler, 2011; Marsh, Hau, & Grayson, 2013). Komparative Fit-Indizes vergleichen den für das implizierte Modell berechneten χ^2-Wert mit dem χ^2-Wert eines restringierteren Basismodells unter Berücksichtigung der Freiheitsgrade (Bentler, 1992; Heene et al., 2012; Reinecke, 2014). Das Basismodell ist meistens ein Nullmodell, bei dem davon ausgegangen wird, dass die beobachteten Variablen unkorreliert sind (Heene et al., 2012). Die Formel für den CFI lautet (Heene et al., S. 39):

$$\mathrm{CFI} = 1 - \left[\frac{(\chi_1^2 - \mathrm{df}_1)}{(\chi_0^2 - \mathrm{df}_0)}\right] \qquad (8)$$

wobei: χ^2_0: χ^2-Statistik des Basismodells

df_0: Freiheitsgrade des Basismodells

χ^2_1: χ^2-Statistik des theoretischen Modells, welche eine gewichtete Funktion des Minimums der Maximum-Likelihood-Diskrepanzfunktion für die implizierte Kovarianzmatrix darstellt

df_1: Freiheitsgrade des theoretischen Modells

Der CFI (Bentler, 1990) korrigiert die Unterschätzung der Modellpassung des NFI (*Normed fit index*; Bentler & Bonett, 1980; dieser Fit-Index ist ebenfalls ein komparativer Index; Bentler, 1990) bei kleinen Stichproben (Schermelleh-Engel et al., 2003). Beim CFI indiziert ein Wert nahe 1 einen sehr guten Fit, wohingegen ein Wert nahe 0 einen sehr schlechten Fit anzeigt. Allerdings wird argumentiert, die komparativen Fit-Indizes, wie der CFI, seien nicht über verschiedene Studien hinweg vergleichbar, da diese nicht nur abhängig von der Modellspezifikation seien, sondern auch von davon, wie schlecht das Nullmodell ist (siehe z.B. Marsh, Balla, und McDonald [1998]).

Hu und Bentler schlugen basierend auf zwei einflussreichen Simulationsstudien (1998, 1999) ab einer Stichprobengröße von 250 Personen folgende Cut-Off-Werte vor: CFI > .95, RMSEA < .06 und SRMR < .08 (Hu & Bentler, 1999, S. 27). Die Autoren empfahlen außerdem eine „two-index-strategy“ (S. 27) für die Evaluation der Modellpassung, da inkrementelle Fit-Indizes, wie der CFI, sowie der RMSEA sensitiv für Missspezifikationen im Messmodell wären und das SRMR sensitiv für Missspezifikationen im Strukturmodell (siehe III. 1). Für die Kombination des CFI oder des RMSEA zusammen mit dem SRMR empfehlen die Autoren (S. 27) einen Cut-Off von > .95 für den CFI oder den Cut-Off von < .06 für den RMSEA zusammen mit einem Cut-Off von < .09 für das SRMR.

Es wird kritisiert, dass die Cut-Off-Werte für die Fit-Indizes von Hu und Bentler (1998, 1999) relativ willkürlich gesetzt wurden (Mahler, 2011). Des Weiteren führten sie ebenfalls zu einer Modellevaluation auf Basis einer dichotomen Entscheidung, ähnlich wie der χ^2-Test, was ursprünglich nicht intendiert war (Marsh, Hau, & Wen, 2004). Die Fit-Indizes sollen hingegen über den Grad der Missspezifikation Auskunft geben (Fan et al., 2009), da dies der χ^2-Test nicht leisten kann. Aus diesen Gründen ist die gleichzeitige Modellevaluation durch lokale Maße der Modellgüte unabdinglich. Lokale Maße der Modellevaluation stellen im Rahmen konfirmatorischer Ansätze unter anderem die Faktorreliablität, die Höhe und Signifikanz der Faktorladungen sowie der Prozentsatz der aufgeklärten Varianz dar (Hooper, Coughlan, & Mullen, 2008).

Nach wie vor wird die Modellpassung von den angewandt Forschenden jedoch primär an den Cut-Off-Werten nach Hu und Bentler (1998, 1999) festgemacht. Diese Daumenregeln zur Interpretation des Modellfit anhand der Fit werden unter III. 1 anhand weiterer Simulationsstudien diskutiert.

3 Bestimmung der Faktorwerte

In den vorherigen beiden Unterkapiteln wurden zum einen die linearen Modelle dargestellt, auf denen Faktorenanalysen und lineare Strukturgleichungsmodelle aufgebaut sind. Zum anderen wurden die Kennwerte zur Beurteilung der Modellpassung eingeführt, die in der ersten Studie hinsichtlich ihrer Sensitivität bezüglich der nicht-spezifizierten Zweidimensionalität in der Faktorenstruktur ausgetestet wurden. Diese konzeptionelle Vorstellung der Strukturgleichungsmodellierung reicht für die Studie 1, die sich auf die Sensitivität der Fit-Indizes fokussiert, ebenso aus wie für die meisten Fälle der angewandten Forschung, bei denen es um die Testung von Theorien und Hypothesen für Populationen geht (Hershberger, 2003; MacCallum & Austin, 2000; Tremblay & Gardner, 1996), aus. Da jedoch mehr und mehr Testverfahren und Fragebögen für die Einzelfalldiagnostik anhand von konfirmatorischen Faktorenanalysen und Strukturgleichungsmodellen konstruktvalidiert werden ("Datenbanksegment PSYNDEX Tests," 2013), liegt die Bedeutung der Bestimmung der Faktorwerte auf der Hand. Die zweite Studie fokussierte sich auf die Einzelfalldiagnostik aus den Faktorwerten. Im Speziellen wurden im Rahmen der zweiten Studie vor dem Hintergrund der Mehrdimensionalitätsthematik der Dissertation die psychometrischen Konsequenzen für die Diagnostik aus den Faktorwerten untersucht, die für die Individuen entstehen würden, wenn ein Modell angewendet wurde, dass fälschlicherweise nur eine latente Dimension anstatt zweier latenter Dimensionen abbildete. Daher wird im Folgenden auf die Faktorwerteschätzung näher eingegangen.

Für die Bestimmung der Faktorwerte sind die drei bekanntesten und am weitesten verbreiteten Möglichkeiten die Faktorwerte nach Thurstone (1935), die Faktorwerte nach Bartlett (1937) sowie die Anderson-Rubin Faktorwerte (Anderson & Rubin, 1956), letztere machen jedoch nur für orthogonale Rotationen Sinn. Da es sich bei beiden Studien im Rahmen dieser Dissertation um oblique Designs handelt, werden im Folgenden nur die Thurstone- und die Bartlett-Faktorwerte erläutert.

Die Faktorwerte nach Thurstone (Grice, 2001b, S. 433; Mulaik, 2009, S. 375) werden mittels Regressionsmethode geschätzt:

$$\widehat{\mathbf{X}}_\mathbf{i} = \boldsymbol{\Phi}_{XX}\,\boldsymbol{\Lambda}\,\mathbf{R}_{YY}^{-1}\,\widehat{\mathbf{Y}}_\mathbf{i} = \mathbf{R}_{XY}\,\mathbf{R}_{YY}^{-1}\,\widehat{\mathbf{Y}}_\mathbf{i} \qquad (9)$$

Dabei ergibt $\widehat{\mathbf{X}}_\mathbf{i}$ die Matrix der geschätzten Thurstone-Faktorwerte der Person i auf den r Faktoren. Diese Faktorwerte werden aus der Inversen der Korrelationsmatrix der beobachteten Variablen $\mathbf{R}_{\mathbf{YY}}^{-\mathbf{1}}$, der Strukturmatrix $\mathbf{R}_{\mathbf{XY}}$ und dem Vektor $\widehat{\mathbf{Y}}_\mathbf{i}$, der standardisierten Werte auf den beobachteten Variablen einer Person i berechnet (Grice, 2001b; Mulaik, 2009).

Die Bartlett-Faktorwerte (Bartlett, 1937; Grice, 2001b) lassen sich wie folgt bestimmen (Grice, 2001b, Gleichung (9), S. 433; die Notation erfolgt in Anlehnung an Mulaik, 2009, Anm. d. Autorin):

$$\widehat{\mathbf{X}}_\mathbf{i} = \widehat{\mathbf{Y}}_\mathbf{i}\, \boldsymbol{\Psi}^{-\mathbf{2}}\, \boldsymbol{\Lambda}\, (\boldsymbol{\Lambda}'\, \boldsymbol{\Psi}^{-\mathbf{2}}\, \boldsymbol{\Lambda})^{-\mathbf{1}} \qquad (10)$$

Die Bartlett-Faktorwerte werden mittels Minimierung der Quadratsummen der uniquen Faktorwerte berechnet (DiStefano, Zhu, & Mindrila, 2009). Diese Art der Berechnung resultiert in dem Vorteil, dass die Bartlett-Faktorwerte erwartungstreue Schätzer für die wahren Faktorwerte sind (Lawley & Maxwell, 1971, S. 110). Ein Schätzer heißt erwartungstreu, wenn sein Erwartungswert dem wahren Wert des geschätzten Parameters entspricht (Kauermann & Küchenhoff, 2011, S. 21). Die Regressions-Faktorwerte erfüllen dieses Kriterium nicht[5]. Sowohl die Thurstone-Faktorwerte, als auch die Bartlett-Faktorwerte stellen standardisierte Faktorwerte dar (DiStefano et al., 2009).

Kriterien für die Güte der geschätzten Faktorwerte sind nach Grice (2001a, S. 68, 2001b, S. 436) Validität, Eindeutigkeit und Orthogonalität. Die Validität der geschätzten Faktorwerte wird nach Grice (2001a, 2001b) an der Korrelation zwischen den geschätzten und den wahren Faktorwerten festgemacht. Eindeutigkeit beschreibt, inwieweit die bestimmten Faktorwerte mit den wahren Faktorwerten anderer Faktoren korrelieren. Orthogonalität bedeutet übertragen auf den Fall eines obliquen Designs, wie es auch in den Studien 1 und 2 verwendet wurde, dass die Uniqueness-Matrix des Faktorenmodells eine Diagonalmatrix darstellen sollte. Die Bartlett-Faktorwerte erfüllen alle drei Kriterien und sind, wie bereits beschrieben, erwartungstreu (McDonald & Burr, 1967). Maathuis (2008) empfiehlt zudem insbesondere die Bartlett-Faktorwerte, wenn Interesse an einer bestimmten Gruppe von Individuen besteht. Im Rahmen der zweiten Studie bestand genau dieses Interesse, es wurden Gruppen von Individuen mit positiven und negativen

[5]Für eine Herleitung siehe Lawley und Maxwell (1971), S. 108-111.

Diagnosen aus den Faktorwerten gebildet. Insofern wurden bei der zweiten Studie (siehe V) die Bartlett-Faktorwerte verwendet.

Die Problematik bei der Schätzung der Faktorwerte[6] liegt in deren sog. „Indeterminacy“ (Mulaik, 2009, S. 375ff). Diese Unbestimmtheit bedeutet, dass es eine unendliche Anzahl an gleichwertigen Lösungen für die Berechnung der Faktorwerte **X** und der uniquen Faktorwerte **E** bei derselben Faktorlösung gibt (Beauducel, 2005; Grice, 2001), wobei meist nur die Faktorwerte **X** der gemeinsamen Faktoren von Interesse sind. Für die Studie 2 spielt die Faktorenunbestimmtheit eine Rolle, da unter anderem auch die Ladungshöhe (Reliabilität) für die Generierung der Populationsdaten variiert wurde. Der Grad der Faktorenunbestimmtheit hängt von der Höhe der Ladungen ab (siehe ρ und Formel (11) auf der nächsten Seite). Dementsprechend hängt auch die Validität der Faktorwerte (für den Begriff vgl. Grice [2001a, 2001b]) von der Höhe der Faktorladungen ab (Erklärung folgt unter V. 2.2.1), aus denen in der zweiten Studien Diagnosen gebildet wurden. Daher wird im Folgenden die Faktorenunbestimmtheit sowohl algebraisch als auch geometrisch illustriert; ebenso die Konvention für die Bestimmung des Grades der Faktorenunbestimmtheit nach Guttman (1955).

Eine simple Betrachtung der Fundamentalgleichung der Faktorenanalyse (Gleichung (1) unter II. 1) verdeutlicht die Faktoren-Unbestimmtheit: Die linke Seite der Gleichung hat n Parameter, nämlich die Anzahl der beobachteten Variablen. Diese n Parameter sollen jedoch durch n + r Parameter, durch die Summe der beobachteten Variablen und der latenten Faktoren, vorhergesagt werden, was zu einer Gleichung mit zwei Unbekannten und insofern mehreren Lösungen führt (vgl. Grice, 2001b; Mulaik, 2009; Schönemann, 1996).

[6]Bei der Bestimmung der Faktorwerte handelt es sich nicht um ein Schätzproblem im klassischen Sinne, welches keine exakte Lösung hervorbringt und deshalb eine Annäherung zum Ziel hat (Guttmann, 1972, zitiert nach Steiger & Schönemann, 1978, S. 137). Maraun (1996a) kritisiert daher die Verwendung der Begriffe „Faktorwerte“, „geschätzte Faktorwerte“ oder „Personenparameter“ in semantischer Hinsicht und empfiehlt den Begriff der „aus den manifesten Variablen vorhergesagten latenten Variablen“ (S. 518f). Trotz Marauns Kritik wird in der vorliegenden Arbeit weiterhin von „geschätzten Faktorwerten“ gesprochen. Der Begriff der „Schätzung der Faktorwerte“ im Rahmen dieser Arbeit soll neben der existierenden Konvention im Sprachgebrauch illustrieren, dass es verschiedene Methoden zur Berechnung der Faktorwerte gibt (vgl. Beauducel, 2005, S. 143).

Wilson (1928, zitiert nach Steiger & Schönemann, 1978), welcher als erstes über das Unbestimmtheitsproblem der Faktorwerte geschrieben hatte, erarbeitete eine geometrische Darstellung des Unbestimmtheitsproblems, indem er Spearmans g (Spearman, 1904, 1927) als Beispiel heranzog. Er veranschaulichte dies anhand eines Kegels, bei dem die Kegelachse die wahren Faktorwerte darstellt (Guttman, 1955; Mulaik, 2009; Steiger & Schönemann, 1978). Die Regressions-Faktorwerte, die lineare Schätzer für die wahren Faktorwerte sind, stellen die Mittelwerte aus den möglichen Faktorwertlösungen dar. Der Cosinus des Winkels zwischen dem Vektor der wahren Faktorwerte, also der Kegelachse, und allen anderen möglichen Faktorwerten, die den Kegelmantel aufspannen, bestimmt die Korrelation zwischen diesen Größen[7]. Daraus wird ersichtlich, dass eine unendliche Zahl an möglichen Faktorwerten durch das lineare Modell errechnet werden kann. Auf die Schätzung der Modellparameter hingegen hat die Faktorenunbestimmtheit keine Auswirkungen (Beauducel, 2007).

Für eine detailliertere Auseinandersetzung mit der Geschichte der Faktoren-Unbestimmtheit werden Steiger und Schönemann (1978), Grice (2001a) sowie Steiger (1996) empfohlen. Technisch empfehlenswerte Bearbeitungen der Faktoren-Unbestimmtheit finden sich bei Guttman (1955), Schönemann und Steiger (1978) sowie Maraun (1996a, 1996b).

Guttman (1955) quantifizierte den Grad der Faktorenunbestimmtheit. Er erarbeitete eine Formel, um die maximal mögliche Unbestimmtheit zwischen zwei alternativen Faktorwertlösungen für denselben Faktor zu berechnen (Guttman, 1955, S. 73):

$$\rho^* = 2\rho^2 - 1 \qquad (11)$$

ρ stellt die multiple Korrelation zwischen dem latenten Faktor und den beobachteten Variablen dar (Guttman, 1955). An dieser Stelle wird ersichtlich, dass der Grad der Faktorenunbestimmtheit eines Faktors nach Guttman alleine durch die Höhe der Faktorladungen determiniert ist. Da diese im Rahmen der zweiten Studie variiert wurde, wurde Guttmans Quantifizierung des Grades der Faktorenunbestimmtheit auch im Rahmen der zweiten Studie bestimmt (siehe V. 2.2.1 und 2.3.1). ρ^2 liegt zwischen 0 und

[7]In einem Euklidischen Vektorraum entspricht der Cosinus des Winkels zwischen zwei Vektoren der Korrelation zwischen diesen beiden Vektoren (Guttman, 1955).

1, wobei ein Wert nahe 1 wünschenswert ist, da er eine hohe Bestimmtheit darstellt (Grice, 2001). Dass ρ^2 bzw. ρ nie ganz Eins werden kann, liegt an der Unbestimmtheit der Faktorwerte an sich (Mulaik, 2009). ρ^* ist die Korrelation zwischen einem möglichen Vektorelement, also einem möglichen Faktorwert, und einem korrespondierenden Vektorelement bzw. einem Faktorwert, der maximal verschieden ist von ersterem. ρ^* bewegt sich zwischen -1 und 1. In Wilsons (1928, zitiert nach Steiger & Schönemann, 1978) geometrischer Veranschaulichung entspricht ρ^* dem Cosinus des Winkels zwischen zwei Faktorwertvektoren, die zusammen mit den anderen möglichen Faktorwertvektoren den Kegelmantel aufspannen, welche aber auf derselben Ebene liegen und insofern maximal weit voneinander entfernt sind.

Die Quantifizierung der Faktoren-Unbestimmtheit Guttmans (1955) zeigt, dass die individuellen Faktorwerte der Personen komplett unterschiedlich sein können, je nachdem, welches Aggregat an Faktorwerten herangezogen wurde (Grice, 2001b). Sofern ρ kleiner als oder gleich .71 ist, kann die Korrelation zwischen verschiedenen Faktorwertlösungen negativ werden (Guttman, 1955) und in substanzieller[8] Hinsicht in die umgekehrte Richtung gehen.

[8]Der Begriff „substanziell“ meint an dieser Stelle die psychometrische Interpretation der individuellen Faktorwerte.

III ABLEITUNG DER FRAGESTELLUNGEN

1 Missspezifikationsforschung

Im vorherigen Kapitel II wurde die Konzeption von Strukturgleichungsmodellen und die Modelltestung anhand des χ^2-Tests und der Fit Indizes dargestellt. In diesem Zuge wurde bereits berichtet, dass die kritische Frage bei der Modelltestung lautet, wie zuverlässig Missspezifikationen, als solche erkannt werden. Auf diese Frage bezieht sich die erste Studie dieser Dissertation.

Die Power/Sensitivität des χ^2-Tests sowie der Fit-Indizes wurde innerhalb der letzten 20 bis 30 Jahre ausführlich hinsichtlich unterschiedlicher Stichprobengrößen, Schätzmethoden und Arten von Missspezifikationen beforscht. Für die Überprüfung der Power/Sensitivität der Fit-Indizes eignet sich das Monte-Carlo-Simulationsverfahren[9] (vgl. Paxton, Curran, Bollen, Kirby, & Chen, 2001), welches auch im Rahmen der Studie 1 verwendet wurde. Dabei werden zunächst Populationsmodelle definiert. Aus diesen spezifizierten Populationsmodellen lassen sich Populations-Kovarianz-Matrizen darstellen. Aus diesen Populationsmodellen oder Populations-Kovarianz-Matrizen wird per Zufall/stichprobenfehlerbedingt eine große Anzahl von Stichprobendaten aus zufällig variierenden Stichproben-Kovarianz-Matrizen erzeugt. Diese so simulierten Stichprobendaten werden dann anhand eines missspezifizierten Modells analysiert. Abschließend wird ausgezählt, für wie viele Stichproben die Fit-Indizes ein missspezifiziertes Modell korrekterweise als solches anzeigen.

Es wäre zwar wünschenswert, die Varianz der Fit-Indizes würde nur von Missspezifikationen beeinflusst werden, jedoch zeigt die Forschungslage, dass die Werte der Fit-Indizes in hohem Maße aufgrund anderer Modellbedingungen variieren, die nichts mit der Missspezifikation zu tun haben (Saris, Satorra, & van der Veld, 2009).

[9]Der Unterschied einer Monte-Carlo-Simulation im Vergleich zu einer sog. Resampling-Methode liegt darin, dass bei ersterer der Daten-Generierungs-Prozess der wahren Population bekannt ist (Carsey & Harden, 2014, S. 4). Dabei werden vom Forschenden alle Aspekte des wahren Populations-Daten-Generierungs-Prozesses kontrolliert, was im genannten Kontext ermöglicht, die Fit-Indizes auf ihre Sensitivität hin auszutesten. Resampling-Methoden hingegen ziehen Stichprobendaten aus empirisch erhobenen Daten, der wahre Populations-Daten-Generierungs-Prozess ist unbekannt.

Es ist augenfällig, dass weniger Simulationsstudien zu den Auswirkungen von Missspezifikationen auf die Fit-Indizes als zum Beispiel zu Stichprobengrößen und Schätzmethoden durchgeführt wurden, die auch Einfluss auf die Fit-Indizes nehmen (Fan et al., 2009). Die Gründe dafür sehen Fan et al. sowie Gerbing und Anderson (1993) darin, dass der Grad einer Missspezifikation nur schwer zu bestimmen sei und Missspezifikationen deshalb schwierig zu quantifizieren seien.

Grundsätzlich wird versucht, Missspezifikationen nach ihrem Typ und ihrem Schweregrad einzuteilen (Fan et al., 2009). Es wird zwischen Missspezifikationen im Messmodell, also zwischen latenter Variable und deren messfehlerbehafteten Indikatoren, sowie Missspezifikationen im Strukturmodell, also zwischen latenten Variablen, unterschieden. Fan und Sivo (2005) empfehlen, den Schweregrad einer Missspezifikation anhand des Nonzentralitätsparameters unter Berücksichtigung der Freiheitsgrade des Modells zu bestimmen. Der Nonzentralitätsparameter indiziert, wie weit die nonzentrale χ^2-Verteilung unter der Alternativhypothese einer Modellabweichung von der zentralen χ^2-Verteilung unter der Nullhypothese der Modellpassung abweicht (Kaplan, 1988). Mit steigendem Nonzentralitätsparameter steigt also der Schweregrad der Missspezifikation. Desto größer der Nonzentralitätsparameter, desto linkssteiler und flacher ist die Verteilung (Erdfelder, Faul, Buchner, & Cüpper, 2010). Bei festgelegtem Alpha-Fehler steigt die Power bei steigendem Nonzentralitätsparameter. Der Nonzentralitätsparameter hängt ferner von der Stichprobengröße ab. Der Vorteil der Bestimmung des Grades der Missspezifikation anhand des Nonzentralitätsparameters bestehe nach Fan und Sivo (2005) darin, dass dieser unabhängig von der Art der Missspezifikation bestimmt werden könne. Curran, Bollen, Paxton, Kirby, und Chen (2002) zeigten allerdings, dass selbst korrekt spezifizierte Modelle nur einer zentralen χ^2-Verteilung folgten, wenn die Stichproben mittelgroß oder groß waren (N = 200 bis N = 1,000). Des Weiteren zeigten die Autoren, dass schwer missspezifizierte Modelle (operationalisiert durch nicht-spezifizierte [Neben-]Faktorladungen unterschiedlicher Anzahl, aber gleicher Faktorladungshöhe) oder auch ein unkorreliertes Basismodell keiner non-zentralen χ^2-Verteilung folgten, genauso wenig jedoch einer zentralen χ^2-Verteilung. Letzterer Befund war unabhängig von der Stichprobengröße. Weiters zeigte sich, dass weniger schwer missspezifizierte Modelle bei kleinen Stichproben auch keiner non-zentralen χ^2-Verteilung folgten, bei mittleren und größeren Stichproben (mindestens N = 200 bis N = 1,000) jedoch schon. Allerdings zeigte sich, trotz der Tatsache, dass die (Non-)Zentralität der erwarteten Verteilungen nicht in allen Modellbedingungen den Vermutungen entsprach, dass die

Varianzen der Verteilungen bei der Anwendung missspezifizierter Modelle stark streuten. Insofern ist zu vermuten, dass die Einordnung von Missspezifikationen in verschiedene Schweregrade anhand des Nonzentralitätsparameters auch nicht unproblematisch ist. Zudem bestätigt dieser Befund, dass das Problem der Quantifizierung einer Missspezifikation weder gelöst ist, noch Konsens hinsichtlich der Herangehensweise besteht. Dennoch wurden diverse Studien durchgeführt, die die Auswirkungen von Missspezifikationen unterschiedlicher Operationalisierungen auf die Fit-Indizes, also den globalen Modellfit, überprüften. Im Folgenden werden die wichtigsten dieser Studien aufgeführt und anhand weiterer Befunde diskutiert, bevor anschließend speziell die Studien aufgeführt werden, aus denen sich die erste Forschungsfrage (siehe III. 3) ableitet.

Hu und Bentler (1998, 1999) führten zwei einflussreiche Simulationsstudien durch, die die Sensitivität gängiger Fit-Indizes gegenüber Missspezifikationen im Mess- und im Strukturmodell austesten sollten. Aus diesen Studien gingen die unter II. 2 genannten Daumenregeln hervor. Hu und Bentler (1999) untersuchten unterschiedliche Fit-Indizes hinsichtlich zwei verschiedener Schweregrade sowie Typen von Missspezifikationen (fälschlicherweise nicht spezifizierte Ladungen auf mehr als einen Faktor [einfache Missspezifikation bzw. Missspezifikation im Messmodell]) versus fälschlicherweise nicht spezifizierte Faktorkovarianzen ungleich Null [komplexe Missspezifikation bzw. Missspezifikation im Strukturmodell] bei einem drei-Faktorenmodell mit insgesamt 15 Indikatoren), welche alle eine Unterparametrisierung in den missspezifizierten Modellen darstellten. Die Stichprobengröße variierte zwischen 150 und 5,000 Fällen. Hu und Bentler (1999) schlussfolgerten aus ihren Ergebnissen, dass es sich empfehlen würde, eine Kombination aus einem der Fit-Indizes RNI (*Relative Noncentrality Index*; Bentler & Bonett, 1980), TLI (*Tucker-Lewis-Index*; Bentler & Bonett, 1980; Tucker & Lewis, 1973), CFI (Bentler, 1990) oder RMSEA (Steiger & Lind, 1980) zusammen mit dem SRMR (Jöreskog & Sörbom, 1981) für die deskriptive Evaluation des Model-Fit zu verwenden, da erstere Fit-Indizes sensitiv für Missspezifikationen im Messmodell wären und letzteres sensitiv für ein missspezifiziertes Strukturmodell. Dafür schlugen sie die unter II. 2 genannten Cut-Off-Regeln vor, die mehr und mehr zu einer dichotomen Entscheidung hinsichtlich der Beurteilung der Modellpassung führten (Marsh et al., 2004). Diese Cut-Offs nach Hu und Bentler (1999) werden im Folgenden vor dem Hintergrund weiterer Forschung diskutiert.

Eine berechtigte Kritik an Hu und Bentlers (1998, 1999) Studien ist, dass Art und Grad der Missspezifikationen konfundiert waren (Fan et al., 2009). Insofern sind die Schlussfolgerungen der Autoren insbesondere für missspezifizierte Modelle nur sehr eingeschränkt haltbar (Fan & Sivo, 2005; Fan et al., 2009). Dies betrifft sowohl die propagierte Sensitivität der Fit-Indizes für die Art der Missspezifikation, als auch die empfohlenen Cut-Off-Werte.

Olsson et al. (2000) zogen aus ihrer Simulationsstudie die Schlussfolgerung, dass empirische Fit-Indizes nicht über verschiedene Schätzmethoden hinweg verglichen werden könnten. Dementsprechend wären die verschiedenen Fit-Indizes nicht austauschbar, unter anderem, weil zum Beispiel der CFI Maximum-Likelihood-basiert ist, der RMSEA jedoch nicht. Olsson et al. argumentierten schlussendlich, dass die Cut-Offs nach Hu und Bentler auch aufgrund der mangelnden Vergleichbarkeit der Fit-Indizes nicht angemessen wären. Ding, Velicer, und Harlow (1995) kamen zu einem ähnlichen Ergebnis, nämlich, dass die Werte der Fit-Indizes durch die Schätzmethode im Rahmen ihrer Simulationsstudie verzerrt waren. Sie variierten ebenfalls die Schätzalgorithmen (Maximum-Likelihood versus General Least Squares), sowie die Anzahl an Indikatoren pro Faktor und die Höhe der Faktorladungen. Die Ergebnisse in Bezug auf den Faktor der Schätzalgorithmen zeigten, dass unter Maximum-Likelihood zwar weniger Heywood Cases[10] auftraten, die Parameterschätzungen aber verzerrter waren als unter General Least Squares. Fan et al. (2009) zogen aus ihren Befunden ähnliche Implikationen wie Olsson et al. (2000), nämlich, dass die Fit-Indizes über verschiedene Schätzmethoden hinweg nicht vergleichbar und damit nicht austauschbar wären. Fan et al. (2009) untersuchten zehn Fit-Indizes auf ihre Sensitivität hinsichtlich Missspezifikationen im Messmodell entweder in Form von fälschlicherweise auf Null gesetzten Faktorladungen oder fälschlicherweise hinzugefügten Faktorladungen ungleich Null, was eine Verletzung der Einfachstruktur darstellte. Es zeigte sich – und dies ist konsistent zu den Befunden von Hu und Bentler – dass der RMSEA am sensitivsten gegenüber beiden Missspezifikationen, also Missspezifikationen im Messmodell, war. CFI und SRMR schnitten diesbezüglich im Vergleich mittelmäßig ab.

[10]Heywood Cases stellen ungültige Lösungen dar (Faktorladungen über Eins oder negative Fehlervarianzen bei Standardisierung), welche aus Schätzproblemen, Missspezifikationen oder auch aus Sampling Fluktuationen resultieren; für eine ausführlichere Erklärung sowie Vorschläge für den Umgang mit Heywood Cases bei der Datenanalyse wird Kolenikov und Bollen (2012) und Dillon, Kumar, und Mulani (1987) empfohlen.

Rigdon (1996) verglich RMSEA und CFI und zeigte analytisch, dass der Gebrauch des CFI problematisch wäre, weil er, wie alle inkrementellen Fit-Indizes, von einem geeigneten Nullmodell abhängt. Alternative Baseline-Modelle wurden im Artikel beschrieben. Der Autor empfahl die Betrachtung des CFI daher im Rahmen von eher explorativen Kontexten und kleinen Stichprobengrößen. Hu und Bentlers (1998, 1999) Studien waren konfirmatorischer Art. Den RMSEA schlug Rigdon für konfirmatorische Ansätze mit größeren Stichproben vor, unter anderem, weil der RMSEA als weniger abhängig von der Stichprobengröße gilt.

Eine weitere angemessene Kritik ist, dass Hu und Bentler in ihren Studien sehr hohe Faktorladungen verwendeten (standardisiert zwischen .70 und .80), welche jedoch im Rahmen der meisten psychologischen Untersuchungen und Testvalidierungen nicht erreicht werden (Fan & Sivo, 2005; Fan et al., 2009; Heene et al., 2011). Peterson (2000) verglich in einer Meta-Analyse Faktorladungen von 803 exploratorischen Faktorenanalysen aus Psychologie- und Marketing-Journals. Er ermittelte eine durchschnittliche standardisierte Faktorladungshöhe von .32, wobei 25% der berichteten Faktorladungen unter .23 lagen und 25% der berichteten Faktorladungen über .37. Der Anteil der erklärten Varianz durch die Faktorenmodelle lag bei durchschnittlich knapp 57%. Diese Kritik lässt die Schlussfolgerung zu, dass die Cut-Offs nach Hu und Bentler (1998, 1999) auch aufgrund unterschiedlicher Faktorladungshöhen in der angewandten Forschung nicht verallgemeinert werden können (Heene et al, 2011).

Heene et al. (2011) untersuchten ein zu Hu und Bentler (1998, 1999) vergleichbares Design, jedoch ohne Konfundierungen zwischen Art und Grad der Missspezifikationen, und erweiterten das Design noch um den Faktor Testlänge. Außerdem wurden anstatt von tau-äquivalenten Indikatoren kongenerische Indikatoren verwendet, deren Faktorladungen zwischen gering, mittelhoch und hoch variierten, wobei die hohen Faktorladungen der von Hu und Bentler (1998, 1999) getesteten Höhe an Faktorladungen entsprach. Bei den komplex missspezifizierten Modellen wurden, wie bei Hu und Bentler, die Faktorkovarianzen auf Null gesetzt, bei den einfach missspezifizierten Modellen hatten drei Indikatoren, ähnlich zu Hu und Bentler, Doppelladungen im Vergleich zum Populationsmodell. Die Stichprobengrößen variierten zwischen 150 und 2,500 Fällen. Die Ergebnisse zeigten einen Haupteffekt für die Höhe der Faktorladungen: Sowohl die χ^2-Werte, als auch die Werte der Fit-Indizes sanken bei sinkenden Ladungen. Das heißt, eine Modellevaluation durch den CFI führte bei geringen Ladungen eher dazu, dass missspezifizierte Modelle verworfen wurden. Bei hohen Ladungen jedoch wurden zu viele Modelle durch den CFI anhand des Cut-Offs nach Hu und Bentler als

passend indiziert. Dass dieser Befund bei Hu und Bentler nicht auftrat, kann durch die Heterogenität der Faktorladungen im Vergleich zu den homogenen Faktorladungen bei Hu und Bentler erklärt werden (Heene et al., S. 322). RMSEA und SRMR[11] akzeptierten bei geringen Faktorladungen zu oft missspezifizierte Modelle, wohingegen sie bei höheren Faktorladungen sensitiver reagierten. Erklärt werden kann dieses Phänomen dadurch, dass bei geringen Faktorladungen und dementsprechend großen uniquen Varianzen der χ^2-Test an Power verliert (Heene et al., S. 328f). Somit verlieren auch die Fit-Indizes CFI und RMSEA, da sie vom χ^2-Wert abhängen, wie auch der SRMR an Sensitivität. Die Befunde sprechen dafür, dass die Cut-Offs von Hu und Bentler für eine realistische Faktorladungshöhe zu niedrig gewählt wurden, sofern man sie als allgemeingültige Regeln verwenden wollte.

Während die Befunde von Heene et al. (2011) eher für noch strengere Cut-Offs sprechen, vertreten Marsh, Hau, und Wen (2004) eine gegensätzliche Meinung. Sie argumentierten, die konventionellen Cut-Offs nach Hu und Bentler wären zu strikt. Marsh et al. schrieben, dass es – ihrer Erfahrung nach (S. 325) – fast unmöglich wäre, mit diesen Cut-Off-Regeln einen akzeptablen Modellfit zu erhalten, selbst, wenn die Testinstrumente psychometrisch gut wären und ausreichend Items zur Abbildung der latenten Faktoren vorhanden wären. Die Autoren argumentierten ferner, dass sich die Klassifikationsregeln nach Hu und Bentler paradox verhalten würden (Marsh et al., S. 327): Die Wahrscheinlichkeit, ein missspezifiziertes Modell zu verwerfen wäre für mindestens eines der Modelle, die Hu und Bentler spezifizierten, geringer für größere Stichproben als für kleinere. Dies wäre laut Marsh et al. vermeidbar gewesen, wenn die Autoren die Cut-Offs anders gesetzt hätten. Marsh et al. kritisierten, dass, sofern Hu und Bentler im Rahmen ihres Designs noch noch extremer missspezifizierte Modelle verwendet hätten, Cut-Offs von .80 für inkrementelle Fit-Indizes sehr exakt zwischen diesen und korrekt spezifizierten Modellen hätten diskriminieren können. Nach Marsh et al. hätten Hu und Bentler missspezifizierte Modelle spezifiziert, die theoretisch durchaus akzeptabel gewesen wären. In der typischen Praxis wären Modelle hingegen deutlich missspezifizierter.

An Marsh et al.s (2004) Argumentationsführung wird deutlich, dass nicht nur kein Konsens darüber besteht, wie Missspezifikationen hinsichtlich Art und Schweregrad einzuordnen sind, sondern vielmehr auch, ab wann ein Modell überhaupt als missspezifiziert

[11] An dieser Stelle ist zu beachten, dass im Rahmen dieser Studie für das SRMR ein Cut-Off von < .11 verwendet wurde.

gilt. Letztere Frage wird jedoch nicht alleine anhand von Simulationsstudien zu klären sein. Hierfür bedarf es den Bezug auf vielfach replizierte substanzielle Theorien und insbesondere auch die kriterielle Evaluation der Schlussfolgerungen, die aus diesen Theorien gezogen werden sollen.

Trotz der Unstimmigkeiten in Bezug auf die Quantifizierung von Missspezifikationen sind sich die Autoren der oben genannten Studien insofern einig, als dass die von Hu und Bentler vorgeschlagenen Cut-Off-Werte zur Beurteilung der Modellpassung nicht als allgemeingültige Regeln angewendet werden können (Fan & Sivo, 2005; Fan et al., 2009; Heene et al., 2011; Marsh et al., 2004). Die Fit-Indizes sind nicht alleine abhängig von der Größe der Modellabweichung, sondern auch von anderen Faktoren, wie Stichprobengröße, Schätzalgorithmus oder Faktorladungshöhe (Beauducel & Wittmann, 2005; Fan & Sivo, 2005; Fan et al., 2009; Heene et al., 2011; Olsson, Foss, Troye, & Howell, 2000; Yang & Green, 2010).

Fan und Sivo (2005) argumentierten, dass noch nicht ausreichend verschiedene Modelle getestet worden wären und somit der Grad der Missspezifikation noch nicht hinreichend kontrolliert worden wäre, um Aussagen über die Sensitivität der Fit-Indizes bezüglich verschiedener Typen und Grade von Missspezifikationen treffen zu können. Eine Schlussfolgerung aus diesen Forschungsbemühungen ist unter anderem, dass die Fit-Indizes für korrekt spezifizierte Modelle vergleichbar sind, für missspezifizierte Modelle aber nicht (Fan et al., 2009). Dementsprechend ist offensichtlich, dass die Cut-Off-Kriterien zur Beurteilung des Modellfits von Hu und Bentler insbesondere bei missspezifizierten Modellen nicht haltbar sind. Wie bereits von Marsh et al. (2004) angedeutet, liegt allerdings die Vermutung nahe, dass missspezifizierte Modelle den Großteil der in der psychologischen Anwendung spezifizierten Modelle repräsentieren.

2 Mehrdimensionalität

2.1 Itemparcels

Innerhalb der Psychologie und verwandter Disziplinen stellt eine nicht-spezifizierte Mehrdimensionalität eine typische Missspezifikation dar (Heene et al., 2012; Little et al., 2002b). Dennoch besteht insgesamt noch wenig Forschung zur Sensitivität der Fit-Indizes für diese Formen der Missspezifikation. Mehrdimensionalität im Kontext der Psychometrie bedeutet, dass ein Test, Fragenbogen oder Item systematisch nicht nur ein Merkmal misst, sondern mehrere[12]. Die Klassische Testtheorie, auf der konfirmatorische Faktorenanalysen und Strukturgleichungsmodelle basieren, ist allerdings für eindimensionale Tests, Fragebögen oder Items konzipiert (Bortz & Döring, 2002/2003). Insofern stellen mehrdimensionale Tests, Fragebögen oder Items, sofern nicht als solche spezifiziert, eine fundamentale Verletzung der statistischen Eindimensionalitätsannahme dar (Bühner, 2011).

Innerhalb der Forschung zu konfirmatorischen Faktorenanalysen und Strukturgleichungsmodellen erregten hinsichtlich der Frage nach der Dimensionalität unter anderem gruppierte (sog. geparcelte [Bandalos, 2000, S. 78]) Items Aufmerksamkeit. Bandalos untersuchte die Auswirkungen von geparcelten Items im Rahmen von Strukturgleichungsmodellen auf die Parameterschätzungen und auf die Fit-Indizes CFI und RMSEA. Sie untersuchte im Rahmen einer ersten Studie ein Populationsmodell mit zwei exogenen latenten Variablen und je 12 Indikatoren und einer endogenen latenten Variablen, die durch 6 Items abgebildet wurde. Alle Indikatoren genügten der Einfachstruktur. Es wurden sowohl normalverteilte als auch nicht normalverteilte Items inkludiert. Die standardisierten

[12]Einfachstruktur hingegen bedeutet, dass Items in Faktorenanalysen möglichst hoch auf einen Faktor, aber möglichst niedrig auf alle anderen Faktoren laden sollten (Bühner, 2011, S. 204). Einfachstruktur garantiert nicht psychologische Eindimensionalität, die nicht gleichzusetzen ist mit statistischer Eindimensionalität. Psychologische Eindimensionalität meint die phänomenologische Eindimensionalität des Messgegenstands. Falls beispielsweise alle Indikatoren, die eine latente Variablen messen sollen, zwei Eigenschaften/Fähigkeiten messen, diese aber in einem konfirmatorischen Modell nicht in Form von Nebenladungen auf eine zweite latente Variable spezifiziert werden, bildet die eine latente Variable psychologisch zwei Eigenschaften/Fähigkeiten ab. In diesem Fall würde statistische Eindimensionalität vorliegen, psychologische Eindimensionalität allerdings nicht. Im Folgenden ist immer, sofern nicht anders gekennzeichnet, von statistischer Ein- und Mehrdimensionalität – im Gegensatz zu psychologischer Ein- oder Mehrdimensionalität – die Rede.

Ladungen der Indikatoren wurden auf .70 festgesetzt, die standardisierten Messfehlervarianzen auf .30. Die Parcels aus den Indikatoren, die aus 2, 3, 4 oder 12 (Gesamtwert) Indikatoren bestanden, wurden anhand der Verteilung der Indikatoren gebildet, woraus Parcels mit normalverteilten Indikatoren, Parcels mit normal- und mit nicht normalverteilten Indikatoren, sowie Parcels, deren Indikatoren schiefer und gewölbter verteilt waren als normalverteilte Indikatoren, resultierten. Die Stichprobengrößen variierten zwischen 100 und 800 Fällen. Die Modelle mit den geparcelten Indikatoren führten insgesamt betrachtet zu niedrigeren Werten des RMSEA und höheren Werten des CFI, also zu besserem Fit, als das korrekte Modell ohne Parcels. Nur für das Modell mit den 12-Indikatoren-Parcels war der RMSEA hingegen leicht höher als beim Modell mit den individuellen Indikatoren. Beim CFI zeigten sich aufgrund von insgesamt sehr hohem Fit nur geringfügig Unterschiede bei kleinen Stichprobengrößen, hier fiel die Modellpassung auch zugunsten der individuell verwendeten Indikatoren aus. Die Ergebnisse wurden allerdings nicht mehr zwischen der Anzahl der Indikatoren in den Parcels getrennt ausgewertet.

Im Rahmen ihrer zweiten Studie spezifizierte Bandalos (2002) in der Grundstruktur dasselbe Populationsmodell wie im Rahmen ihrer ersten Studie, allerdings mit dem wichtigen Unterschied, als dass jeweils sechs Indikatoren der exogenen latenten Variablen gleichzeitig einen dritten Faktor widerspiegelten. Insgesamt hatten also 12 Indikatoren Nebenladungen, die auf .40 (standardisiert) gesetzt wurden. Diese Indikatoren folgten also keiner Einfachstruktur. Die Indikatoren wurden wieder in Parcels zu 2, 3, 4 oder 12 Indikatoren zusammengefasst, wobei unterschieden wurde zwischen Parcels mit Indikatoren, die nur auf eine latente Variable luden oder nur auf zwei latente Variablen luden (isoliert; S. 87), oder Parcels, die sowohl Indikatoren ohne Nebenladungen als auch Indikatoren mit Nebenladungen enthielten (aufgeteilt; S. 87). Die Stichprobengrößen waren die gleichen wie in der ersten beschriebenen Studie der Autorin. Die Lösungen für die Itemparcels wurden verglichen mit dem missspezifizierten Modell, bei dem die dritte exogene latente Variable, und somit die Nebenladungen der Indikatoren, nicht spezifiziert wurden. Es zeigte sich, dass der RMSEA für die Lösungen mit aufgeteilten Parcels kleiner war als für die Lösungen mit isolierten Parcels, der RMSEA zeigte also bei den aufgeteilten Parcels besseren Fit an. Die Modellpassung anhand des RMSEA bei den ungeparcelten Indikatoren im missspezifizierten Modell lag zwischen den RMSEA-Werten für die aufgeteilten und isolierten Parcels. Insgesamt lagen die Werte des RMSEA entweder unter oder knapp über der Cut-Off-Regel von Hu und Bentler (1998, 1999). Der CFI zeigte bessere Modellpassung bei den Parcels bestehend aus 12 Indikatoren im

Vergleich zu Parcels bestehend aus weniger Indikatoren; ebenso, wie der RMSEA, bei den aufgeteilten Parcels im Vergleich zu den isolierten Parcels (Bandalos, 2002). Die individuellen Indikatoren im missspezifizierten Modell führten zu niedrigeren CFI-Werten als die geparcelten Indikatoren. Insgesamt lag der CFI jedoch immer über .90 und, sofern die aufgeteilten Parcels verwendet wurden, über .95, also über dem Cut-Off nach Hu und Bentler.

Die Schlussfolgerung aus Bandalos' (2002) Studien lautete, dass die statistischen Dimensionen, auf die die Indikatoren luden, für das Parceling eine entscheidende Rolle spielten. Sofern die Indikatoren der Einfachstruktur genügten, konnte das Zusammenfassen der Indikatoren in höherer Modellpassung hinsichtlich der Fit-Indizes resultieren, und dies selbst bei nicht normalverteilten Indikatoren. Sofern jedoch Indikatoren mit Nebenladungen auftraten, welche nicht als solche spezifiziert wurden, maskierten aufgeteilte Itemparcels diese Missspezifikation sogar noch im Vergleich zu nicht-geparcelten Indikatoren. Sie führten zu akzeptablen Werten der Fit-Indizes, sofern die Kriterien nach Hu und Bentler (1998, 1999) zur Beurteilung der Modellpassung herangezogen wurden (Bandalos, 2002).

Marsh, Lüdtke, Nagengast, Morin, und von Davier (2013) untersuchten Itemparcels zunächst anhand von zwei empirischen Datensätzen, welche sie anschließend mit zwei simulierten Datensätzen verglichen. Der erste empirische Datensatz bestand aus zehn beobachteten Variablen, die eine Dimension anhand einer Skala messen sollten. Die Daten stammten von 2,175 Schülerinnen und Schülern aus unterschiedlichen Messzeitpunkten. Itemparcels wurden einerseits anhand der Mittelwerte zusammengestellt, andererseits anhand nur positiv oder nur negativ formulierter Indikatoren (homogene Parcels; S. 263), sowie anhand positiv und negativ formulierter Indikatoren (aufgeteilte Parcels; S. 263). Wurde ein korrektes Strukturgleichungsmodell auf die empirischen Daten angewandt, welches die unterschiedlich formulierten Indikatoren nicht in Form von Methodenfaktoren berücksichtigte, resultierte daraus ein schlechter Fit, das korrekte Modell wurde an den Daten abgelehnt. Die Modelle mit den aufgeteilten Parcels führten zur Akzeptanz des Modells nach den Kriterien von Hu und Bentler (1998, 1999). Die homogenen Itemparcels führten zu schlechterem Fit, allerdings war der Fit anhand der konventionellen Kriterien immer noch gut (Marsh et al., 2013).

Der zweite empirische Datensatz bestand aus 24 beobachteten Variablen, von denen je 12 eine von zwei Dimensionen darstellten (Marsh et al., 2013). Der verwendete Datensatz mit 3,390 Schülerinnen und Schülern stammte aus einer früheren Publikation der ersten beiden Autoren (Marsh et al., 2010). Auch hier führte die Anwendung eines korrekten Zwei-Faktoren-

Modells anhand der individuellen Items auf den Datensatz zu einem schlechten Fit anhand des CFI und des RMSEA, ein Ein-Faktoren-Modell wurde klar verworfen (Marsh et al., 2013). Die Parcels aus jeweils 2, 3 oder 4 Items wurden auch in dieser Studie entweder aufgeteilt oder homogen gebildet und die Parcels stellten entweder eine Zwei-Faktoren-Lösung oder eine einfaktorielle Lösung dar. Die Modelle mit den homogenen Itemparcels, die eine Ein-Faktoren-Lösung widerspiegelten, wurden nach den Kriterien von Hu und Bentler (1998, 1999) durch die Fit-Indizes abgelehnt. Die Ein-Faktoren-Modelle mit den aufgeteilten Parcels, bei denen Neurotizismus und Extraversion konfundiert waren, resultierten jedoch in der Akzeptanz des Modells anhand der Cut-Off-Kriterien für den CFI und den RMSEA (Marsh et al., 2013).

Im Rahmen der dritten Studie wurden Daten simuliert (Populationssimulation mit je 100,000 Fällen pro Bedingung), die das Antwortverhalten auf einer Skala von 24 Items widerspiegeln sollten (Marsh et al., 2013). Variiert wurde in den Populationsmodellen der Grad der Einfachstruktur (Einfachstruktur [alle standardisierten Nebenladungen auf zweiten Faktor .00], annähernd Einfachstruktur [alle standardisierten Nebenladungen auf zweiten Faktor zwischen .00 und .10], gute Einfachstruktur [vier standardisierte Nebenladungen auf zweiten Faktor zwischen .00 und .20], akzeptable Einfachstruktur [sechs standardisierte Nebenladungen auf zweiten Faktor zwischen .00 und .40]), als auch die Faktorkorrelation (.25 oder .60). In den missspezifizierten Modellen wurden die Nebenladungen nicht spezifiziert und insofern wurde bei diesen fälschlicherweise von Einfachstruktur ausgegangen. Alle missspezifizierten Modelle ohne Parcels führten zur Annahme der Modelle durch die Fit-Indizes CFI und RMSEA. Die Item-Parcels wurden genauso gebildet wie in Studie 2. Einfaktorielle Modelle mit aufgeteilten Itemparcels führten genauso wie in Studie 2 anhand des CFI und des RMSEA zu akzeptabler Modellpassung. Die Ein-Faktoren-Modelle mit homogenen Parcels wurden hingegen durch die Fit-Indizes verworfen.

In Studie 4 wurde ein Populationsmodell mit drei korrelierten exogenen latenten Variablen und einer endogenen latenten Variablen untersucht, wobei nur die latente endogene Variable Indikatoren ohne Nebenladungen widerspiegelte (Marsh et al., 2013). Bei dieser Studie handelte es sich ebenfalls um eine Populationssimulation. Zudem wurde als weitere Populationsbedingung ein Methodenfaktor spezifiziert und die Korrelation zwischen erstem und zweitem exogenen Faktor sowie deren Strukturpfade auf dem endogenen Faktor in der Höhe variiert. Es wurden wie in den Studien 2 und 3 homogene und aufgeteilte Itemparcels gebildet. Auch im Rahmen dieser Studie führten insbesondere die aufgeteilten Parcels zu

akzeptablen Werten hinsichtlich der Modellpassung anhand der Cut-Off-Werte für den CFI und den RMSEA. Der Modellfit für die homogenen Parcels war nur geringfügig schlechter als der Modellfit für die aufgeteilten Parcels und führte bei der überwiegenden Mehrheit an Bedingungen zur Akzeptanz der Modelle.

Sowohl die Studien von Bandalos (2002) als auch die Studien von Marsh et al. (2013) zeigten, dass der Gebrauch von Item-Parcels bei unbekannter wahrer Faktorenstruktur im Rahmen konfirmatorischer Faktorenanalysen oder im Rahmen von Strukturgleichungsmodellen nicht zu empfehlen ist. Dies hat den Grund, dass im Rahmen der beschriebenen Studien die Fit-Indizes eine missspezifizierte Faktorenstruktur eines implizierten Modells bei geparcelten Indikatoren noch weniger erkannten als bei missspezifizierten Modellen, deren Indikatoren nicht geparcelt wurden. Insbesondere kaschierte der Gebrauch von aufgeteilten Parcels, also Parcels, deren Items hinsichtlich einer Gegebenheit variierten, eine Verletzung der Einfachstruktur in den Daten bzw. im wahren Modell. Sofern also homogene und aufgeteilte Itemparcels zu unterschiedlichen Resultaten führen, sei es hinsichtlich der Parameterschätzungen oder auch hinsichtlich der Modellpassung anhand der Fit-Indizes, ist Vorsicht geboten (Marsh et al., 2013, S. 276). Dies kann darauf hinweisen, dass die Einfachstruktur in den empirischen beobachteten Variablen verletzt ist. In diesem Fall kann dann auch die Varianz der entsprechenden latenten Variablen nicht mehr als Varianz eines einzigen Faktors interpretiert werden (Little et al., 2002a; Raykov, 2001).

2.2 Missspezifikationen in Form von Mehrdimensionalität im Messmodell

Im Folgenden werden zunächst Studien beschrieben, die die Fit-Indizes hinsichtlich Modellabweichungen in Form von Mehrdimensionalität im Messmodell untersuchten, bevor die Studien beschrieben werden, die sich auf nicht-spezifizierte Mehrdimensionalität im Strukturmodell fokussierten und aus denen die Fragestellung für die erste Studie abgeleitet wurde.

Mehrdimensionalität im Messmodell kann sich äußern durch Verletzung der Einfachstruktur (Indikatoren einer latenten Variablen haben Nebenladungen auf eine andere latente Variable; Beauducel & Wittmann, 2005). Eine Verletzung der Einfachstruktur bezeichnet also Mehrdimensionalität innerhalb der Indikatoren (Von Davier & Carstensen,

2007; Wei, 2008), die Dimensionalität im Strukturmodell bleibt erhalten (Wei, 2008).[13] Eine weitere Form der Mehrdimensionalität im Messmodell stellen Messfehlerkorrelationen der Indikatoren dar. Die Studien zur Sensitivität der Fit-Indizes bei Verletzung der Einfachstruktur und beim Vorliegen von Messfehlerkorrelationen werden im Folgenden näher beschrieben.

Beauducel und Wittmann (2005) untersuchten die Auswirkungen von Missspezifikationen in Form von Verletzungen der Einfachstruktur auf die gängigen Fit-Indizes. Deren Design bestand aus vier orthogonalen oder obliquen Faktoren mit insgesamt 20 Indikatoren, die unterschiedlich hoch, aber homogen auf die Faktoren luden und von denen vier Indikatoren positive oder negative Nebenladungen hatten. In den orthogonalen Faktorenmodellen konnte also keine Einfachstruktur durch oblique Rotation erreicht werden, in den obliquen Faktorenmodellen jedoch schon. Bei den missspezifizierten Modellen, die auf die Daten, die aus den beschriebenen Populationsmodellen erzeugt wurden, gefittet wurden, wurden die Nebenladungen nicht spezifiziert, was Missspezifikationen im Messmodell darstellte. Die Stichproben enthielten 250, 500 oder 1,000 Fälle. Die Ergebnisse waren ähnlich zu denen Heene et al.s (2011): Bei geringen Ladungen wurden die missspezifizierten Modelle durch RMSEA und SRMR zu oft akzeptiert. Einen wichtigen Befund von Beauducel und Wittmann stellte außerdem dar, dass die Höhe des Cut-Offs für den CFI mit der Ladungshöhe interagierte: Während bei einem Cut-Off von .90 bei steigender Ladungshöhe auch die Modellakzeptanz stieg, sank bei einem Cut-Off von .95 die Modellakzeptanz bei steigender Ladungshöhe. Die Beurteilung des Modellfits anhand des χ^2-Werts unter Berücksichtigung der Freiheitsgrade sowie anhand des RMSEA und SRMR, wie von Hu und Bentler (1998, 1999) vorgeschlagen, kann nach Beauducel und Wittmann (2005) bei geringen Ladungen nur empfohlen werden, wenn die Einfachstruktur leicht verletzt ist. Bei einer gröberen Verletzung der Einfachstruktur sei davon auszugehen, dass diese bei geringen Ladungen durch die Fit-Indizes nicht angezeigt wird und somit zu einer falschen Akzeptanz des missspezifizierten Modells führt. Dieses Resultat geht einher mit Heene et al.s (2011) Ergebnissen, nach denen sich die Cut-Offs für die Fit-Indizes nach Hu und Bentler (1998, 1999) über unterschiedliche Faktorladungen und eine Verletzung der Einfachstruktur hinweg nicht bewährten.

[13]Von Mehrdimensionalität zwischen den Indikatoren spricht man hingegen, wenn jede latente Variable nur Indikatoren abbildet, die nicht gleichzeitig auch eine andere latente Variable abbildet. Insofern ist im Falle der Mehrdimensionalität zwischen den Indikatoren Einfachstruktur gegeben (Stout et al., 1996; Wei, 2008).

Problematisch am Design von Beauducel und Wittmann (2005) war allerdings, dass mit ansteigender Faktorladungshöhe in den spezifizierten Populationsmodellen auch die Nebenladungen stiegen, welche im missspezifizierten Modell nicht spezifiziert wurden (Savalei, 2012). Mit höheren Ladungen stieg also auch der Grad der Missspezifikation. Nach den Ergebnissen von Heene et al. (2011) reagieren die absoluten Fit-Indizes aber genau konträr dazu. Bei hohen Ladungen im Vergleich zu niedrigen Ladungen wird eine Modellabweichung durch die absoluten Fit-Indizes eher als solche erkannt, durch die inkrementellen Fit-Indizes weniger. Bei niedrigen Ladungen bedarf es also einer schwerwiegenderen Missspezifikation, damit die absoluten Fit-Indizes diese indizieren (Savalei, 2012).

Eine andere häufige Form der Mehrdimensionalität im Messmodell stellen korrelierte Messfehler bzw. korrelierte unique Faktorwerte dar (Heene et al., 2012). Das Konzept der statistischen Eindimensionalität erfordert, dass, sobald die latente Variable spezifiziert wurde, die Korrelationen zwischen den Items Null werden sollten, da die Ursache der Korrelationen zwischen den Items die latente Variable ist (Bühner, 2011). Korrelierte Messfehler entstehen insofern durch Kovarianzen, die durch die latente Variable nicht erklärt werden können. Sie repräsentieren – im Gegensatz zu unsystematischen Messfehlern – systematische Messfehler (Brown & Moore, 2012). Sofern korrelierte Messfehler in den Daten vorhanden sind, stellt die Matrix $\boldsymbol{\Psi^2}$ aus Formel (2) unter II. 1 keine Diagonalmatrix mehr dar, die Off-Diagonal-Elemente sind dann nicht mehr signifikant verschieden von Null (Mulaik, 2009). Korrelierte Messfehler entstehen beispielsweise durch die gleichen Wörter innerhalb verschiedener Fragebogenitems, logische Abhängigkeiten zwischen den Items oder auch Mehrdimensionalität (Brown & Moore, 2012; Heene et al., 2012; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Korrelierte Fehler führen zu verzerrten Schätzungen der internen Konsistenz, der Reliabilitätsschätzung im Rahmen der Strukturgleichungsmodellierung und der Minderungskorrektur der Kriteriumsvalidität (Green & Hershberger, 2000; Gu, Little, & Kingston, 2013; Heene et al., 2012; Maxwell, 1968; Osburn, 2000; Zimmermann & Williams, 1977a). Außerdem kann für die Beurteilung der Validität dann nicht mehr der Korrelationskoeffizient betrachtet werden, da die Korrelation zwischen beobachteten Variablen durch die korrelierten Messfehlervarianzen verzerrt ist (Zimmermann & Williams, 1977b).

Trotz der hohen Relevanz korrelierter Messfehler wurde bisher kaum untersucht, ob Modellabweichungen in Form von korrelierten Fehlern von den gängigen Fit-Indizes anhand der Cut-Offs nach Hu und Bentler (1998, 1999) zuverlässig entdeckt werden. Heene et al.

(2012) konstruierten ein Design aus Populationsmodellen mit zwei korrelierten Faktoren mit je 12 Indikatoren, von denen die Messfehlervarianzen von drei oder sechs Indikatoren des einen Faktors entweder nur positiv oder positiv und negativ mit drei oder sechs Messfehlervarianzen der Indikatoren des anderen Faktors korreliert waren. Die Höhe der standardisierten Faktorladungen rangierte entweder zwischen .30 und .60 oder zwischen .50 und .80, die Höhe der Messfehlerkorrelationen bewegte sich zwischen .30 und .50 sowie -.30 und .30. Die Stichprobengrößen variierten zwischen 150 und 2,500 Fällen. Die Missspezifikation bestand in nicht-spezifizierten Messfehlerkorrelationen. Heene et al. konnten zeigen, dass SRMR und RMSEA selbst schwerwiegende Missspezifikationen in Form von nicht spezifizierten sechs positiv korrelierten faktorübergreifenden Messfehlervarianzen anhand der von Hu und Bentler (1998, 1999) vorgeschlagenen Cut-Off-Werte nicht als solche indizierten, und dies selbst bei einem Cut-Off von < .11 für das SRMR (Heene et al. 2012). Auch diese Befunde waren ähnlich zu den Befunden unter III. 1 abhängig von Stichprobengröße und Faktorladungshöhe. Der CFI stieg bei zunehmender Stichprobengröße an, wohingegen der SRMR und der RMSEA bei zunehmender Stichprobengröße geringer ausfielen (dieselben Befunde zeigten sich bei Fan et al. [2009] sowie Marsh et al. [2004]). Im Gegensatz zu RMSEA und SRMR erkannte der CFI die Missspezifikationen größtenteils, insbesondere bei niedrigen Faktorladungen. Letzterer Befund kann dadurch erklärt werden, dass ein Modell mit niedrigen Faktorladungen näher am Nullmodell ist (Heene et al., 2012, S. 43).

Savalei (2012) untersuchte den RMSEA auf seine Sensitivität hinsichtlich einer im fehlspezifizierten Modell nicht inkludierten Messfehlerkorrelation. Sie spezifizierte ein Populationsmodell mit einem Faktor, der von 8 Indikatoren widergespiegelt wurde. Die Autorin variierte die Ladungshöhen und die Höhe der Messfehlerkorrelation zwischen].00, 1.00[, wobei bei höheren Ladungen die Messfehlerkorrelation niedriger angesetzt wurde[14]. Bei homogenen Ladungen von .90 zeigte der RMSEA selbst bei einer nicht-spezifizierten Messfehlerkorrelation von .06 eine Modellabweichung an. Mathematisch-analytisch wurde die Modellablehnung eines Modells mit hohen Faktorladungen und kleinen Messfehlerkorrelationen durch die Fit-Indizes bereits von Browne, MacCallum, Kim, Andersen, und Glaser (2002) gezeigt. Bei homogenen Ladungen von .40 jedoch konnte die

[14]Die Stichprobengröße wurde nicht genannt.

Messfehlerkorrelation im Populationsmodell .26 erreichen, sodass der RMSEA bei .05[15] lag (Savalei, 2012). Dieser Effekt der Ladungen auf die Werte des RMSEA verstärkte sich, allerdings nur geringfügig, wenn die Ladungen heterogen waren und um .10 bis .15 von der durchschnittlichen Ladung abwichen. Der Effekt der Höhe der Ladungen auf den RMSEA reduzierte sich, wenn anstatt der 8 Indikatoren 16 Indikatoren verwendet wurden.

In einer zweiten Monte-Carlo-Studie untersuchte Savalei (2012) ein zweifaktorielles Populationsmodell mit zu .30 korrelierten Faktoren und 8 oder 16 Indikatoren insgesamt[16]. Weiters wurde eine Messfehlerkorrelation in das Populationsmodell einbezogen, die sich entweder zwischen zwei Indikatoren desselben Faktors bewegte oder faktorübergreifend spezifiziert wurde. Ladungshöhe und Höhe der Messfehlerkorrelation wurden wie in Savaleis erster Studie gewählt. Im Vergleich zum einfaktoriellen Modell aus der ersten Studie reduzierte sich die Sensitivität des RMSEA hinsichtlich der nicht-spezifizierten Messfehlerkorrelation im missspezifizierten Modell stark; dieser Befund war unabhängig von der Itemanzahl pro Faktor oder insgesamt. Bei homogenen Ladungen und Messfehlerkorrelationen zwischen Indikatoren desselben Faktors von .40 wurde die Missspezifikation bei einem Cut-Off von .05 nicht erkannt, selbst bei Messfehlerkorrelationen von fast 1.00 im Populationsmodell. Bei faktorübergreifenden Messfehlerkorrelationen jedoch zeigte der RMSEA eine deutlich höhere Sensitivität: Bei 8 Indikatoren insgesamt und homogenen Ladungen von .40 lag der RMSEA bei einer nicht-spezifizierten Messfehlerkorrelation von .20 bei .05, bei 16 Items insgesamt musste die Messfehlerkorrelation .40 erreichen, damit der RMSEA über .05 lag. Das Ergebnis, dass der RMSEA im Populationsmodell mit mehr Indikatoren und mehr Freiheitsgraden weniger sensitiv für nicht-spezifizierte Messfehlerkorrelationen war, steht im Kontrast zu dem bekannten Befund, nach dem der RMSEA die Sparsamkeit eines spezifizierten Modells belohnt (Beauducel & Wittmann, 2005; Breivik & Olsson, 2001; Savalei, 2012). Basierend auf den Ergebnissen der zweiten Studie ist allerdings einerseits zu vermuten, dass die Sensitivität des RMSEA primär von der tatsächlichen Faktorenstruktur im Populationsmodell, also vom Strukturmodell und weniger vom Messmodell, abhängt (Savalei). Andererseits konnte die nicht-spezifizierte Messfehlerkorrelation möglicherweise durch die Größe des Modells kompensiert werden (Savalei).

[15]Man beachte, dass Savalei (2012) nicht die Cut-Off-Regel von Hu und Bentler (1998, 1999) von < .06 verwendete, sondern eine ältere, die auf Browne & Cudeck (1993) zurückgeht.

[16]Die Stichprobengröße der Simulationsstudie wurde nicht genannt.

Während kleine Verletzungen der Einfachstruktur in Form von Nebenladungen bei Beauducel & Wittmann (2005) insbesondere bei hohen Ladungen zur Ablehnung der Modelle anhand des CFI führten, akzeptierten RMSEA und SRMR die missspezifizierten Modelle anhand der Cut-Offs nach Hu und Bentler (1998, 1999) oft. Bei Heene et al. (2012) sowie Savalei (2012) zeigte sich, dass der RMSEA insbesondere bei nicht-spezifizierten faktorübergreifenden Messfehlerkorrelationen vor dem Hintergrund von Ladungen realistischer Höhe nicht sensitiv für diese Art der Modellabweichung war, sofern diese Cut-Off-Werte als Kriterien zur Modellevaluation herangezogen wurden.

2.3 Missspezifikationen in Form von Mehrdimensionalität im Strukturmodell

Während zu den Auswirkungen von nicht spezifizierter Mehrdimensionalität im Messmodell auf die Fit-Indizes schon einige Simulationsstudien durchgeführt wurden, hinkt die Forschung zu Missspezifikationen im Strukturmodell hinterher. Doch gerade die Forschung von Heene et al. (2012) zu RMSEA und SRMR sowie von Savalei (2012) zum RMSEA lässt die Schlussfolgerung zu, dass die statistische Dimensionalität der Faktorenstruktur im Populationsmodell für die Sensitivität dieser Fit-Indizes eine entscheidende Rolle spielt. Es kann zudem argumentiert werden, dass Missspezifikationen im Strukturmodell schwerwiegender sind als im Messmodell (Mahler, 2011).

Kenny und McCoach (2009) untersuchten die Auswirkungen einer Missspezifikation in Form von Mehrdimensionalität im Strukturmodell auf die Fit-Indizes sowohl anhand einer Simulationsstudie, als auch berechneten sie die Werte der Fit-Indizes analytisch. Das Hauptaugenmerk der Studie lag allerdings darin, die Auswirkungen unterschiedlicher Indikatorenanzahlen auf die Fit-Indizes auszutesten. Die Autoren kreierten ein Populationsmodell mit zwei zu .80 korrelierten latenten Variablen und wandten ein einfaktorielles Modell auf die aus dem Populationsmodell erzeugten Stichproben an. Variiert wurde im Populationsmodell die Anzahl der Indikatoren: Jeweils 3, 6 oder 10 Indikatoren spiegelten je eine der beiden latenten Variablen wider, wobei die Faktorladungen bei .70 (standardisiert) lagen. Die Stichprobengröße betrug in allen erzeugten Stichproben $N = 200$. Die Ergebnisse der Simulationsstudie zeigten, dass der CFI mit steigender Anzahl an Indikatoren sank, also schlechtere Modellpassung anzeigte, wohingegen der RMSEA auch

sank, also besseren Modellfit anzeigte. Genauso verhielt es sich auch bei den analytisch berechneten Werten für die Fit-Indizes.

Mahler (2011) untersuchte den Effekt von nicht-spezifizierter statistischer Mehrdimensionalität hinsichtlich der Faktorenstruktur auf die Fit-Indizes. Dazu spezifizierte sie Populationsmodelle mit zwei oder drei latenten Variablen, die zwischen Null (orthogonales Modell) und Eins (obliques Modell bei einer Faktorkorrelation > 0) korrelierten. Die Anzahl der Indikatoren lag bei 12 und die Faktorladungen lagen entweder alle bei .40, .50., .60, .70, .80 oder .90[17]. Sie wandte ein einfaktorielles Modell auf die aus den Zwei- oder Drei-Faktorenmodellen erzeugten Stichprobendaten an. Sowohl CFI, als auch RMSEA und SRMR zeigten eine höhere Modellpassung, wenn die Faktorkorrelation stieg und das Populationsmodell insofern näher am eindimensionalen Modell lag. Der CFI zeigte besseren Modellfit, wenn ein einfaktorielles Modell auf zweifaktorielle (im Vergleich zu dreifaktoriellen) Stichprobendaten angewandt wurden, der SRMR zeigte bessere Modellpassung bei Anwendung eines einfaktoriellen Modells auf eine Drei-Faktoren-Struktur (im Vergleich zu einer zweifaktoriellen) in den Daten. Der RMSEA zeigte keinen Unterschied hinsichtlich der Modellpassung, wenn ein eindimensionales Modell auf zwei- oder dreidimensionale Stichprobendaten angewandt wurde. Die Ergebnisse zeigten außerdem wiederum, dass die Ladungshöhe einen entscheidenden Einfluss auf die Indizierung der Missspezifikation durch die Fit-Indizes anhand der Cut-Offs nach Hu und Bentler (1998, 1999) hatte. Bei einer Ladungshöhe von .40 lag der SRMR, egal, wie hoch die Faktorkorrelation war, nicht über dem Cut-Off von Hu und Bentler (1998, 1999) von .08 (Mahler, 2011). Bei derselben Ladungshöhe zeigte der CFI bis zu einer Faktorkorrelation von .60 eine Modellabweichung nach Hu und Bentlers Daumenregeln an, darüber, unabhängig von der Anzahl der Faktoren im Populationsmodell, nicht mehr. Der RMSEA zeigte bei einer Ladungshöhe von .40 erst Misfit im Sinne von Hu und Bentlers Daumenregeln an, wenn die Faktorkorrelation kleiner als .30 war.

Mahlers (2011) Ergebnisse zeigten, dass die Fit-Indizes, vor allem die absoluten Fit-Indizes, bei geringen Faktorladungen, die dennoch überdurchschnittlich hoch ausfielen, große Schwierigkeiten bei der Entdeckung der Missspezifikation im Strukturmodell in Form von nicht-spezifizierten Faktoren hatten, sofern man die Cut-Offs nach Hu und Bentler (1998, 1999)

[17]Die Stichprobengröße für die Simulationsstudie wurde nicht genannt.

heranzog. Dabei interagierte außerdem die Ladungshöhe mit der Faktorkorrelation im Populationsmodell (Mahler, 2011).

Savalei (2012) führte zwei weitere Studien durch, die die Auswirkungen von Missspezifikationen im Strukturmodell in Form von nicht spezifizierter Mehrdimensionalität auf den RMSEA untersuchen sollten. Im Vergleich zur Studie von Mahler (2011) schränkte sie einerseits die Anzahl der betrachteten Fit-Indizes auf den RMSEA ein, erweiterte aber andererseits das Design von Mahler. Die Autorin spezifizierte Populationsmodelle mit obliquen Faktoren, deren Korrelation aus dem Bereich [.00, 1.00][18] kam und die von insgesamt 8 Indikatoren repräsentiert wurden[19]. Die homogenen Ladungen in den Populationsmodellen variierten zwischen .40, .50, .60., .70 oder .80. Sie wandte dann einerseits ein orthogonales Modell auf die Daten an (bei der Faktorkorrelation von .00 ein korrektes Modell), die aus den Populationsmodellen gezogen wurden, andererseits ein eindimensionales Modell (bei der Faktorkorrelation von 1.00 ein korrektes Modell). Wurde ein orthogonales Modell auf die eigentlich oblique Faktorenstruktur in den Daten angewandt, stieg der RMSEA mit steigender Faktorkorrelation im wahren Modell, zeigte also zunehmend Modellmisfit an. Sofern ein eindimensionales Modell auf die erzeugten Stichprobendaten gefittet wurde, stieg der RMSEA mit sinkender Faktorkorrelation im Populationsmodell. Waren die Faktorladungen im Populationsmodell gering (.40), lag der RMSEA bei einer Faktorkorrelation von .50 im Populationsmodell gerade noch unter dem Cut-Off nach Hu und Bentler (1998, 1999), zeigte also Modellpassung an, unabhängig davon, ob ein orthogonales Modell oder ein eindimensionales Modell angewandt wurde (Savalei, 2012). Bei höheren Faktorladungen wurden die missspezifizierten Modelle bereits bei niedrigerer (Anwendung eines annähernd orthogonalen Modells) bzw. bei höherer Faktorkorrelation (Anwendung eines annähernd eindimensionalen Modells) durch den RMSEA-Cut-Off abgelehnt. Bei in der Psychologie realistischen Faktorladungen, die sogar noch unter den von Savalei festgelegten Ladungen liegen (vgl. Peterson, 2000), wäre nach Savaleis Ergebnissen eine schwere Missspezifikation im Strukturmodell notwendig, damit der RMSEA eine Modellabweichung anzeigen würde. Über alle Faktorladungshöhen hinweg betrachtet bestrafte der RMSEA ein missspezifiziertes eindimensionales Modell eher mit einem schlechten Modellfit als ein missspezifiziertes orthogonales Modell: Einen Faktor nicht zu spezifizieren, stellte also in der Metrik des RMSEA

[18]Das Populationsmodell mit der Faktorkorrelation von .00 spiegelte insofern ein orthogonales Modell wider.

[19]Die Stichprobengröße ist hier ebenfalls nicht bekannt.

eine schwerwiegendere Missspezifikation dar, als zwei Faktoren fälschlicherweise als unabhängig zu betrachten (Savalei).

In einem zweiten Schritt untersuchte Savalei (2012), ob die Anzahl der latenten Dimensionen im Populationsmodell eine Auswirkung auf den RMSEA hätten, wenn ein eindimensionales Modell auf die aus dem Populationsmodell erzeugten Stichproben-Kovarianzmatrizen angewandt würden. Sie spezifizierte Populationsmodelle mit zwei bis acht latenten Variablen, die sie entweder zu .30 oder zu .50 korrelieren ließ. Die Höhe der Faktorladungen setzte sie in gleicher Höhe wie in den vorherigen Studien fest. Die Anzahl der Indikatoren lag unabhängig von der Anzahl der Faktoren immer bei 24[20]. Die Befunde sind alarmierend: Wenn die Anzahl der Faktoren im Populationsmodell stieg, sanken die Werte des RMSEA, wenn das eindimensionale Modell angewandt wurde (Savalei). Bei niedrigen Ladungen von .40 lag der RMSEA immer unter dem Cut-Off-Wert nach Hu und Bentler (1998, 1999). Bei Faktorladungen von .50 lag er über .06, wenn die Faktorkorrelation im Populationsmodell bei .50 lag; wenn die Faktorkorrelation bei .30 lag, resultierte der RMSEA bei Ladungen von .50 in Werten um .06, wenn zwei bis vier Faktoren im Populationsmodell vorlagen (Savalei, 2012).

Die Studien von Savalei (2012) zeigten im Einklang mit den Ergebnissen von Mahler (2011), dass der RMSEA bei niedrigen Faktorladungen, die dennoch für die angewandte Psychologie relativ hoch sind (vgl. Peterson, 2000), bei Anwendung eines eindimensionalen Modells kaum Modellabweichung anhand der Cut-Off-Werte anzeigte, und dies unabhängig davon, ob die tatsächliche Faktorenstruktur zwei oder sogar acht Dimensionen beinhaltete (Savalei, 2012).

[20]Die Stichprobengröße der Simulationsstudie wurde nicht genannt.

3 Erste Fragestellung: Auswirkungen nicht-spezifizierter Mehrdimensionalität im Strukturmodell auf die Fit-Indizes

Die beschriebenen Studien lassen den Schluss zu, dass insbesondere die Faktorladungen, also inzidentelle Parameter eines Modells, großen Einfluss auf die Sensitivität der Fit-Indizes hinsichtlich der Indizierung von Missspezifikationen haben. Dieser Einfluss der Faktorladungen führt von einer Modellablehnung bei sehr geringen Modellabweichungen bei hohen Ladungen bis hin zu einer Verschleierung nicht vorhandener Modellpassung bei geringen Ladungen (siehe III. 1 und 2). Der Schluss, dass die Faktorladungen einen großen Einfluss auf die Fit-Indizes haben, kann sowohl für Missspezifikationen im Messmodell als auch für Missspezifikationen im Strukturmodell gezogen werden (Heene et al., 2011; Savalei, 2012).

Savalei (2012), Mahler (2011) und insbesondere Kenny und McCoach (2009) untersuchten Missspezifikationen in Form von Mehrdimensionalität im Strukturmodell bei unrealistisch hohen Faktorladungen. Wie bereits unter III. 1, basierend auf Petersons (2000) Metaanalyse, erwähnt, fallen Faktorladungen in der angewandten Psychologie und verwandten Disziplinen deutlich geringer aus. Das Forschungsdesign von Kenny und McCoach war des Weiteren vor allem auf die Austestung des Einflusses der Indikatorenanzahl angelegt und die Autoren verwendeten nur homogene Faktorladungen. Auch letzteres ist in der angewandten Forschung so gut wie nie der Fall (Buzick, 2010). Bei Savalei kommt neben den ebenso als homogen festgelegten Faktorladungen hinzu, dass sie nur den Einfluss von Missspezifikationen in Form von Mehrdimensionalität auf den RMSEA überprüfte. Mahler untersuchte zwar auch CFI und RMSEA, allerdings auch nur bei homogenen und relativ hohen Faktorladungen.

Im Rahmen der ersten Studie wurde ein Forschungsdesign konzipiert, welches die oben genannten Forschungslücken im Fokus hatte: Es wurde die Sensitivität der gängigen Fit-Indizes CFI, RMSEA und SRMR hinsichtlich nicht-spezifizierter Mehrdimensionalität in Form von zwei obliquen Faktoren im Populationsmodell (eindimensionales missspezifiziertes Modell) untersucht. Die Faktorladungen wurden sowohl heterogen, als auch für die angewandte Psychologie realistisch hoch definiert. Der Grad der Missspezifikation wurde einerseits durch die Höhe der Faktorkorrelation im Populationsmodell, andererseits durch das Modell(un)gleichgewicht hinsichtlich der Anzahl der Indikatoren pro Faktor variiert; ein niedrigerer Grad an Missspezifikation bestand in einer höheren Faktorkorrelation im

Populationsmodell (im Vergleich zu einer niedrigeren Faktorkorrelation) und in einer ungleichen Indikatorenaufteilung im Populationsmodell (im Vergleich zu einer ausgewogenen). Die Auswirkungen einer ungleichen Indikatorenaufteilung auf zwei Faktoren auf die Fit-Indizes wurden bisher nur im Kontext von Missspezifikationen im Messmodell von Mahler (2011) untersucht. Dies allerdings nur in Form eines inzidentellen Parameters, bei Mahler wurde die Missspezifikation nicht durch die Indikatorenaufteilung bestimmt. Unter IV wird die Studie genauer beschrieben.

Nachdem der Forschungsbedarf hinsichtlich der Sensitivität der gängigen Fit-Indizes gegenüber Missspezifikationen in Form von Mehrdimensionalität im Strukturmodell aufgezeigt wurde, stellt sich im folgenden Kapitel III. 4 als nächstes die Frage, inwiefern sich Missspezifikationen dieser Form substanziell und auf die Individuen auswirken würden. Die Beurteilung der Güte der Modellpassung anhand der Fit-Indizes kann als eine Frage der Reliabilität betrachtet werden, letztere Fragestellung bezieht sich auf die Validität der Faktorwerte (für den Begriff vgl. Grice [2001a, 2001b]) und darauf aufbauend auf die Validität der Diagnosen aus den Faktorwerten im Rahmen der zweiten Studie. Wünschenswert wäre, dass die Fit-Indizes eine Modellabweichung anzeigen, bevor die Missspezifikation hinsichtlich der Validität der Faktorwerte der Individuen kritisch wird.

4 Diagnostische Entscheidungen

Wie bereits beschrieben, werden mehr und mehr psychologische Tests und Fragebögen anhand von konfirmatorischen Faktorenanalysen und linearen Strukturgleichungsmodellen konstruktvalidiert ("Datenbanksegment PSYNDEX Tests," 2013). Ein Grund dafür liegt sicherlich in der Möglichkeit der Modellierung der latenten Variablen (Fan et al., 2009; Tomarken & Waller, 2005) aufgrund der Kontrolle der Messfehlerkovarianz der manifesten Variablen (Bollen, 1989; Oberski & Satorra, 2013). Die Möglichkeit, latente Variablen zu modellieren, hat insbesondere für die Psychometrie große Vorteile (Fan et al., 2009; Tomarken & Waller, 2005), da die Test- und Fragebogenkonstruktion auf eine valide Erfassung der Ausprägungen der Individuen auf den latenten Variablen mittels der Faktorwerte abzielt.

Bei der Anwendung von Testverfahren im Rahmen psychologischer Einzelfalldiagnostik werden sehr oft diagnostische Entscheidungen hinsichtlich quantitativer Klassifikationen[21] getroffen. Dieser werden entweder auf Basis der Faktorwerte[22] oder, deutlich häufiger, auf Basis von Summenwerten[23] (siehe III. 5.2) getroffen. Derartige diagnostische Entscheidungen betreffen beispielsweise klinische Diagnosestellungen („krank" versus „gesund") oder auch die Personalauswahl („geeignet" versus „ungeeignet"). Diese diagnostischen Entscheidungen werden entweder normorientiert oder kriterienorientiert getroffen (Amelang & Schmidt-Atzert, 2006, S. 16). Eine normorientierte Diagnostik bestimmt (inter-)individuelle Unterschiede hinsichtlich eines Merkmals, wohingegen eine kriterienorientierte Diagnostik die individuelle Position in Relation zu einem Merkmalskriterium angibt (S. 16). Eine normorientierte Diagnostik stellt beispielsweise eine Top-Down-Klassifikation dar, nach der ein festgelegter Prozentsatz an Personen ausgewählt wird, die die höchsten Werte auf einer Skala/in einem Testverfahren erreichen[24], oder auch eine Bottom-Up-Klassifikation, nach der ein festgelegter Prozentsatz an Personen ausgewählt wird, die die niedrigsten Werte auf einer Skala/in einem Testverfahren erreichen (Gatewood, Feild,

[21]Es sei darauf hingewiesen, dass weder an dieser Stelle, noch an folgenden Textstellen eine qualitative Wertung mit dem Begriff „Klassifikation" einhergeht.

[22]So kann zum Beispiel beim I-S-T 2000 R (Liepmann, Beauducel, Brocke, & Amthauer, 2007) ein Faktorwert für die Individuen berechnet werden, wobei keine weiteren quantitativen Klassifikationen getroffen werden.

[23]beispielsweise beim BDI-II (Beck et al., 2006)

[24]Beim I-S-T 2000 R (Liepmann et al., 2007) wird beispielsweise der Prozentrang des individuellen IQ-Wertes in Relation zur Normstichprobe bestimmt.

& Barrick, 2016, S. 662). Eine kriterienorientierte diagnostische Entscheidung kann beispielsweise auf einem Cut-Off-Wert basieren, also einer Mindestanforderung, die die Personen überschreiten müssen, um ausgewählt zu werden[25] (Gatewood et al., S. 663).

Sofern die Diagnostik auf Basis der Faktorwerte (im Gegensatz zu den Summenwerten; siehe III. 5.2) erfolgt, resultiert die Frage nach der Validität einer diagnostischen Entscheidung aus der Validität der Faktorwerte aus einem Modell, das dem Testverfahren zugrunde liegt. Im Gegensatz dazu bezieht sich die Frage nach der Modellpassung, auf die sich Studie 1 fokussierte, auf die Reliabilität eines Modells. Die Frage nach den psychometrischen Auswirkungen einer Modellabweichung auf die Validität der Faktorwerte bzw. die Validität diagnostischer Entscheidungen aus den Faktorwerten heraus wurde erstmals im Rahmen der zweiten Studie dieser Dissertation gestellt. Bevor jedoch die zweite Fragestellung näher beschrieben wird (siehe III. 5.1), wird die bisherige Forschung zur Güte diagnostischer Entscheidungen beschrieben. Diese zeigt, dass die diagnostische Präzision nicht nur vom theoretischen Modell und dessen (in-)korrekter Spezifikation (Konstruktvalidität) abhängt, auf dessen Basis die diagnostische Entscheidung erfolgt (Emons, Sijtsma, & Meijer, 2007; Kruyen, Emons, & Sijtsma, 2012; Schönemann, 1997; Schönemann & Thompson, 1996; Taylor & Russell, 1939). Weitere Einflussfaktoren stellen neben der Validität eines Testverfahrens auch die Basisrate (Anteil an Personen in der Population, die zu einem bestimmten Zeitpunkt ein bestimmtes Merkmal aufweisen; Eid, Gollwitzer, & Schmitt, 2013, S. 163), die Selektionsrate (Anteil der ausgewählten Individuen; S. 163) sowie Reliabilität und Trennschärfe des Testverfahrens dar (Emons et al., 2007; Kruyen et al., 2012; Meehl & Rosen, 1955; Schönemann, 1997; Schönemann & Thompson, 1996; Taylor & Russell, 1939).

Taylor und Russell (1939) zeigten für den Fall von bivariat normalverteilten Variablen, dass sich die Güte der Vorhersage einer Variable bei steigender Validität eines Testverfahrens (operationalisiert durch den Pearson-Korrelationskoeffizienten) exponentiell verbesserte. Die Autoren entwickelten Tabellen für binäre Entscheidungen, an denen abzulesen ist, inwiefern die Rate an erfolgreich ausgewählten Kandidatinnen und Kandidaten bei vorgegebener Basisrate mit der Validität und der Selektionsrate variiert. Außerdem machen diese Tabellen

[25]Beim BDI-II (Beck et al., 2006) wird beispielsweise ab einem Cut-Off von 29 basierend auf den Gesamtsummenwerten eine major depressive Episode diagnostiziert.

deutlich, dass die Rate an korrekt als „geeignet“ Eingeordneten bei gleich großen Basis- und Selektionsraten sowie einer Validität von Null der Basis- bzw. Selektionsrate, also einer Zufallsentscheidung, entspricht; bei nicht gleich großen Basis- wie Selektionsraten der Selektionsrate.

Meehl und Rosen (1955), Schönemann und Thompson (1996) sowie Schönemann (1997) argumentierten genauso wie Taylor und Russell (1939), dass die Güte einer diagnostischen Entscheidung stark von der Basisrate, der Validität des Tests und der Selektionsrate abhängt. Schönemann und Thompson (1996) betonten für den spezifischen Fall von dichotomen Variablen und sowohl unter Kontrolle des Alpha- als auch des Beta-Fehlers, was bereits von Meehl und Rosen geschrieben wurde: Testverfahren würden unabhängig von deren Validität nur für Basisratensplits von 50%/50% bei dichotomen Entscheidungen hinsichtlich der Hit Rate[26] deutlich häufiger korrekt als der Zufall klassifizieren (S. 14). Für Tests mit geringen Validitäten würden die Basisraten entscheidender für die Korrektheit der diagnostischen Entscheidung. Für Basisraten von 30% (versus 70% in der Gruppe der Nicht-Merkmalsträgerinnen und Nicht-Merkmalsträger) und einer geringen, aber laut den Autoren realistischen Validität von kleiner als .50 würde sich ein Test zur Klassifikation nicht besser als der Zufall eignen (S. 14). Schönemann (1997) sowie Meehl und Rosen (1955) warnten zudem, dass für Populationen mit extrem kleinen Basisraten, wie sie in der klinischen Psychologie vorkommen, oder extrem großen Basisraten das Einsetzen eines Testverfahrens hinsichtlich der korrekten diagnostischen Einordnung sogar schlechter ausfallen könne als der Zufall, wenn die Validität des Testverfahrens gering ist.

Meehl und Rosen kritisierten bereits 1955, dass die Basisraten nicht ausreichend berichtet würden und es insofern schwierig wäre, psychometrische Entscheidungen überhaupt zu evaluieren, nach Schönemann (1997) hätte sich in der Zwischenzeit daran nichts geändert. Im Folgenden werden zwei Simulationsstudien berichtet, die die Güte diagnostischer Entscheidungen auf der Basis probabilistischer Modelle[27] u.a. mit verschiedenen Basisraten

[26]Dieser von Schönemann und Thompson (1996, S. 8) sowie Schönemann (1997, S. 175) als „Hit Rate“ bezeichnete Kennwert ist mittlerweile unter dem Begriff der „Sensitivität“ bekannt und bezeichnet die Rate der korrekt als Merkmalsträgerinnen und Merkmalsträger erkannten Fälle an den Fällen aller Merkmalsträgerinnen und Merkmalsträger (Amelang & Schmidt-Atzert, 2006, S. 422; Eid, Gollwitzer, & Schmitt, 2013, S. 163; Ziegler & Bühner, 2012, S. 146).

[27]Im Gegensatz zur Strukturgleichungsmodellierung, die auf der Klassischen Testtheorie basiert

untersuchten, welche auch im Rahmen der zweiten Studie variiert wurden. Allerdings lag der Hauptfokus dieser Studien – im Gegensatz zur vorliegenden Arbeit, bei der die Auswirkungen von Missspezifikationen untersucht wurden – auf der Austestung von Kurzskalen hinsichtlich diagnostischer Entscheidungen unter der Voraussetzung der Modellgültigkeit.

Emons et al. (2007) untersuchten die Konsistenz von disjunkten diagnostischen Entscheidungen („braucht Behandlung“ versus „braucht keine Behandlung“) im Rahmen einer Simulationsstudie. Die Skalen, anhand derer die diagnostischen Entscheidungen getroffen wurden, waren entweder Kurz- oder Lang-Skalen und bestanden entweder aus dichotomen oder polytomen Items. Als Analysemethoden wurden Rasch-Modelle ausgewählt. Extreme Cut-Off-Werte, d.h. kleine Basisraten für die definierte Behandlungsgruppe, führten zu höheren Raten an diagnostischen Konsistenzen für die Gruppe, die keine Behandlung brauchte (Korrekt Negative) und zu kleineren Raten an diagnostischen Konsistenzen für die Gruppe, die Behandlung brauchte (Korrekt Positive). Insbesondere bei Basisraten von 10% und 5%, wie sie auch in der klinischen Psychologie vorkommen, in Interaktion mit geringen Trennschärfen und einer für die Psychologie typischen Itemanzahl von 20 dichotomen Items, wurden nur 47% (Basisrate 10%) bzw. 42% (Basisrate 5%) der Fälle, die in Wahrheit in der Behandlungsgruppe waren, mit einer 90%igen Sicherheit korrekt in diese Behandlungsgruppe eingeordnet; dagegen wurden 90% (Basisrate 10%) bzw. 94% (Basisrate 5%) der Individuen, die in Wahrheit in der Gruppe ohne Behandlungsbedarf waren, bei einem Sicherheitslevel von 90% korrekt klassifiziert (S. 113). Für im Rahmen der Simulation definierte hohe Trennschärfen der Items oder auch polytome Items verbesserte sich die Klassifikation für die Behandlungsgruppe, für die nicht behandlungsbedürftige Gruppe nur unwesentlich. Für die Kurzskalen fielen die diagnostischen Konsistenzen deutlich niedriger aus.

In einer weiteren Simulationsstudie untersuchten Kruyen et al. (2012) die Güte verschiedener Arten von dichotomen diagnostischen Entscheidungen (Top-Down-Klassifikationen wie auch Cut-Score-basierte Klassifikationen). Sie verwendeten unterschiedliche Basis- und Selektionsraten, dichotome und polytome manifeste Variablen sowie unterschiedliche Testlängen. Als Datenanalysemethoden wurden wiederum Raschmodelle und Graded-Response-Modelle verwendet. Top-Down-Klassifikationen nach den höchsten Werten auf jeder der fünf gleich langen Einzelskalen, welche zu .20 korrelierten, führten bei einer typischen Testlänge von 20 Items bei einer Basisrate von 50% und der

kleinsten Selektionsrate von 10% zu einer Sensitivität[28] von .65 und einer Spezifität von .96 (Kruyen et al., 2012, S. 332). Die Ergebnisse fielen wie bei Emons et al. (2007) klar zugunsten der Langskalen mit höheren Reliabilitäten aus, wobei die Testlänge mit der Basisrate und der Selektionsrate interagierte. Die Befunde für die Cut-Score-basierten Klassifikationsentscheidungen fielen sehr ähnlich im Vergleich zu den Top-Down-Zuordnungen aus. Die diagnostischen Entscheidungen fielen basierend auf polytomen Items ähnlich aus wie auf Basis von dichotomen Items. Auch dieser Befund zeigte sich bereits bei Emons et al. (2007).

Die Autoren schlussfolgerten aus ihren Simulationsstudien, dass – bei realistischen Trennschärfen und guten Reliabilitäten – Langskalen erforderlich wären, um ausreichend hohe Trefferquoten zu erreichen, und dies insbesondere für die Gruppe der Merkmalsträgerinnen und Merkmalsträger im Vergleich zur Gruppe der Nicht-Merkmalsträgerinnen und Nicht-Merkmalsträger (Emons et al., 2007; Kruyen et al., 2012) Diese Befunde stellen die neueren Entwicklungen in der Testkonstruktion in Richtung der Kurzskalen (z.B. das BDI-FS; Beck, Steer, & Brown, 2013) stark in Frage.

Die Ergebnisse von Emons et al. (2007) sowie Kruyen et al. (2012) können u.a. aufgrund anderer Verteilungseigenschaften der untersuchten Variablen und aufgrund eines sehr unterschiedlichen Designs nicht mit den Taylor-Russell-Tafeln (Taylor & Russell, 1939) verglichen werden. Ein Vergleich mit den Berechnungen von Schönemann und Thompson (1996) ist nicht möglich, da diese Autoren einerseits nur Berechnungen für Validitäten bzw. Korrelationskoeffizienten bis .50 anstellten und sich auch zwischen diesen beiden Studien das Design stark unterschied. Jedoch zeigen die Befunde aus den beiden genannten Simulationsstudien, was bereits von Schönemann und Thompson (1996) sowie Taylor und Russell (1939) berechnet wurde: Die Basis- und Selektionsraten haben einen entscheidenden Einfluss auf die Güte von diagnostischen Entscheidungen (Emons et al., 2007; Kruyen et al.,

[28]Die Begriffe Sensitivität und Spezifität wurden erstmals im Rahmen der Signalentdeckungstheorie beschrieben (Green & Swets, 1966) und später in die psychologische Diagnostik übertragen. Die Sensitivität bezeichnet die Rate an korrekt klassifizierten Merkmalsträgerinnen und Merkmalsträger relativiert an allen Merkmalsträgerinnen und Merkmalsträgern und die Spezifität bezeichnet die Rate an korrekt klassifizierten Nicht-Merkmalsträgerinnen und Nicht-Merkmalsträger relativiert an allen Nicht-Merkmalsträgerinnen und Nicht-Merkmalsträgern (siehe z.B. Amelang & Schmidt-Atzert, 2006, S. 422; Eid, Gollwitzer, & Schmitt, 2013, S. 163; Ziegler & Bühner, 2012, S. 146).

2012). Außerdem zeigte sich, dass neben der Validität, wie von den genannten Autoren sowie von Meehl und Rosen (1955) als auch Schönemann (1981) beschrieben, auch die unterschiedlichen Reliabilitäten der Skalen sowie die Höhe der Trennschärfe einen wichtigen Faktor hinsichtlich der Sensitivität darstellte (Emons et al., 2007; Kruyen et al., 2012).

5 Zweite Fragestellung: Auswirkungen nicht-spezifizierter Mehrdimensionalität im Strukturmodell auf diagnostische Entscheidungen

5.1 Diagnostische Entscheidungen basierend auf missspezifizierten Modellen

Es wurde bereits berichtet, dass die Validierung von Testverfahren durch konfirmatorische Faktorenanalysen und lineare Strukturgleichungsmodelle zunimmt ("Datenbanksegment PSYNDEX Tests," 2013). Diese bringt psychometrische Vorteile mit sich (Fan et al., 2009; Tomarken & Waller, 2005), welche sich wiederum positiv auf die Güte diagnostischer Entscheidungen auswirken können. Dass bisher nur wenig Forschung zur Güte diagnostischer Entscheidungen existiert, wurde unter III. 4 aufgezeigt. Diese Studien untersuchten unter anderem den Einfluss der Validität eines diagnostischen Instruments auf die Güte diagnostischer Entscheidungen (Schönemann & Thompson, 1996; Schönemann, 1997; Taylor & Russell, 1939). Unter III. 1 wurde beschrieben, dass Missspezifikationen mehr die Regel als die Ausnahme bei der Strukturgleichungsmodellierung darstellen (Marsh et al., 2004). Die im Rahmen der Dissertation untersuchte Art der Missspezifikation (nicht-spezifizierte Zweidimensionalität im Strukturmodell) stellt eine Missspezifikation in Form einer Verletzung der Konstruktvalidität dar und ist somit von inzidentellen Parametern des Modells abzugrenzen, die Einfluss auch auf die Güte diagnostischer Entscheidungen basierend aus einem Modell heraus nehmen. Daher stellt sich als nächstes die Frage, was es in substanzieller Hinsicht für die getesteten Individuen bedeuten würde, wenn ein Modell in Form der beschriebenen Verletzung der Konstruktvalidität missspezifiziert ist. Diese Fragestellung betrifft im Kontext der Strukturgleichungsmodellierung die Validität der Faktorwerteschätzung bzw. darauf aufbauend die Validität der diagnostischen Entscheidungen basierend auf den Faktorwerten und ist insbesondere für die Testkonstruktion und die Einzelfalldiagnostik von Interesse. Im Rahmen der zweiten Studie wurde daher untersucht, inwieweit die Güte der Diagnostik beeinträchtigt werden würde, wenn Diagnosen auf Basis der Faktorwerte fälschlich als einfaktoriell spezifizierter Modelle vergeben wurden, die Faktorenstruktur der True Scores (der wahren Faktorwerte der Individuen; Eid et al., 2013, S. 818) jedoch zweifaktoriell ist. Aus Referenzgründen wurde außerdem untersucht, inwieweit die Diagnosen auf Basis der Faktorwerte korrekter zweifaktorieller Modelle den wahren Diagnosen basierend auf den True

Scores entsprechen würden. Der Grad der Missspezifikation wurde durch die Höhe der Faktorkorrelation sowie die (Un-)Ausgewogenheit der Indikatorenaufteilung auf die Faktoren im Populationsmodell variiert. Neben des Einflusses der Validität wurde die Relevanz von Parametern wie der Basisrate und der Reliabilität eines Testverfahrens/Fragebogens hinsichtlich der Güte diagnostischer Entscheidungen unter III. 4 aufgezeigt. Daher wurden die Diagnosen im Rahmen der zweiten Studie basierend auf unterschiedlichen Basisraten vergeben sowie zudem als inzidenteller Modellparameter die Reliabilität (Höhe der Faktorladungen) in einem realistischen Ausmaß (vgl. Peterson, 2000) variiert. Da die klinische Psychologie einen großen Teil angewandter Forschung innerhalb der Psychologie ausmacht und tagtäglich in Deutschland im Rahmen der Einzelfalldiagnostik allen voran klinische Tests und Fragebögen im Einsatz sind ("Datenbanksegment PSYNDEX Tests," 2013), wurden für die Diagnosenvergabe Basisraten in klinischen Größenordnungen herangezogen. Die Relevanz klinischer Diagnostik wird ferner an Wittchen et al.s (2011, S. 656) EU-Studie deutlich, nach der pro Jahr etwa 38% aller EU-Bürger an mindestens einer psychischen Störung leiden. Zudem wurden aus Vergleichsgründen auch größere Basisraten, wie sie in der Eignungsdiagnostik vorkommen (vgl. Schuler, 2014), für die Vergabe der Diagnosen verwendet.

Für die Evaluation der Güte psychologischer Diagnostik wurden Sensitivität, Spezifität, Positiver und Negativer Prädiktionswert berechnet (Amelang & Schmidt-Atzert, 2006, S. 422; Eid, Gollwitzer, & Schmitt, 2013, S. 163; Ziegler & Bühner, 2012, S. 146), wie sie in Tabelle 2 erklärt werden. Diese diagnostischen Kennwerte werden zumeist zur Beurteilung der diagnostischen Präzision verwendet, so auch im Rahmen der Studie 2 (siehe V). Sensitivität und Spezifität sowie Positiver und Negativer Prädiktionswert verhalten sich komplementär: Steigt die Sensitivität oder steigt der Positive Prädiktionswert, sinkt die Spezifität und sinkt der Negative Prädiktionswert und umgekehrt. Diese diagnostischen Kennwerte lassen sich aus den diagnostischen Konsistenzen berechnen (Korrekt Positive, Korrekt Negative, Falsch Positive sowie Falsch Negative; Ziegler & Bühner, 2012, S. 147), welche in Tabelle 1 aufgeführt sind.

Tabelle 1

Konsistenzen diagnostischer Entscheidungen

		Tatsächlicher Zustand	
		Krank	Gesund
Testergebnis	Krank	Korrekt Positiv/True Positive (TP)	Falsch Positiv/False Positive (FP)
	Gesund	Falsch Negativ/False Negative (FN)	Korrekt Negativ/True Negative (TN)

Anmerkungen. Aus Gründen des allgemeinen Sprachgebrauchs in der Literatur werden die Begriffe auch auf Englisch eingeführt (Ziegler & Bühner, 2012, S. 147).

Tabelle 2

Diagnostische Kennwerte

Positiver Prädiktionswert = TP / (TP + FP)
Negativer Prädiktionswert = TN / (TN + FN)
Sensitivität = TP / (TP + FN)
Spezifität = TN / (TN + FP)

Anmerkungen. Die Diagnostischen Kennwerte (Amelang & Schmidt-Atzert, 2006, S. 422; Eid, Gollwitzer, & Schmitt, 2013, S. 163; Ziegler & Bühner, 2012, S. 146) lassen sich aus den Konsistenzen diagnostischer Entscheidungen (siehe Tabelle 1) berechnen.

5.2 Diagnostische Entscheidungen basierend auf Gesamtsummenwerten

Eine Nebenfragestellung für die genannte zweite Simulationsstudie leitet sich aus der gängigen Praxis der Testkonstruktion ab (Estabrook & Neale, 2013). Die Auswertung vieler Testverfahren stützt sich auf Summenwerte, die über einzelne Skalen oder den gesamten Test oder Fragebogen (z.B. *Beck-Depressionsinventar – BDI-II*; Beck et al., 2006) hinweg gebildet werden. Dabei ist anzumerken, dass Summenwerte immer ungewichtet sind, bei der Berechnung von Summenwerten wird kein itemspezifischer Gewichtungsfaktor angewandt (Eid et al., 2013). Anhand der Summenwerte werden bei diesen Testverfahren norm- oder kriterienorientierte Aussagen über die Merkmalsausprägung getroffen. Demgegenüber erlauben Faktorwerte eine durch die Faktorladungen (Reliabilität) der Indikatoren gewichtete Einordnung des Grads der Merkmalsausprägung (Eid et al.).

Sofern das theoretische Modell dem tau-äquivalenten Messmodel (Bühner, 2011, S. 125; Eid et al., 2013, S. 831) genügt und insofern die Faktorladungen homogen sind, führen Gesamtsummenwert- und Faktorwertdiagnostik zum selben Ergebnis hinsichtlich der Schätzung der Merkmalsausprägungen der Individuen (DiStefano et al., 2009; Eid et al., 2013; Skrondal & Rabe-Hesketh, 2014). Der kritische Punkt bei der Verwendung der Gesamtsummenwerte zur Diagnostik ist jedoch der, dass die Faktorladungen in der angewandten psychologischen Forschung selten homogen sind (Buzick, 2010; Peterson, 2000)[29]. Sind die Faktorladungen heterogen bzw. die Indikatoren der Skala/des Testverfahrens unterschiedlich gewichtet (vgl. das tau-kongenerische Messmodell; Bühner, 2011, S. 125; Eid et al., 2013, S. 835), ergeben die Faktorwerte das bessere Abbild der Merkmalsausprägung (Eid et al., 2013; Estabrook & Neale, 2013; Skrondal & Rabe-Hesketh, 2014).

Die im vorherigen Absatz genannten Studien bezogen sich auf die Diagnostik auf Basis der Faktorwerte korrekt spezifizierter Modelle, die mit den Summenwerten verglichen wurden. Dabei wurden die Summenwerte der Indikatoren eines Faktors mit den Faktorwerten der einzelnen Faktoren verglichen. Daher stellt sich die weiterführende Frage, wie die Summenwerte im Vergleich zu den Faktorwerten eines missspezifizierten Modells hinsichtlich der Güte der Diagnostik abschneiden würden. Eingebettet in das Design der zweiten Studie lautet diese Nebenfragestellung konkret, ob ein Gesamtsummenwert oder die Faktorwerte eines einfaktoriellen, unterschiedlich stark missspezifizierten Modells zu besserer Diagnostik führen, wenn das Populationsmodell zweifaktoriell ist und die Faktorladungen des Populationsmodells heterogen definiert wurden.

Bevor jedoch in Kapitel V die psychometrischen Auswirkungen auf die Individuen beschrieben werden, die sich aus der Diagnostik basierend auf Faktorwerten missspezifizierter Modelle und auf Basis der Gesamtsummenwerte ergeben, wird im folgenden Kapitel IV zunächst der Forschungsfrage nachgegangen, inwieweit die gängigen Fit-Indizes die unterschiedlichen Schweregrade der Missspezifikation im Strukturmodell in Form einer nicht-

[29]Wie bereits unter III. 1 beschrieben, lagen in Petersons Metaanalyse bei einem Mittelwert von $\lambda = .32$ (standardisiert) 25% der Faktorladungen unter $\lambda = .23$ und 25% der Faktorladungen über $\lambda = .37$, die Faktorladungen fielen also heterogen aus.

spezifizierten Zweidimensionalität anhand der Cut-Off-Regeln nach Hu und Bentler (1998, 1999) als solche erkennen.

IV STUDIE 1

1 Methode

1.1 Stichprobenziehungen

Die unter III. 3 ausgeführte Fragestellung, inwiefern die Fit-Indizes CFI, RMSEA und SRMR fälschlicherweise als einfaktoriell spezifizierte Modelle anhand der Cut-Off-Regeln nach Hu und Bentler (1998, 1999) als nicht passend erkennen, wurde anhand einer Simulationsstudie untersucht. Es handelte sich dabei um eine Monte-Carlo-Simulation (vgl. Paxton, Curran, Bollen, Kirby, & Chen, 2001). Diese wurde mithilfe der „R"-Pakete (R Core Team, 2015) „lavaan" (Rosseel, 2012) und „simsem" (Pornprasertmanit, Miller, & Schoemann, 2015) realisiert. Um eine optimale Rechenleistung durch die Nutzung aller verfügbarer Prozessorkerne zu erreichen, wurde das Paket „parallel" (R Core Team, 2015) eingesetzt[30]. Für jede der zwölf Bedingungen, die im Folgenden beschrieben werden, wurden aus den Populationsmodellen jeweils 1,000 Stichproben-Kovarianz-Matrizen aus multivariat normalverteilten Daten erzeugt. Es wurden aus Replikationsgründen Startwerte für die Ziehung der Zufallszahlen gesetzt.

1.2 Design

Die Basis für die verschiedenen Populationsmodelle bildete ein lineares Strukturgleichungsmodell bestehend aus zwei korrelierten Faktoren und insgesamt 20 Indikatoren. Die Anzahl der Indikatoren wurde auf 20 festgelegt, da diese eine gängige Fragebogenlänge in der psychologischen Forschung darstellt (Peterson, 2000; Shrout & Yager, 1989). Für die zwölf verschiedenen experimentellen Bedingungen, aus denen die Populationsmodelle zusammengesetzt wurden, eignete sich ein 2(Faktorladungen: hoch versus typisch) × 2(Indikatorenaufteilung: 10:10 versus 15:5) × 3(Faktorkorrelation: .30 versus .50 versus .80) Studiendesign. Die Faktorladungshöhe wurde in Anlehnung an die Meta-Analyse

[30]An dieser Stelle sei Terrence Jorgensen ("lavaan Google Groups," 2015) für die Hilfe beim Schreiben der ineinander verschachtelten Schleifen-Funktion in R gedankt.

von Peterson (2000) vorab festgelegt. Demnach wurden die typischen Faktorladungen (standardisiert) für die Simulationsstudie gleichverteilt und zufällig aus dem Bereich [.20, .40] gezogen. Die hohen Faktorladungen (standardisiert) wurden aus dem Bereich [.40, .60] zufällig und gleichverteilt gezogen. Die Varianzen der latenten Variablen wurden bei allen für die Simulation verwendeten und bei allen auf die Simulationsdaten angewandten Modellen auf Eins gesetzt. Die Bedingungen, die sich aus den Kombinationen der Faktorstufen der unabhängigen Variablen Itemaufteilung und Faktorkorrelation ergaben, bestimmten den Grad der Missspezifikation. Die 20 Indikatoren wurden im Rahmen von sechs der zwölf Bedingungen gleichmäßig auf beide Faktoren aufgeteilt und im Rahmen der anderen sechs Bedingungen wurde der erste Faktor durch 15 Items repräsentiert und der zweite Faktor durch 5 Items. Eine unausgewogene Indikatorenaufteilung sollte daher einen geringeren Grad an Missspezifikation darstellen als eine ausgewogene Aufteilung, da erstere einen Faktor stärker repräsentiert als den anderen und diese Bedingung somit näher am einfaktoriellen Modell liegt. Weiters wurde in den Populationsmodellen die Höhe der Korrelation zwischen beiden Faktoren variiert. Oblique Faktorenstrukturen sind typisch in der Psychologie, die durchschnittliche Korrelation zwischen zwei Variablen beträgt $r = .30$, was einem mittelhohen Zusammenhang entspricht (Cohen & Manion, 1980). Somit stellte im Rahmen dieser Studie eine Faktorkorrelation von $r = .30$ eine in der Psychologie typisch hohe Korrelation[31] und einen hohen Grad an Missspezifikation dar. Eine in der Psychologie eher seltene hohe Korrelation wurde durch $r = .50$[32] repräsentiert, diese entsprach einem mittleren Grad an Missspezifikation. Die Faktorkorrelation von $r = .80$ wurde aufgenommen, um Bedingungen zu untersuchen, die noch näher an der Eindimensionalität liegen bzw. latente Variablen darstellen, die annähernd dasselbe messen (geringe Missspezifikation).

1.3 Durchführung

Fälschlicherweise als einfaktoriell spezifizierte Messmodelle mit 20 Indikatoren wurden auf die aus den zweifaktoriellen Populationsmodellen erzeugten Stichproben angewandt. Die Schätzung der Modellpassung wurde mit dem Maximum-Likelihood-Algorithmus

[31]Die Metaanalyse von Steel, Schmidt, und Shultz (2008) zeigte beispielsweise über 17 Studien hinweg, dass die Big-Five-Faktoren Neurotizismus und Extraversion zu $r = -.33$ korrelierten.

[32]So wurde zum Beispiel eine Korrelation von $r = .50$ zwischen Intelligenz und Schulerfolg bereits mehrfach bestätigt (Rost, 2009).

vorgenommen, da dieser die am häufigsten genutzte Schätzmethode darstellt (Beauducel & Wittmann, 2005; Eid et al., 2013; Mahler, 2011; Reinecke, 2014; Schermelleh-Engel et al., 2003). Es wurde ausgewertet, wie oft das missspezifizierte Modell an den erzeugten Stichprobendaten durch die Fit-Indizes CFI, RMSEA und SRMR anhand der Cut-Off-Kriterien und der Kombinationsregel von Hu und Bentler (1998, 1999) zurückgewiesen wurde.

2 Ergebnisse

2.1 Nonzentralitätsparameter

Der Nonzentralitätsparameter stieg mit sinkender Faktorkorrelation, d.h. der Nonzentralitätsparameter zeigte mit sinkender Faktorkorrelation zunehmende Modellabweichung an (siehe Tabelle 3). Ebenso fiel der Nonzentralitätsparameter bei der ungleichmäßigen Indikatorenaufteilung niedriger aus als bei der ausgewogenen Indikatorenaufteilung (siehe Tabelle 3). Insgesamt passte also die intuitive Ordnung des Grades der Missspezifikation mit derjenigen durch den Nonzentralitätsparameter zusammen.

Tabelle 3

Nonzentralitätsparameter bei den missspezifizierten Modellen

	Modell 1: .80 [.20, .40] 10:10	Modell 2: .50 [.20, .40] 10:10	Modell 3: .30 [.20, .40] 10:10	Modell 4: .80 [.20, .40] 15:5	Modell 5: .50 [.20, .40] 15:5	Modell 6: .30 [.20, .40] 15:5
NCP	18.263	88.588	154.873	8.653	34.199	51.352
	Modell 7: .80 [.40, .60] 10:10	Modell 8: .50 [.40, .60] 10:10	Modell 9: .30 [.40, .60] 10:10	Modell 10: .80 [.40, .60] 15:5	Modell 11: .50 [.40, .60] 15:5	Modell 12: .30 [.40, .60] 15:5
NCP	98.99	435.531	710.532	47.058	182.389	258.958

Anmerkungen. NCP = gemittelter Nonzentralitätsparameter über alle 1000 Stichprobenkovarianzmatrizen hinweg. Die Zellen der Beschreibung der Populationsmodelle enthalten als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

2.2 Korrekte Modelle

2.2.1 Ergebnisse hinsichtlich des χ^2-Tests bezüglich der korrekten Modelle

Es wurde die exploratorische Fragestellung untersucht, inwieweit die gängigen Fit-Indizes CFI, RMSEA und SRMR eine fälschlicherweise als einfaktoriell spezifiziertes Modell im Gegensatz zu einem zweifaktoriellen Modell, das den Stichprobendaten zugrunde liegt,

anhand der Cut-Off-Kriterien nach Hu und Bentler (1998, 1999) korrekterweise zurückwiesen. Aus Vollständigkeits- und Vergleichsgründen wurde zunächst die Modellpassung korrekter Modelle in den Stichproben anhand des χ^2-Tests inspiziert.

Die korrekten Modelle wurden an allen Stichprobenkovarianzmatrizen durch den χ^2-Test angenommen, wobei der Mittelwert des χ^2-Werts über alle Bedingungen und Stichprobenziehungen hinweg bei $M = 168.404$ ($SD = 3.937$), also sehr nahe an der Anzahl der Freiheitsgrade ($df = 169$) der korrekten Modelle, lag.

2.2.2 Ergebnisse hinsichtlich des CFI bezüglich der korrekten Modelle

Aus Vergleichsgründen wurden die drei Fit-Indizes zunächst in Bezug auf korrekt spezifizierte Modelle untersucht. Dabei wurden korrekt spezifizierte Modelle mit frei zu schätzenden Parametern auf die aus den Populationsmodellen erzeugten Stichprobendaten angewandt und ausgezählt, wie oft die korrekten Modelle durch die Fit-Indizes angenommen wurden.

Die Mittelwerte des CFI lagen in allen zwölf Bedingungen über dem von Hu und Bentler (1998, 1999) vorgeschlagenen Cut-Off von .95. In den sechs Bedingungen mit den hohen Ladungen wurde das korrekte Modell an allen jeweils 1000 Stichproben korrekterweise angenommen. In den sechs Bedingungen mit den typischen Faktorladungen wurden mindestens 931 korrekte Modelle pro Bedingung angenommen, wobei die Anzahl der korrekterweise angenommenen Modelle durch den CFI mit der Faktorkorrelation im Populationsmodell stieg und außerdem zugunsten der ungleichen Indikatorenaufteilung ausfiel.

2.2.3 Ergebnisse hinsichtlich des RMSEA bezüglich der korrekten Modelle

Innerhalb der zwölf verschiedenen Simulationsbedingungen zeigte der RMSEA sehr konsistent eine sehr gute Modellpassung anhand der Cut-Off-Kriterien, wenn ein korrektes, frei zu schätzendes Modell auf die jeweils aus dem Populationsmodell erzeugten Stichprobenkovarianzmatrizen angewandt wurde. In allen zwölf Bedingungen wurde das korrekte Modell an allen jeweils 1000 Stichproben durch den RMSEA korrekterweise als passend indiziert. Ebenfalls lag der Mittelwert des RMSEA in den einzelnen zwölf Bedingungen über die jeweils 1000 Stichprobendaten hinweg unter dem Cut-Off von Hu und Bentler (1998, 1999).

2.2.4 Ergebnisse hinsichtlich des SRMR bezüglich der korrekten Modelle

Das SRMR verhielt sich ähnlich wie der RMSEA und zeigte über alle zwölf Bedingungen hinweg eine sehr gute Modellpassung. Die korrekten Modelle wurden an allen 1000 Stichprobenmatrizen je Bedingung nach dem Cut-Off von Hu und Bentler (1998, 1999) angenommen. Weiters lagen die Mittelwerte des SRMR stets unter dem Cut-Off von Hu und Bentler.

2.3 Missspezifizierte Modelle

2.3.1 Ergebnisse hinsichtlich des χ^2-Tests bezüglich der missspezifizierten Modelle

Das missspezifizierte Modell wurde in allen Bedingungen über alle Stichprobenkovarianzmatrizen hinweg abgelehnt. Die unterschiedlichen Grade an Missspezifikation zeigten sich anhand der χ^2-Werte, mit steigendem Grad an Missspezifikation stieg der χ^2-Wert (siehe Tabelle 4): Der χ^2-Wert stieg mit sinkender Faktorkorrelation im Populationsmodell (höherer Grad an Missspezifikation) und war im Rahmen der Bedingungen mit der ungleichen Indikatorenaufteilung auf die Faktoren im Populationsmodell (niedrigerer Grad an Missspezifikation) niedriger als in den Bedingungen mit der ausgewogenen Aufteilung. Bei den hohen Faktorladungen im Populationsmodell waren auch die χ^2-Werte höher als bei den typischen Faktorladungen (siehe Tabelle 4). Dies zeigt, dass der χ^2-Test bei hohen Faktorladungen sensitiver war für die Missspezifikation.

Tabelle 4

Ergebnisse bezüglich des χ^2-Tests hinsichtlich der missspezifizierten Modelle

	Modell 1: .80 [.20, .40] 10:10	Modell 2: .50 [.20, .40] 10:10	Modell 3: .30 [.20, .40] 10:10	Modell 4: .80 [.20, .40] 15:5	Modell 5: .50 [.20, .40] 15:5	Modell 6: .30 [.20, .40] 15:5
M χ^2(170)	188.263	258.588	324.873	178.653	204.199	221.352
SD χ^2(170)	20.365	27.774	33.873	19.120	22.451	23.903
	Modell 7: .80 [.40, .60] 10:10	Modell 8: .50 [.40, .60] 10:10	Modell 9: .30 [.40, .60] 10:10	Modell 10: .80 [.40, .60] 15:5	Modell 11: .50 [.40, .60] 15:5	Modell 12: .30 [.40, .60] 15:5
M χ^2(170)	268.990	605.531	880.532	217.058	352.389	428.958
SD χ^2(170)	28.954	56.755	74.487	23.045	35.252	40.719

Anmerkungen. M = Mittelwert des χ^2-Werts über alle 1000 Stichprobenkovarianzmatrizen hinweg, SD = Standardabweichung des χ^2-Werts über alle 1000 Stichprobenkovarianzmatrizen hinweg. Die Zellen der Beschreibung der Populationsmodelle enthalten als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

2.3.2 Ergebnisse hinsichtlich des CFI bezüglich der missspezifizierten Modelle

Es zeigten sich Haupteffekte zugunsten der Anzahl an abgelehnten missspezifizierten Modellen sowohl für die Höhe der Faktorkorrelation als auch die Balancierung der Indikatoren in den Populationsmodellen und die Höhe der Faktorladungen (letzterer Haupteffekt ist allerdings nicht interpretierbar; Erklärung folgt).

Hinsichtlich der Ausgewogenheit der Anzahl der Indikatoren zeigte sich, dass die ausgewogene Itemaufteilung (hohe Missspezifikation) vorteilhafter hinsichtlich der Entdeckung der Missspezifikation war als die unausgewogene Itemverteilung (geringe Missspezifikation), die missspezifizierten Modelle wurden an den Daten der ersteren Populationsmodelle öfter abgelehnt (siehe Tabelle 5). Ebenso führte die unausgewogene Itemaufteilung im Mittel zu einem höheren CFI als die ausgewogene Itemaufteilung.

Desto höher die Faktorkorrelation, desto weniger häufig wurden die missspezifizierten Modelle an den Stichprobendaten abgelehnt (siehe Tabelle 5). Ebenso stieg mit der Faktorkorrelation (sinkender Grad an Missspezifikation) im Populationsmodell auch der Mittelwert des CFI bei Anwendung des missspezifizierten Modells auf die Stichprobendaten.

Tabelle 5

Ergebnisse bezüglich des CFI hinsichtlich der missspezifizierten Modelle

	Modell 1: .80 [.20, .40] 10:10	Modell 2: .50 [.20, .40] 10:10	Modell 3: .30 [.20, .40] 10:10	Modell 4: .80 [.20, .40] 15:5	Modell 5: .50 [.20, .40] 15:5	Modell 6: .30 [.20, .40] 15:5
M(SD) CFI	.971(.025)	.848(.043)	.713(.051)	.983(.020)	.946(.033)	.916(.037)
Korr. Zurückw.	191	996	1000	81	539	826
	Modell 7: .80 [.40, .60] 10:10	Modell 8: .50 [.40, .60] 10:10	Modell 9: .30 [.40, .60] 10:10	Modell 10: .80 [.40, .60] 15:5	Modell 11: .50 [.40, .60] 15:5	Modell 12: .30 [.40, .60] 15:5
M(SD) CFI	.960(.011)	.805(.024)	.668(.030)	.982(.009)	.923(.015)	.886(.017)
Korr. Zurückw.	179	1000	1000	1	974	1000

Anmerkungen. M = Mittelwert, SD = Standardabweichung, Korr. Zurückw. = Anzahl korrekter Zurückweisungen durch den CFI. Die Zellen der Beschreibung der Populationsmodelle enthalten als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

Hinsichtlich der Anzahl an korrekt zurückgewiesenen missspezifizierten Modellen durch den Cut-Off nach Hu und Bentler (1998, 1999) für den CFI interagierte die Höhe der Faktorladungen hybrid mit der Höhe der Faktorkorrelation (siehe Tabelle 5). Bei geringer und mittlerer Faktorkorrelation (hoher und mittlerer Grad an Missspezifikation) im Populationsmodell war die Anzahl an korrekt zurückgewiesenen missspezifizierten Modellen bei den hohen Ladungen höher als bei den typischen Ladungen. Bei einer hohen Faktorkorrelation (geringer Grad an Missspezifikation) drehte sich dieses Muster um; bei hohen Ladungen wurden die missspezifizierten Modelle weniger oft zurückgewiesen als bei typischen Ladungen. Die Mittelwerte des CFI über die Stichproben eines Populationsmodelles hinweg lagen jedoch bei hohen Ladungen stets unter denen bei typischen Ladungen. Dieser Befund wird unter 3.2 diskutiert.

2.3.3 Ergebnisse bezüglich des RMSEA hinsichtlich der missspezifizierten Modelle

Insgesamt betrachtet zeigte der RMSEA eine äußerst geringe Sensitivität hinsichtlich der Indizierung des missspezifizierten eindimensionalen Modells (siehe Tabelle 6). Lediglich

in der am höchsten missspezifizierten Bedingung (Modell 9) mit hohen Faktorladungen, einer ausgewogenen Indikatorenaufteilung sowie einer geringen Faktorkorrelation im Populationsmodell wurde das missspezifizierte Modell in 914 von 1000 Fällen als nicht passend indiziert. Ebenso lag der Mittelwert über die 1000 Stichprobendaten pro Bedingung hinweg nur in dieser Bedingung über dem Cut-Off von .06.

Tabelle 6

Ergebnisse hinsichtlich des RMSEA bezüglich der missspezifizierten Modelle

	Modell 1: .80 [.20, .40] 10:10	Modell 2: .50 [.20, .40] 10:10	Modell 3: .30 [.20, .40] 10:10	Modell 4: .80 [.20, .40] 15:5	Modell 5: .50 [.20, .40] 15:5	Modell 6: .30 [.20, .40] 15:5
M(SD) RMSEA	.009(.006)	.023(.004)	.030(.003)	.006(.006)	.013(.005)	.017(.004)
Korr. Zurückw.	0	0	0	0	0	0
	Modell 7: .80 [.40, .60] 10:10	Modell 8: .50 [.40, .60] 10:10	Modell 9: .30 [.40, .60] 10:10	Modell 10: .80 [.40, .60] 15:5	Modell 11: .50 [.40, .60] 15:5	Modell 12: .30 [.40, .60] 15:5
M(SD) RMSEA	.024(.004)	.051(.003)	.065(.003)	.016(.005)	.033(.003)	.034(.003)
Korr. Zurückw.	0	3	914	0	0	0

Anmerkungen. M = Mittelwert, SD = Standardabweichung, Korr. Zurückw. = Anzahl der korrekten Zurückweisungen durch den RMSEA. Die Zellen der Beschreibung der Populationsmodelle enthalten als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

Die Mittelwerte des RMSEA lagen in den Bedingungen mit den typischen Ladungen im Populationsmodell geringfügig unter denen der hohen Ladungen. Allerdings ist dieser Unterschied vernachlässigbar, zumal die Mittelwerte des RMSEA in allen Bedingungen außer der genannten neunten unter dem Cut-Off von Hu und Bentler (1998, 1999) lagen, die Modellabweichung also nicht identifizierten. Die Unterschiede hinsichtlich der Mittelwerte bezüglich des RMSEA bei einer geringen Faktorkorrelation (höherer Wert des RMSEA im Vergleich zu den Bedingungen mit einer mittleren oder hohen Faktorkorrelation im Populationsmodell) und bei einer ausgewogenen Indikatorenaufteilung (höherer Wert des RMSEA im Vergleich zu einer unausgewogenen Indikatorenaufteilung) sind ebenfalls nur marginal.

2.3.4 Ergebnisse bezüglich des SRMR hinsichtlich der missspezifizierten Modelle

Die Ergebnisse hinsichtlich des SRMR sind noch kritischer als die bezüglich des RMSEA (siehe Tabelle 7).

Der Mittelwert des SRMR lag in allen Bedingungen unter dem Cut-Off nach Hu und Bentler (1998, 1999). Ebenfalls wurden nur 70 von 1000 Stichprobenkovarianzmatrizen in der neunten Bedingung, also der am wenigsten eindimensionalen Bedingung kombiniert mit hohen Faktorladungen, durch das SRMR zurückgewiesen. Betrachtet man die Mittelwerte, wird zwar deutlich, dass die Sensitivität des SRMR mit sinkender Faktorkorrelation (höherer Missspezifikation), hohen Ladungen (im Vergleich zu typischen Ladungen) und einer ausgewogenen Indikatorenaufteilung (höhere Missspezifikation im Vergleich zu einer unausgewogenen) stieg, dies aber äußerst geringfügig.

Tabelle 7

Ergebnisse hinsichtlich des SRMR bezüglich der missspezifizierten Modelle

	Modell 1: .80 [.20, .40] 10:10	Modell 2: .50 [.20, .40] 10:10	Modell 3: .30 [.20, .40] 10:10	Modell 4: .80 [.20, .40] 15:5	Modell 5: .50 [.20, .40] 15:5	Modell 6: .30 [.20, .40] 15:5
M(SD) SRMR	.028(.002)	.034(.002)	.040(.003)	.027(.001)	.029(.002)	.031(.002)
Korr. Zurückw.	0	0	0	0	0	0
	Modell 7: .80 [.40, .60] 10:10	Modell 8: .50 [.40, .60] 10:10	Modell 9: .30 [.40, .60] 10:10	Modell 10: .80 [.40, .60] 15:5	Modell 11: .50 [.40, .60] 15:5	Modell 12: .30 [.40, .60] 15:5
M(SD) SRMR	.030(.002)	.055(.003)	.073(.004)	.027(.002)	.038(.002)	.045(.002)
Korr. Zurückw.	0	0	70	0	0	0

Anmerkungen. M = Mittelwert, SD = Standardabweichung, Korr. Zurückw. = Anzahl korrekter Zurückweisungen durch den SRMR. Die Zellen der Beschreibung der Populationsmodelle enthalten als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

Wendet man den Cut-Off von < .09 für das SRMR im Rahmen der Kombinationsregel nach Hu und Bentler (1999) zusammen mit dem RMSEA oder dem CFI an, wurde das missspezifizierte Modell auch in der neunten Bedingung an allen 1,000 Stichprobenkovarianzmatrizen durch das SRMR angenommen. Das heißt, bei den 70

Stichproben, an denen das missspezifizierte Modell bei einem Cut-Off von < .08 für den SRMR noch abgelehnt wurde, lag das SRMR zwischen .08 und .09.

3 Diskussion

3.1 Zusammenfassung der Ergebnisse

Es wurde die Fragestellung untersucht, inwieweit drei der am meist genutzten Fit-Indizes CFI, RMSEA und SRMR Missspezifikationen im Strukturmodell im Kontext realistisch hoher und heterogener Faktorladungen anhand der Cut-Offs nach Hu und Bentler (1998, 1999) zurückweisen würden. Die untersuchte Art der Missspezifikation stellte eine einfaktorielles Modell im Gegensatz zu einem obliquen zweifaktoriellen Populationsmodell dar. Um unterschiedliche Schweregrade an Missspezifikationen im Strukturmodell zu erreichen, wurde die Höhe der Korrelation zwischen den latenten Variablen im Populationsmodell variiert, ebenso die (Un-)Ausgewogenheit der Indikatoren pro latenter Variable. Pro Populationsmodell wurden 1,000 Stichprobenkovarianzmatrizen erzeugt, auf die das eindimensionale missspezifizierte Modell angewandt wurde.

Alle drei Fit-Indizes zeigten bei Anwendung korrekter Modelle auf die erzeugten Daten anhand der Kriterien nach Hu und Bentler (1998, 1999) eine gute Modellpassung. Die Fit-Indizes konnten also unter der besagten Cut-Off-Bedingung ohne Probleme korrekte Modelle als korrekt identifizieren. Hinsichtlich des missspezifizierten Modells zeigte sich ein anderes Bild.

Der CFI wies das schwer und mittelgradig missspezifizierte Modell anhand des Cut-Offs von Hu und Bentler ausreichend oft zurück. Die Höhe der Faktorkorrelation in Interaktion mit der Höhe der Faktorladungen stellte einen wichtigeren Einflussfaktor als die Indikatorenaufteilung auf den CFI dar, die missspezifizierten Modelle abzulehnen. Bei geringer und mittlerer Faktorkorrelation (hoher und mittlerer Grad an Missspezifikation) lag der Mittelwert des CFI immer unter dem Cut-Off von Hu und Bentler (1998, 1999) und das missspezifizierte Modell wurde an der überwiegenden Mehrheit an Stichprobenkovarianzmatrizen abgelehnt. Bei hohen Ladungen lagen die Mittelwerte des CFI stets unter den Mittelwerten bei typischen Ladungen und mit Ausnahme der Bedingungen mit der hohen Faktorkorrelation (niedrige Missspezifikation) wurden in ersteren Bedingungen auch so gut wie alle Modelle an den Stichprobendaten durch den CFI abgelehnt. Eine ausgewogene Indikatorenaufteilung erwies sich als vorteilhafter hinsichtlich der Entdeckung des Misfit (niedrigere Mittelwerte des CFI und mehr Zurückweisungen des missspezifizierten Modells)

als die unausgewogene Aufteilung, da erstere näher an der Zwei-Faktoren-Struktur lag als an der Ein-Faktoren-Struktur.

Während sich die Sensitivität des CFI hinsichtlich der Zurückweisung des missspezifizierten Modells – mit Ausnahme der Bedingungen mit hohen Faktorkorrelationen – als vergleichsweise hoch erwies, zeigte sich bezüglich des RMSEA und des SRMR ein anderes Ergebnis. Beide Fit-Indizes bewährten sich hinsichtlich der Ablehnung des missspezifizierten Modells nicht, sofern die Cut-Off-Werte nach Hu und Bentler (1998, 1999) als Kriterium galten. Während der Mittelwert des RMSEA in der am höchsten missspezifizierten neunten Bedingung noch über dem Cut-Off nach Hu und Bentler lag und das missspezifizierte Modell am Großteil der Stichprobenmatrizen in dieser Bedingung abgelehnt wurde, lag der Mittelwert des SRMR in allen Bedingungen unter dem Cut-Off von Hu und Bentler und wies das missspezifizierte Modell nicht zurück. Beide Indizes zeigten hohe Typ-2-Fehlerraten und insofern eine äußerst geringe Sensitivität hinsichtlich der Entdeckung der Modellabweichungen.

Die Kombinationsregel aus CFI oder RMSEA zusammen mit dem SRMR wäre im Rahmen des Designs der vorliegenden Studie nur für den CFI zusammen mit dem SRMR und nur bei einer mittleren oder schwerwiegenden Missspezifikation (geringe und mittelhohe Faktorkorrelation) zielführend bei der Identifikation der Missspezifikation gewesen.

Im Gegensatz zu den Fit-Indizes konnte die Missspezifikation an allen Stichprobendaten der verschiedenen Populationsmodelle anhand des χ^2-Tests als solche identifiziert werden. Ebenso stieg der Grad der Missspezifikation mit dem absoluten Wert des Nonzentralitätsparameters.

3.2 Diskussion der Ergebnisse

Während der CFI im Rahmen dieser Studie eine ausreichende Sensitivität zur Entdeckung des Misfit zeigte, sofern die Faktorkorrelation im Populationsmodel nicht zu hoch (die Missspezifikation zu niedrig) war, zeigten die untersuchten absoluten Fit-Indizes keine Modellabweichung an, sofern die Cut-Offs nach Hu und Bentler (1998, 1999) als Kriterien zur Modellevaluation herangezogen wurden. Das SRMR schnitt im Rahmen des untersuchten Designs am schlechtesten ab. Dabei wurde von Hu und Bentler propagiert, dass das SRMR gerade für Missspezifikationen im Strukturmodell sensitiv wäre. Bei Mahler (2011) und Savalei (2012) zeigten sich ähnliche Ergebnisse wie die vorliegenden für die untersuchten Fit-Indizes.

Insgesamt schnitten die absoluten Fit-Indizes RMSEA und SRMR besser ab, wenn es sich bei der Art der Missspezifikation um nicht-spezifizierte Messfehlerkovarianzen handelte (vgl. Heene et al., 2012; Mahler, 2011; Savalei, 2012) als um die Art der Missspezifikation im Strukturmodell, die im Rahmen dieser Simulationsstudie untersucht wurde. Diese Befunde sind vor dem Hintergrund, dass Missspezifikationen auf der Ebene der latenten Variablen konzeptuell als schwerwiegender angesehen werden können als Missspezifikationen im Messmodell (Mahler, 2011), bedenklich. Der Umstand, dass die Stichprobengröße in dieser Studie für die angewandte Forschung sehr hoch gewählt wurde und auch die Voraussetzung der multivariaten Normalverteilung erfüllt war, macht die Ergebnislage noch gravierender.

Die Sensitivität dieser drei Fit-Indizes für die Missspezifikation im Strukturmodel fiel in dieser Studie in absoluten Werten geringfügig niedriger aus als in den Studien von Mahler (2011) und Savalei (2012) in den hinsichtlich der Missspezifikation äquivalenten Bedingungen. Letzteres kann möglicherweise auf die in der Hälfte der Bedingungen noch niedrigeren Faktorladungen sowie die Heterogenität der Faktorladungen oder auch die unterschiedliche Indikatorenanzahl[33] im Rahmen der vorliegenden Studie zurückgeführt werden. Andererseits wurden in den Studien von Mahler und Savalei keine Stichprobengrößen genannt, sodass auch die Stichprobengröße als möglicher Grund für den Unterschied genannt werden kann.

Hinsichtlich der Wichtigkeit der Einflussfaktoren auf die Sensitivität der Fit-Indizes zeigten sich ähnliche Ergebnisse wie bei Savalei (2012) für den RMSEA. Die Höhe der Faktorladungen sowie die Höhe der Faktorkorrelation, insbesondere in Interaktion, stellten sich als wichtige Einflussfaktoren auf die Sensitivität der Indizes heraus. In Savaleis Studien hatte die Höhe der Faktorladungen den höchsten Einfluss auf die Sensitivität des RMSEA. Dieser Unterschied kann ebenfalls dadurch erklärt werden, dass die untersuchte Spannweite an Faktorladungen in der vorliegenden Studie kleiner war als bei Savalei und die Faktorladungen absolut niedriger waren als bei Savalei. Dies führte zu Bodeneffekten für den RMSEA und das SRMR, welche eine klare Ordnung der Einflussfaktoren hinsichtlich ihrer Wirkung erschwerten, zumal die Höhe der Faktorkorrelation ähnlich wie bei Mahler (2011) mit der Höhe der Faktorladungen interagierte. Jedoch hatte auch die Indikatorenaufteilung auf die beiden latenten Variablen im Populationsmodell vor allem einen Einfluss auf die Sensitivität des CFI, die ausgewogene Aufteilung auf die latenten Faktoren erwies sich als vorteilhafter hinsichtlich

[33]Savalei (2012) verwendete insgesamt 8 Indikatoren, Mahler (2011) insgesamt 12, wohingegen in der vorliegenden Studie 20 Indikatoren insgesamt verwendet wurden.

der Sensitivität für die Missspezifikation. Bei RMSEA und SRMR ging die Tendenz zwar in dieselbe Richtung, allerdings waren die Unterschiede aufgrund der genannten Bodeneffekte dieser beiden Indizes marginal.

Hinsichtlich der Höhe der Faktorladungen bestätigte sich für den RMSEA und das SRMR, was von Heene et al. (2011) gezeigt wurde und was auch im Rahmen der Studien von Mahler (2011) und Savalei (2012) auftrat: Bei niedrigeren Faktorladungen waren RMSEA und SRMR weniger sensitiv gegenüber Missspezifikationen. Dieser Befund resultiert aus dem positiven Zusammenhang zwischen Faktorladungshöhe und der Höhe des χ^2-Werts, der sich auch im Rahmen dieser Studie zeigte. Sofern die Uniqueness-Matrix durch niedrigere Faktorladungen größer wird, sinken die Eigenwerte dieser Matrix und die χ^2-Test-Statistik, die diese Eigenwerte enthält, sinkt auch (Heene et al., S. 329). Insofern verfehlen Fit-Indizes wie der RMSEA und das SRMR, die auf der Differenz zwischen beobachteter und implizierter Kovarianzmatrix basieren, ihre Funktion, missspezifizierte Modelle abzulehnen.

Der CFI hingegen, der als inkrementeller Fit-Index den Vergleich mit einem Nullmodell heranzieht, markierte bei Heene et al. (2011) bei sinkenden Ladungen das missspezifizierte Modell in höherem Maße als abweichend (vgl. die Ergebnisse von Beauducel und Wittmann, 2005), da die Differenz zwischen impliziertem Modell und Nullmodell geringer wird. Allerdings zeigte sich dieses Muster in der vorliegenden Studie nicht. In der vorliegenden Studie waren die Mittelwerte des CFI über die jeweils 1,000 Stichproben eines Populationsmodells hinweg bei den typischen Ladungen geringfügig höher als bei den hohen Ladungen. Dies zeigte sich auch in der Studie von Mahler (2011) mit vergleichbarem Design. Hinsichtlich der Anzahl an korrekten Zurückweisungen des missspezifizierten Modells durch den CFI zeigte sich in der vorliegenden Studie ein komplexeres Muster: Bei hohen Ladungen und einer hohen Faktorkorrelation wurde das missspezifizierte Modell weniger oft an den Stichprobendaten abgelehnt als bei typischen Ladungen und einer hohen Faktorkorrelation (dieser Befund deckt sich mit Heene et al.s [2011] Ergebnissen; steht jedoch in Kontrast zum Verhalten der Mittelwerte des CFI in der vorliegenden Studie). Bei hohen Ladungen und einer geringen oder mittleren Faktorkorrelation im Populationsmodell wurde das missspezifizierte Modell anhand der Cut-Offs nach Hu und Bentler (1998, 1999) öfter zurückgewiesen als bei den typischen Ladungen und einer geringen und mittleren Faktorkorrelation im Populationsmodell (dieser Befund steht in Kontrast zu Heene et al.s [2011] Befunden, deckt sich allerdings mit den Mittelwerten des CFI in der vorliegenden Studie). Mahler (2011) berechnete nur Mittelwerte für die Fit-Indizes über die Stichproben pro Populationsmodell

hinweg und nicht noch zusätzlich die Anzahl an korrekten Zurückweisungen des missspezifizierten Modells durch die Cut-Off-Werte nach Hu und Bentler (1998, 1999), sodass letzterer Befund nicht mit Mahler verglichen werden kann. Eine mögliche Erklärung für das Ergebnismuster des CFI stellt dar, dass im Prinzip nichts über die Verteilung der Teststatistik des unkorrelierten Baseline-Modells bekannt ist, auf der die inkrementellen Fit-Indizes, so auch der CFI, basieren (Curran et al., 2002). Insofern ist auch nicht bekannt, ob die Teststatistik einer (non-)zentralen χ^2-Verteilung folgt. Abgesehen davon führen Curran et al. an, dass die relativen Fit-Indizes keine linearen Funktionen der Teststatistik für das unkorrelierte Basismodell und das implizierte Modell darstellen. Demnach verkompliziert sich die Einordnung des Grades der Missspezifikation. Vermutet wird von den Autoren außerdem, dass die (Non-)Zentralität der Verteilung der Teststatistik auch abhängig ist von der Art der Missspezifikation, nicht nur von deren Schweregrad.

Hinsichtlich der Anwendung der korrekten Modelle zeigten sich die Fit-Indizes homogen in der Hinsicht, als dass sie alle drei eine gute Modellpassung nach den Cut-Off-Kriterien anzeigten. Hinsichtlich der Anwendung des missspezifizierten Modells zeigte sich im Rahmen des untersuchten Designs, was bereits von Beauducel und Wittmann (2005) sowie Fan et al. (2009) im Kontext deren Designs beschrieben wurde: Die Fit-Indizes reagierten heterogen auf die verschiedenen Modellbedingungen. Bei Beauducel und Wittmann stieg mit höheren Ladungen allerdings auch der Grad der Missspezifikation (Savalei, 2012). Dies führte dazu, dass die Fit-Indizes mit steigenden Ladungen homogenere Ergebnisse zeigten, da die Fit-Indizes bei höheren Ladungen sensitiver für die Missspezifikation wurden.

Als Schlussfolgerung aus dieser wie auch aus den unter III. 1 und 2 beschriebenen vorherigen Simulationsstudien kann wiederum gezogen werden, dass allgemeingültige Cut-Offs für die Fit-Indizes kaum zu definieren sind. Erstens haben inzidentelle Parameter eines Modells, wie die Faktorladungshöhe (oder auch die Stichprobengröße; siehe z.B. Beauducel und Wittmann, 2005) einen Einfluss auf die Modellpassung anhand der Fit-Indizes (vgl. Heene et al., 2011, wie auch die vorliegende Studie), zweitens die Art und der Schweregrad einer Missspezifikation (vgl. Fan et al., 2009, wie auch die vorliegende Studie) und drittens die Interaktion aus inzidentellen und die Missspezifikation determinierenden Modellbedingungen (vgl. Savalei, 2012, wie auch die vorliegende Studie). Die zusätzliche Verwendung der Modifikationsindizes kann zur Verbesserung der Beurteilung der Modellpassung nicht empfohlen werden, da diese keine reliablen Indikatoren für den Ort der Missspezifikation darstellen (Kaplan, 1988).

Entgegen der Befunde zu den Fit-Indizes identifizierte der χ^2-Test im Rahmen dieser Studie die Missspezifikation über alle Bedingungen hinweg und an allen Stichprobenkovarianzmatrizen. Aufgrund dessen, dass der χ^2-Wert die unterschiedlichen Grade der Missspezifikation anzeigte, konnte auch der Nonzentralitätsparameter, wie von Fan und Sivo (2005) sowie Fan et al. (2009) vorgeschlagen, zur Bestimmung des Grades der Missspezifikation dienen.

Die Empfehlungen, die sich aus dieser Simulationsstudie ableiten lassen, sind keineswegs neu. Zum einen wird generell empfohlen, mehrere Fit-Indizes (siehe II. 2) nur in Kombination mit dem χ^2-Test unter Berücksichtigung seiner Freiheitsgrade zur Beurteilung der Modellpassung heranzuziehen (Schermelleh-Engel et al., 2003). Die Befunde aus der vorliegenden Studie legen insbesondere nahe, dass bei der Modellevaluation auf den χ^2-Test zu achten ist, da dieser die Modellabweichung im Gegensatz zu den Fit-Indizes anzeigte. Als Empfehlung für die Anwendung kann daher gegeben werden, dass Vorsicht geboten ist, sobald die Fit-Indizes Modellpassung anzeigen, der χ^2-Test aber nicht. Der χ^2-Test gibt allerdings nicht die Größe der Missspezifikation an (Saris et al., 1987), daher ist die gleichzeitige Betrachtung der lokalen Fitmaße unabdingbar.

Zum anderen bestätigen diese und andere Befunde (z.B. Beauducel & Wittmann, 2005; Heene et al., 2011; Savalei, 2012), wie wichtig eine gute Test- und Skalenkonstruktion im Vorfeld ist. Es wurde bereits im Rahmen zahlreicher Studien (Heene et al., 2011; Savalei, 2012; siehe III. 1 und 2) bestätigt, dass hohe Ladungen die Sensitivität der Fit-Indizes für Missspezifikationen erhöhen und dass die Ladungen im Rahmen dieser Studie zwar für die Anwendung typisch hoch waren (vgl. Peterson, 2000), für die Entdeckung der Missspezifikation aber zu niedrig. Ebenso wurde im Rahmen dieses Studiendesigns gezeigt, dass sowohl die Höhe der Faktorkorrelation als auch die Ausgewogenheit der Anzahl der Indikatoren pro latenter Variable im Populationsmodell insbesondere für die Sensitivität des CFI relevant war, Missspezifikationen im Strukturmodell in Form von Dimensionalitätsverletzungen in der Faktorenstruktur zu entdecken. Um den letzteren Befund verallgemeinern zu können, bedarf es allerdings weiterer Forschung mit anderen Formen von Missspezifikationen im Strukturmodell sowie auch unterschiedlichen Graden an Unausgewogenheit der Aufteilung der Indikatoren auf die latenten Variablen.

3.3 Limitationen und Implikationen

Die Befunde aus dem Design dieser Simulationsstudie lassen vermuten, dass insbesondere die Fit-Indizes RMSEA und SRMR nicht geeignet dafür sind, Missspezifikationen in Form von Dimensionalitätsverletzungen in der Faktorenstruktur zu entdecken, sofern die Cut-Off-Werte nach Hu und Bentler (1998, 1999) das Kriterium darstellen. Allerdings sollte diese Vermutung umfassender untersucht werden, um die Befunde bestätigen zu können, zumal, wie bereits unter III. 2.3 beschrieben, Fehlspezifikationen im Strukturmodell und deren Auswirkungen auf die Fit-Indizes bisher noch wenig untersucht wurden.

Wie bereits erwähnt, wurde die Stichprobengröße in dieser Studie konstant gehalten, um die Missspezifikation im Strukturmodell möglichst spezifisch, ohne zu viele Mehrfach-Interaktionen mit anderen Modellparametern, zu untersuchen. Vermutlich schneiden die Fit-Indizes bei der Entdeckung des Misfit bei einer größeren Stichprobe besser ab. Eine größere Stichprobe ist allerdings im Kontext der angewandten Forschung innerhalb der Psychologie kaum realistisch.

Im Rahmen dieser Studie fiel der absolute Wert des RMSEA geringer aus als in derselben Missspezifikationsbedingung bei Savalei, ähnlich verhielten sich SRMR und CFI im Vergleich zu Mahlers (2011) Studie. Wie bereits erwähnt wurde, liegt die Vermutung nahe, dass die Heterogenität der Faktorladungen im Rahmen dieser Simulationsstudie, die unterschiedliche Anzahl an Items und möglicherweise aber auch Unterschiede in der Stichprobengröße[34] verantwortlich dafür waren, dass die Modellpassung schlechter ausfiel als bei Savalei und Mahler. Daher sollten im Kontext weiterer Studien homogene und heterogene Ladungen sowohl im Rahmen korrekter Modelle verglichen werden, als auch im Rahmen von Missspezifikationen im Strukturmodell und im Messmodell hinsichtlich ihrer Fähigkeit, die Modellabweichung zu entdecken, untersucht werden.

Außerdem ist im Rahmen weiterer Simulationsstudien zu Missspezifikationen im Strukturmodell zu untersuchen, ob der CFI tatsächlich auch abhängig von der Art der Missspezifikation und der (Non-)Zentralität deren Teststatistik ist, wie von Curran et al. (2002) vorgeschlagen wurde und wie in der vorliegenden Studie und der Studie von Mahler (2011) in

[34]Mahler (2011) und Savalei (2012) nannten ihre Stichprobengrößen nicht.

Kontrast zu den Befunden von Heene et al. (2011) zur Höhe der Faktorladungen angenommen wurde.

Im Rahmen dieses Designs stellte die Höhe der Faktorkorrelation in den Populationsmodellen einen Parameter dar, der neben der Indikatorenaufteilung den Schweregrad der Missspezifikation determinierte. Dieser Parameter interagierte hinsichtlich seines Einflusses auf die Fit-Indizes mit der Höhe der Faktorladungen, einem inzidentellen Parameter im Rahmen dieser Studie. Es liegt daher nahe, den Einfluss der Faktorkorrelationshöhe und der Faktorladungshöhe danach getrennt hinsichtlich ihrer Auswirkungen auf die Fit-Indizes zu untersuchen, ob sie beide inzidentelle Parameter darstellen oder beide für den Grad der Missspezifikation verantwortlich sind.

Ebenso wurden im Rahmen dieses Designs die Nebenladungen auf Null gesetzt. Dies stellt allerdings in der angewandten Forschung einen sehr seltenen Fall dar (Beauducel & Wittmann, 2005). Nebenladungen, die fälschlicherweise auf Null gesetzt wurden, sollten daher als weitere inzidentelle Bedingung im Rahmen dieses Designs, oder auch als Bedingung, die u.a. den Grad der Missspezifikation ausmacht, in ihren Auswirkungen auf die Fit-Indizes untersucht werden.

Weiters sind mehr und verschiedenere Missspezifikationen im Strukturmodell zu untersuchen, da Missspezifikationen im Strukturmodell, insbesondere Missspezifikationen hinsichtlich der Dimensionalität der Faktorenstruktur, bisher noch kaum untersucht wurden. Es bietet sich an, weitere Fit-Indizes im Rahmen des umfassenden Designs von Savalei (2012) zu untersuchen: Eine Missspezifikation in Form von statistischer Eindimensionalität, wenn im Populationsmodell oblique oder orthogonale Faktorenstrukturen vorliegen, sowie Populationsmodelle mit mehr als zwei latenten Variablen. Allerdings sollte das Design von Savalei, wie in der vorliegenden Studie, anhand von realistisch hohen und realistisch heterogenen Faktorladungen bezüglich der Fit-Indizes ausgetestet werden.

Die oben aufgeworfene Frage nach den Nebenladungen führt zu einer weiteren Implikation für künftige Simulationsstudien, die die Auswirkungen von Missspezifikationen auf die Fit-Indizes untersuchen. Es ist zu vermuten, dass in der angewandten Forschung nicht nur entweder Missspezifikationen im Messmodell oder im Strukturmodell, sondern beide Arten von Missspezifikationen gleichzeitig auftreten. Die Forschung bisher beschränkte sich vermutlich aus Gründen der Handhabbarkeit entweder auf Missspezifikationen im Mess-, oder im Strukturmodell. Die Vermutung liegt nahe, dass miteinander einhergehende Modellverletzungen im Mess- und Strukturmodell die Modellevaluation anhand der Fit-Indizes

noch zusätzlich erschweren. Insbesondere für den RMSEA sollte ein derartiges Design herausfordernd sein, da dieser Fit-Index zusätzlich noch von der Anzahl der Indikatoren abhängt (Kenny & McCoach, 2009; Savalei, 2012). Savalei (2012) schlussfolgerte dazu aus ihren Studien, dass drei Indikatoren pro latenter Variable dazu führen, dass die Sensitivität des RMSEA zur Entdeckung von Modellabweichungen in Form von Messfehlerkovarianzen innerhalb eines Messmodells minimal wäre, wohingegen die Sensitivität zur Entdeckung von Missspezifikationen im Strukturmodell in Form nicht-spezifizierter Faktorkorrelationen maximiert wäre.

3.4 Ausblick auf die zweite Studie

Im Rahmen der ersten Simulationsstudie wurden die Auswirkungen von Missspezifikationen auf die Güte der Modellpassung untersucht. Die Ergebnisse implizieren, dass SRMR und RMSEA im Kontext des verwendeten Forschungsdesigns ungeeignet dafür waren, Missspezifikationen auf der Ebene der latenten Variablen zu erkennen. Die Befunde favorisieren zwar den CFI zur Beurteilung des Modellfits, sofern das zu evaluierende Modell ein oder mehrere Strukturmodelle enthält, doch erkannte der CFI geringe Missspezifikationen in Form von hohen Faktorkorrelationen im Populationsmodell im Rahmen des untersuchten Designs auch nicht.

Vor dem Hintergrund dieser Befunde stellt sich nun die Frage, inwiefern sich Missspezifikationen im Strukturmodell, welche insbesondere von den absoluten Fit-Indizes im Gegensatz zum CFI anhand der gängigen Cut-Offs nach Hu und Bentler (1998, 1999) nicht erkannt wurden, auf die Validität eines Modells – genauer gesagt auf die Validität der diagnostischen Entscheidungen aus den Faktorwerten – auswirken. Diese Fragestellung wurde im Rahmen einer zweiten Simulationsstudie untersucht, welche im folgenden Kapitel V beschrieben wird. Dazu wurden diagnostische Entscheidungen basierend auf korrelierten Populationsfaktorwerten mit diagnostischen Entscheidungen basierend auf den geschätzten Faktorwerten aus missspezifizierten einfaktoriellen Modellen verglichen.

V STUDIE 2

1 Methode

1.1 Populationsgenerierung

Die unter III. 5.1 ausgeführte Fragestellung lautete, inwiefern sich Modellabweichungen, wie sie bereits im Rahmen von Studie 1 unter IV spezifiziert wurden, auf diagnostische Entscheidungen basierend auf den Faktorwerten auswirken würden. Diese Fragestellung wurde ebenfalls anhand einer Simulationsstudie untersucht. Es wurden die „R"-Pakete (R Core Team, 2015) „lavaan" (Rosseel, 2012), „MASS" (Venables & Ripley, 2002) sowie „psych" (Revelle, 2015) verwendet. Die Simulationsstudie stellte eine Populationssimulation dar. Bei einer Populationssimulation (Kaplan, 1988) werden keine Stichproben aus dem Populationsmodell gezogen, wie es in einer Monte-Carlo-Simulation üblich ist (vgl. Paxton, Curran, Bollen, Kirby, & Chen, 2001), sondern eine Population definiert. Für jede der Bedingungen, welche im Folgenden näher beschrieben werden, wurde eine Population mit 1 Million Faktorwerten (True Scores; Eid et al., 2013, S. 818) generiert. Aus Replikationsgründen wurden Startwerte für die Generierung der Zufallszahlen gesetzt.

1.2 Design

Der Vergleich beruhte auf dichotomen diagnostischen Entscheidungen, die auf der Basis wahrer bzw. definierter Faktorwerte erstellt wurden und dichotomen Klassifikationen, die auf Basis der Faktorwerte aus den zu untersuchenden Modellen berechnet wurden. Das Design für die Populationsmodelle wurde aus Studie 1 (siehe Kapitel IV) übernommen: Es handelte sich wiederum um ein obliques Zwei-Faktoren-Populationsmodell, bei dem der Grad an Missspezifikation durch die Höhe der Faktorkorrelation und die Aufteilung der Indikatoren auf die beiden latenten Variablen im Populationsmodell bestimmt war. Die beiden Faktoren korrelierten entweder zu .30, .50 oder .80 und entweder wurden beide Faktoren von jeweils 10 Indikatoren gemessen oder der erste Faktor von 15 Indikatoren und der zweite von 5. Der entscheidende Unterschied zur Methode der Simulation im Rahmen der Studie 1 lag darin, dass die Datengenerierungen nicht aus einem Populations- oder wahren Strukturgleichungsmodell

heraus erfolgten, sondern basierend auf definierten bzw. wahren Faktorwerten[35]. Auf Basis dieser definierten Faktorwerte wurden die Strukturgleichungsmodelle aufgebaut[36]. In einem ersten Schritt wurden daher bivariat verteilte wahre Faktorwerte für zwei latente Variablen generiert, die entweder zu .30, zu .50 oder zu .80 korrelierten. Außerdem wurden 20 standardnormalverteilte unique Faktorwerte erzeugt, die weder untereinander korrelierten, noch mit den Faktorwerten korrelierten. Weiters wurden die Faktorladungen aus den unter II. 1 und 3 sowie III 5.1 genannten Gründen zufällig aus dem Bereich [.20, .40] oder aus dem Bereich [.40, .60] gleichverteilt gezogen und somit festgelegt. Dadurch konnte die Koeffizientenmatrix der uniquen Faktorwerte errechnet werden. In einem zweiten Schritt konnten dann anhand der Fundamentalgleichung der Faktorenanalyse (siehe (1) unter II. 1; Mulaik, 2009) aus den wahren Faktorwerten, deren Mustermatrix sowie den uniquen Faktorwerten und deren Koeffizientenmatrix die beobachteten Werte für die 1 Million Individuen erstellt werden. Die generierten Populationsmodelle stellten True-Score-Modelle dar (vgl. Eid, Gollwitzer, & Schmitt, 2013, S. 856).

1.3 Durchführung

Zur Beantwortung der ersten Fragestellung wurden fälschlicherweise als einfaktoriell spezifizierte Messmodelle mit 20[37] Indikatoren auf die Daten angewandt, die basierend auf den wahren Faktorwerten erzeugt wurden. Die Varianzen der latenten Variablen der untersuchten Modelle wurden auf Eins gesetzt und die Faktorladungen frei geschätzt. Die Schätzung der Modellpassung erfolgte wiederum mit dem Maximum-Likelihood-Algorithmus (für eine Begründung siehe IV. 1.3). Die Schätzung der Faktorwerte, welche sich bei Anwendung der korrekten zweifaktoriellen Modelle und des missspezifizierten einfaktoriellen Modells auf die Daten ergaben, erfolgte nach der Methode von Bartlett (1937; für die Begründung siehe Kapitel II. 3).

[35]Diese Methode wurde erstmals von Grice und Harris (1998) sowie Grice (2001a, 2001b) beschrieben.

[36]An dieser Stelle sei sehr herzlich Herrn Prof. Dr. André Beauducel sowohl für die Idee zu dieser Art der Datengenerierung als auch für die Hilfestellung beim Aufbau der entsprechenden Funktion zur Datengenerierung gedankt.

[37]Wie bereits unter IV erwähnt, stellt dies eine typische Fragebogenlänge dar (vgl. Peterson, 2000; Shrout & Yager, 1989).

Die dichotome Top-Down-Klassifizierung (Gatewood et al., 2016, S. 662; „Störung liegt vor“ versus „Störung liegt nicht vor“) erfolgte sowohl auf Basis der wahren Faktorwerte, als auch auf Basis der aus dem missspezifizierten Modell errechneten Bartlett-Faktorwerte nach den höchsten Faktorwerten. Es wurden unterschiedliche Basisraten für die diagnostischen Entscheidungen nach den höchsten Faktorwerten berücksichtigt. Die klinischen Basisraten wurden an den 12-Monatsprävalenzen psychischer Störungen in der Europäischen Union orientiert (Wittchen et al., 2011). Die Basisrate gibt den Prozentsatz der Fälle an der erzeugten Gesamtpopulation an, die eine positive Diagnose („Störung liegt vor“) bekamen. Die kleinste für die Simulationsstudie verwendete Basisrate von 2.5% wurde an der Prävalenz der Posttraumatischen Belastungsstörung, der sozialen Phobie oder der generalisierten Angststörung festgemacht (S. 656). Die 5%-Basisrate deckt Suchterkrankungen oder auch das Aufmerksamkeits-Defizit-Hyperaktivitäts-Syndrom bei Kindern und Jugendlichen ab (S. 664). Die nächstgrößere verwendete Basisrate von 7.5% wurde anhand der Prävalenzen für Insomnie oder Major Depression festgelegt (S. 666). Einer Basisrate von 10% entspricht einer 12-Monatsprävalenz für eine depressive Störung (S. 666). Aus Vergleichsgründen wurden zu den kleinen Basisraten in Größenordnungen für die Diagnose einer einzelnen Störung zusätzlich noch größere Basisraten von 30%, 50% und 70% in die Studie miteinbezogen. Einer Prävalenz von 30% entspricht eine Diagnose aus den Bereichen der Suchterkrankungen, der Angst- und Belastungsstörungen, der affektiven Störungen und der somatoformen Störungen im Gesamten (S. 666). Gleichzeitig stellt eine Basisrate von 30% auch die Lebenszeitprävalenz für eine Angststörung dar (Meyer, Rumpf, Hapke, Dilling, & John, 2000, S. 537); ebenso die 12-Monats-Prävalenz für zwei oder mehr psychische Störungen (Wittchen & Jacobi, 2001, S. 999). Die Lebenszeitprävalenz für irgendeine psychische Störung beträgt nahezu 50% (Meyer et al., 2000, S. 540). Außerdem entsprechen die größeren verwendeten Basisraten Grundquoten, wie sie in der Eignungsdiagnostik vorkommen (Schuler, 2014, S. 359). Eine positive Diagnose („krank“) wurde gestellt, wenn ein Individuum auf beiden wahren Faktoren (konjunktive Entscheidungsstrategie; Amelang & Schmidt-Atzert, 2006, S. 399) unter den entsprechenden höchsten Perzentilen (97.5%, 95%, 92.5%, 90%, 70%, 50%, 30%) rangierte. Um diese bivariaten Basisraten trotz der unterschiedlichen Faktorkorrelationen konstant zu halten, da mit der Faktorkorrelation auch die Anzahl der Fälle stieg, die auf beiden Faktoren hohe Werte erzielten (vgl. Gardner & Neufeld, 2013), wurden unterschiedliche univariate Cut-Offs

ermittelt[38]. Für die Vergabe der Diagnosen basierend auf den Faktorwerten des eindimensionalen missspezifizierten Modells wurde der univariate Cut-Off bei den höchsten 97.5%, 95%, 92.5%, 90%, 70%, 50% und 30% der Bartlett-Faktorwerte angesetzt. Analog dazu wurde der Cut-Off für die Vergabe der Diagnosen basierend auf dem Gesamtsummenwert gesetzt.

Für die Beurteilung der Güte der Klassifikation wurden Sensitivität, Spezifität sowie Positiver und Negativer Prädiktionswert berechnet, da diese Kennwerte im Rahmen der Psychometrie weit verbreitete Größen zur Evaluation der Güte der Diagnostik darstellen (Amelang & Schmidt-Atzert, 2006) und sich gegenseitig komplementieren. Diese ließen sich aus den diagnostischen Konsistenzen (True Positives, True Negatives, False Positives und False Negatives) berechnen, die zur Erklärung der Befunde zu den diagnostischen Kennwerten herangezogen wurden.

Weiters wurden für einen ordinalen Vergleich der wahren Faktorwerte mit den Faktorwerten aus dem missspezifizierten Modell deren Korrelationen gebildet, indem die Faktorwerte aus dem missspezifizierten Modell einerseits mit den wahren Faktorwerten des ersten Faktors, andererseits mit den wahren Faktorwerten des zweiten Faktors korreliert wurden. Diese Korrelationen stellen nach Grice (2001a, 2001b) Maße für die Validität der geschätzten Faktorwerte dar.

Zur Beantwortung der Nebenfragestellung wurden in den 12 verschiedenen Bedingungen Gesamtsummenwerte über die 20 erzeugten beobachteten Variablen hinweg gebildet. Die Diagnosevergabe erfolgte bei den Gesamtsummenwerten wie bei den missspezifizierten Modellen nach den höchsten Werten basierend auf den genannten Basisraten. Die Diagnostik auf Basis der Gesamtsummenwerte wurde ebenso anhand der genannten diagnostischen Kennwerte mit der Diagnostik basierend auf den wahren Faktorwerten verglichen. Außerdem wurden auch die Summenwerte jeweils mit den wahren Faktorwerten auf dem ersten Faktor wie auch mit denen auf dem zweiten Faktor korreliert.

[38]An dieser Stelle sei ganz herzlich Felix Naumann und Florian Pargent für die Hilfe beim Schreiben der entsprechenden R-Funktion gedankt.

2 Ergebnisse

2.1 Sanity Checks

2.1.1 Eigenwerte

Zur Prüfung der Plausibilität der Generierung der beobachteten Daten aus den wahren Faktorwerten heraus wurden zunächst die Eigenwerte der beobachteten Kovarianzmatrix der Populationsdaten inspiziert (siehe Tabelle 8). Die Eigenwerte bestimmen zusammen mit den Eigenvektoren die Berechnung der Faktorladungen eines Modells (Lawley & Maxwell, 1971; Mulaik, 2009); die (Höhe der) Faktorladungen wirken sich wiederum als primärer Einflussfaktor auf die Berechnung der Bartlett-Faktorwerte aus (Erklärung folgt unter 2.2.1).

Die Eigenwerte hingen von der definierten Faktorkorrelation wie auch von der vorgegebenen Aufteilung der Indikatoren auf die beiden Faktoren im Populationsmodell ab. Die Eigenwerte verhielten sich plausibel: Die ersten beiden Eigenwerte variierten mit der Höhe der definierten Faktorkorrelation und mit der vorgegebenen Indikatorenaufteilung auf die beiden Faktoren (siehe Tabelle 8). Die Eigenwerte 3 bis 20 waren über die verschiedenen Bedingungen hinweg gleich und wurden aus diesem Grund nicht in die Tabelle inkludiert. Mit sinkender definierter Faktorkorrelation sank der erste Eigenwert, wohingegen der zweite stieg (siehe Tabelle 8), was zeigt, dass sich die Eigenwerte bei sinkender Faktorkorrelation mehr und mehr anglichen und sich der Zwei-Faktoren-Struktur im Populationsmodell annäherten. Bei der ungleichmäßigen Aufteilung der Indikatoren auf die Faktoren war der erste Eigenwert größer und der zweite Eigenwert kleiner als in den vergleichbaren Bedingungen mit der ausgewogenen Indikatorenaufteilung (siehe Tabelle 8). Dies spiegelt wider, dass die ungleichmäßige Aufteilung der Indikatoren näher am einfaktoriellen Modell lag als die gleichmäßige Aufteilung der Indikatoren auf die Faktoren.

Tabelle 8

Erste zwei Eigenwerte der erzeugten beobachteten Kovarianzmatrix

	Modell 1: .80 [.20, .40] 10:10	Modell 2: .50 [.20, .40] 10:10	Modell 3: .30 [.20, .40] 10:10	Modell 4: .80 [.20, .40] 15:5	Modell 5: .50 [.20, .40] 15:5	Modell 6: .30 [.20, .40] 15:5
1. EW	2.685	2.387	2.188	2.744	2.561	2.470
2. EW	1.095	1.391	1.590	1.038	1.213	1.303
	Modell 7: .80 [.40, .60] 10:10	Modell 8: .50 [.40, .60] 10:10	Modell 9: .30 [.40, .60] 10:10	Modell 10: .80 [.40, .60] 15:5	Modell 11: .50 [.40, .60] 15:5	Modell 12: .30 [.40, .60] 15:5
1. EW	5.475	4.684	4.157	5.624	5.133	4.887
2. EW	1.260	2.048	2.575	1.111	1.596	1.842

Anmerkungen. EW = Eigenwert. Die Zellen der Beschreibung der Populationsmodelle enthalten als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

2.1.2 Modellfit der korrekten Modelle

Aus Vergleichsgründen wurden die diagnostischen Entscheidungen auf Basis der wahren Faktorwerte zunächst mit den diagnostischen Entscheidungen verglichen, die auf Basis der Bartlett-Faktorwerte aus einem korrekten Modell getroffen wurden. Daher wird an dieser Stelle die Modellpassung der korrekten Modelle berichtet. Tabelle 9 gibt den Modellfit korrekter Modelle wider, die auf die erzeugten Populationsdaten angewandt wurden, welche basierend auf den wahren Faktorwerten generiert wurden. Alle korrekten Modelle zeigten sehr gute Modellpassung auf die erzeugten Populationsdaten.

Tabelle 9

Modellpassung der korrekten Modelle

	Modell 1: .80 [.20, .40] 10:10	Modell 2: .50 [.20, .40] 10:10	Modell 3: .30 [.20, .40] 10:10	Modell 4: .80 [.20, .40] 15:5	Modell 5: .50 [.20, .40] 15:5	Modell 6: .30 [.20, .40] 15:5
$\chi^2(169)$	172.746	168.343	166.434	178.452	179.147	179.839
p	.406	.500	.501	.294	.282	.270
CFI	1.000	1.000	1.000	1.000	1.000	1.000
RMSEA	< .001	< .001	< .001	< .001	< .001	< .001
SRMR	.001	.001	.001	.001	.001	.001
	Modell 7: .80 [.40, .60] 10:10	Modell 8: .50 [.40, .60] 10:10	Modell 9: .30 [.40, .60] 10:10	Modell 10: .80 [.40, .60] 15:5	Modell 11: .50 [.40, .60] 15:5	Modell 12: .30 [.40, .60] 15:5
$\chi^2(169)$	168.686	164.904	164.722	177.066	176.824	177.873
p	.492	.575	.579	.320	.324	.305
CFI	1.000	1.000	1.000	1.000	1.000	1.000
RMSEA	< .001	< .001	< .001	< .001	< .001	< .001
SRMR	.001	.001	.001	.001	.001	.001

Anmerkungen. Die Zellen der Beschreibung der Populationsmodelle als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

2.1.3 Modellfit der missspezifizierten Modelle

Tabelle 10 zeigt die Parameter der Modellpassung für die missspezifizierten einfaktoriellen Modelle, die auf die erzeugten Daten angewandt wurden, welche basierend auf den wahren Faktorwerten der beiden obliquen Faktoren generiert wurden. Der χ^2-Test zeigte keine Modellpassung an und der χ^2-Wert erzielte in Relation zu der Anzahl der Freiheitsgrade von 170 sehr hohe Werte, was eine hohe Modellabweichung indiziert. Die Fit-Indizes zeigten ein ähnliches Muster wie in Studie 1 unter VI, wobei insbesondere RMSEA und SRMR in vielen Fällen Modellpassung anzeigten.

Wie in der ersten Studie (siehe Kapitel IV) konnte der χ^2-Wert Auskunft über den Grad der Missspezifikation geben. Mit zunehmender Missspezifikation (sinkende Faktorkorrelation im Populationsmodell) stieg der χ^2-Wert. Ebenso war der χ^2-Wert bei der ungleichen Indikatorenaufteilung auf die beiden Faktoren im Populationsmodell (geringerer Grad an

Missspezifikation) niedriger als bei der ausgewogenen Indikatorenaufteilung. Dementsprechend verhielt sich auch der Nonzentralitätsparameter (Differenz aus χ^2-Wert und Freiheitsgraden des Modells; für beide Werte im Einzelnen siehe Tabelle 10): Der Nonzentralitätsparameter stieg mit höherem Schweregrad der Missspezifikation, operationalisiert durch geringere Faktorkorrelation und ausgewogene Indikatorenaufteilung auf die Populationsfaktoren.

Tabelle 10

Modellpassung der missspezifizierten Modelle

	Modell 1: .80 [.20, .40] 10:10	Modell 2: .50 [.20, .40] 10:10	Modell 3: .30 [.20, .40] 10:10	Modell 4: .80 [.20, .40] 15:5	Modell 5: .50 [.20, .40] 15:5	Modell 6: .30 [.20, .40] 15:5
$\chi^2(170)$	21193.601	108171.383	189767.004	9181.598	40472.649	60537.001
p	< .001	< .001	< .001	< .001	< .001	< .001
CFI	.974	.842	.703	.989	.946	.915
RMSEA	.011	.025	.033	.007	.015	.019
SRMR	.010	.024	.033	.006	.014	.018
	Modell 7: .80 [.40, .60] 10:10	Modell 8: .50 [.40, .60] 10:10	Modell 9: .30 [.40, .60] 10:10	Modell 10: .80 [.40, .60] 15:5	Modell 11: .50 [.40, .60] 15:5	Modell 12: .30 [.40, .60] 15:5
$\chi^2(170)$	170692.725	707129.913	1127164.083	86337.547	292464.751	404395.746
p	< .001	< .001	< .001	< .001	< .001	< .001
CFI	.953	.785	.645	.978	.915	.879
RMSEA	.032	.064	.081	.022	.041	.049
SRMR	.025	.063	.088	.017	.038	.049

Anmerkungen. Die Zellen der Beschreibung der Populationsmodelle enthalten als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

2.1.4 Faktorladungen

Die standardisierten Faktorladungen der korrekten Modelle bewegten sich innerhalb der Spannweiten, wie sie in den Populationsmodellen definiert wurden (Bedingungen mit den typischen Ladungen: $M = .310$, $SD = .063$; Bedingungen mit den hohen Ladungen: $M = .520$, $SD = .063$).

Die Faktorladungen der missspezifizierten einfaktoriellen Modelle waren geringer als die in den Populationsmodellen definierten Faktorladungen. Die Mittelwerte der Ladungen der missspezifizierten Modelle in den Bedingungen mit den typischen Ladungen in den Populationsmodellen bewegten sich zwischen M = .246 und M = .293 (SD = .050 bis SD = .103), wohingegen sich die Mittelwerte der Bedingungen mit den hohen Ladungen in den Populationsmodellen zwischen M = .404 und M = .490 (SD = .053 bis SD = .157) bewegten. Außerdem zeigte sich, dass die Ladungen in den missspezifizierten Modellen mit sinkender Faktorkorrelation in den Populationsmodellen in zunehmendem Maße geringer wurden (Reduktion der standardisierten Faktorladungen im missspezifizierten Modell von .02 bis .11 bei sinkender Faktorkorrelation).

2.2 Faktorwerte korrekter Modelle

2.2.1 Korrelationen

Bevor aus den Bartlett-Faktorwerten korrekter Modelle Diagnosen gebildet wurden und diese mit den Diagnosen aus den wahren Faktorwerten verglichen wurden, wurde zunächst die Validität (für den Begriff vgl. Grice [2001a, 2001b]) der Faktorwerteschätzung aus den korrekten Modellen bestimmt, auf der wiederum die Validität der diagnostischen Entscheidungen aus den Faktorwerten basiert. Dazu wurden die wahren Faktorwerte der beiden definierten obliquen Faktoren mit den Bartlett-Faktorwerten der beiden spezifizierten Faktoren aus korrekten Modellen korreliert.

Tabelle 11 zeigt, dass die Höhe der Ladungen in den Populationsmodellen einen Einfluss auf die Korrelationen der Faktorwerte hatte. Hohe Faktorladungen in den Populationsmodellen führten zu höheren Korrelationen der True Scores und der Bartlett-Faktorwerte aus den korrekten Modellen. Die unausgewogene Indikatorenaufteilung führte zu geringfügig höheren Korrelationen (Unterschied auf der zweiten Dezimalstelle, siehe Tabelle 11) der wahren und der aus dem korrekten Modell geschätzten Faktorwerte der ersten Faktoren (überrepräsentiert durch 15 Indikatoren) und niedrigeren Korrelationen der Faktorwerte der zweiten Faktoren (5 Indikatoren) im Vergleich zur gleichmäßigen Indikatorenaufteilung auf die Faktoren (siehe Tabelle 11). Für letzteren Befund liegt folgende Erklärung nahe: Estabrook und Neale (2013) zeigten anhand einer Simulationsstudie mit einem mehrfaktoriellen Modell, dass die geschätzten Faktorwerte bei einer höheren Anzahl an Indikatoren näher an den wahren

Faktorwerten lagen als bei einer geringen Anzahl an Indikatoren pro Faktor. Lawley und Maxwell (1971) zeigten außerdem, dass sich die Verteilungseigenschaften der geschätzten Faktorwerte mehr denen der wahren Faktorwerte annähern, wenn mehr Indikatoren verwendet wurden.

Tabelle 11

Korrelationen der wahren Faktorwerte und der Faktorwerte aus einem korrekten Modell

	Modell 1: .80 [.20; .40] 10:10	Modell 2: .50 [.20; .40] 10:10	Modell 3: .30 [.20; .40] 10:10	Modell 4: .80 [.20; .40] 15:5	Modell 5: .50 [.20; .40] 15:5	Modell 6: .30 [.20; .40] 15:5
$r_{W1\text{-}K1}$	$.733^{***}$	$.733^{***}$	$.733^{***}$	$.795^{***}$	$.795^{***}$	$.795^{***}$
$r_{W2\text{-}K2}$	$.723^{***}$	$.724^{***}$	$.724^{***}$	$.591^{***}$	$.591^{***}$	$.591^{***}$
	Modell 7: .80 [.40; .60] 10:10	Modell 8: .50 [.40; .60] 10:10	Modell 9: .30 [.40; .60] 10:10	Modell 10: .80 [.40; .60] 15:5	Modell 11: .50 [.40; .60] 15:5	Modell 12: .30 [.40; .60] 15:5
$r_{W1\text{-}K1}$	$.889^{***}$	$.889^{***}$	$.889^{***}$	$.921^{***}$	$.921^{***}$	$.921^{***}$
$r_{W2\text{-}K2}$	$.885^{***}$	$.885^{***}$	$.885^{***}$	$.800^{***}$	$.800^{***}$	$.800^{***}$

Anmerkungen. $r_{W1\text{-}K1}$ = Korrelation der wahren Faktorwerte des ersten definierten Faktors mit den Bartlett-Faktorwerten des ersten Faktors des korrekten Modells, $r_{W1\text{-}K1}$ = Korrelation der wahren Faktorwerte des zweiten definierten Faktors mit den Bartlett-Faktorwerten des zweiten Faktors des korrekten Modells, *** = höchst signifikanter Zusammenhang. Die Zellen der Beschreibung der Populationsmodelle enthalten als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

Die Höhe der Korrelationen zwischen den wahren Faktorwerten und den Faktorwerten aus den korrekten Modellen war insgesamt unerwartet niedrig, wurden doch korrekte Modelle auf die erzeugten Daten angewandt. Der Grund für dieses Phänomen liegt in den niedrigen definierten Faktorladungen in den Populationsmodellen. Desto niedriger die Faktorladungen eines Modells, desto mehr ist die Berechnung der Faktorwerte fehlerbehaftet. Wie aus Gleichung (10) unter II. 3 (Grice, 2001b, S. 433) ersichtlich, werden für die Berechnung der Bartlett-Faktorwerte (wie auch der anderen Faktorwerte) die Faktorladungen benötigt; die Berechnung der Bartlett-Faktorwerte erfolgt durch Minimierung der Off-Diagonal-Elemente der Uniqueness-Matrix (DiStefano et al., 2009). Die uniquen Faktorwerte sind proportional zu

den uniquen Faktorladungen[39]. Das heißt, große Einträge in den Off-Diagonal-Elementen der Uniqueness-Matrix senken die Validität (Grice, 2001a, 2001b) der berechneten Faktorwerte.

Dass primär die Faktorladungen Einfluss auf die Berechnung der Bartlett-Faktorwerte haben und als Konsequenz niedriger Ladungen die Faktorenunbestimmtheit größer wird, wurde bereits unter II. 3 beschrieben und zeigte sich bereits an der Validität der Faktorwerte (siehe Tabelle 11). Dieser Befund hat wiederum negative Auswirkungen auf die Güte der Diagnostik auf Basis der Bartlett-Faktorwerte, die an späterer Stelle berichtet werden. Daher wird an dieser Stelle zunächst der Grad der Unbestimmtheit anhand der Formel von Guttman (1955, S. 73; siehe Gl. (11) unter II. 3) berichtet.

Dem Befund zu den Korrelationen der True Scores und der Bartlett-Faktorwerte entsprechend war die Unbestimmtheit der Faktorwerte, operationalisiert durch ρ^2 (siehe Gl. (11) unter II. 3; Guttman, 1955, S. 73), in den Bedingungen mit den typischen Ladungen höher als in den Bedingungen mit den hohen Ladungen (siehe Tabelle 12). In allen Bedingungen mit den typischen Ladungen im Populationsmodell war der Grad an Faktorenbestimmtheit kleiner als 50% (Quadratwurzel aus ρ^2; siehe Tabelle 12). Ein Grad an Faktorenunbestimmtheit von unter 70% bedeutet, dass es Sets aus Faktorwerten gibt, die nicht einmal positiv miteinander korrelieren, d.h., dass die Sets aus Faktorwerten substanziell sogar in die entgegengesetzte Richtung gehen können (Guttman). Außerdem war bei den Bedingungen mit den typischen Ladungen im Populationsmodell die maximal mögliche Unbestimmtheit der Faktorwerte, operationalisiert durch ρ^* (siehe Gl. (11) unter II. 3; Guttman, S. 73), größer als in den Bedingungen mit den hohen Ladungen (siehe Tabelle 12). Dies geht einher mit dem Befund zu den Korrelationen, nach denen die Faktorwerte der Populationsmodelle mit hohen Ladungen höher korrelierten als die der Populationsmodelle mit typischen Ladungen. Bei der unausgewogenen Indikatorenaufteilung ging die Bestimmtheit der Faktoren zwischen den beiden Faktoren weiter auseinander als bei der ausgewogenen Indikatorenaufteilung. Auch dieses Ergebnismuster steht im Einklang zu den Befunden zu den Korrelationen der Faktorwerte.

[39]Diese Proportionalität $\boldsymbol{\Psi \cdot E}$ (siehe Fundamentalgleichung der Faktorenanalyse; Gleichung (1) unter II. 1; Mulaik, 2009) wird an den Bartlett-Faktorwerten ersichtlich, sobald man in die Gleichung zur Berechnung der Matrix der Bartlett-Faktorwerte (siehe Gl. (10) unter II. 3) die Fundamentalgleichung der Faktorenanalyse einsetzt:

$$\mathbf{X = Y \Psi^{-2} \Lambda (\Lambda' \Psi^{-2} \Lambda)^{-1} = (\Lambda X + \Psi E) (\Psi E)^{-1} \Lambda (\Lambda' (\Psi E)^{-1} \Lambda)^{-1}}$$

Tabelle 12

Unbestimmtheit der Faktorwerte im korrekten Modell

	Modell 1: .80 [.20, .40] 10:10	Modell 2: .50 [.20, .40] 10:10	Modell 3: .30 [.20, .40] 10:10	Modell 4: .80 [.20, .40] 15:5	Modell 5: .50 [.20, .40] 15:5	Modell 6: .30 [.20, .40] 15:5
ρ^2_{1F}	.624***	.568***	.548***	.667***	.645***	.637***
ρ^2_{2F}	.618***	.556***	.534***	.549***	.420***	.374***
ρ^*_{1F}	.249***	.136***	.097***	.335***	.290***	.274***
ρ^*_{2F}	.236***	.112***	.069***	.097***	-.160***	-.253***
	Modell 7: .80 [.40, .60] 10:10	Modell 8: .50 [.40, .60] 10:10	Modell 9: .30 [.40, .60] 10:10	Modell 10: .80 [.40, .60] 15:5	Modell 11: .50 [.40, .60] 15:5	Modell 12: .30 [.40, .60] 15:5
ρ^2_{1F}	.827***	.800***	.793***	.863***	.853***	.850***
ρ^2_{2F}	.822***	.793***	.786***	.748***	.671***	.650***
ρ^*_{1F}	.653***	.600***	.587***	.726***	.705***	.700***
ρ^*_{2F}	.644***	.587***	.572***	.495***	.343***	.300***

Anmerkungen. ρ^2 = Maß für die Höhe der Faktorenunbestimmtheit (Guttman, 1955, S. 73), ρ^* = Maß für die maximal mögliche Unbestimmtheit der Faktorwerte (Guttman, S. 73), 1F = 1. Faktor des korrekten Modells, 2F = 2. Faktor des korrekten Modells, *** = höchst signifikanter Zusammenhang. Die Zellen der Beschreibung der Populationsmodelle enthalten als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

Exkurs: Populationsmodelle mit extrem hohen Ladungen

Wurden im Rahmen desselben Designs Populationsmodelle mit Faktorladungen aus dem Bereich [.80, 1.00[[40] zusammengesetzt, beträgt ρ^2 nahezu Eins (Range ρ^2 .993-.996) und die Bartlett-Faktorwerte aus dem korrekt spezifizierten Modell korrelierten mit den wahren Faktorwerten auch zu fast Eins (Range der Korrelationen .994-.997). Das heißt, sehr hohe Ladungen (hohe Reliabilität) führten auch zu einer hohen Validität (für den Begriff vgl. Grice [2001a, 2001b]) der berechneten Faktorwerte.

[40]Für die Berechnung der Bartlett-Faktorwerte wird die Inverse der Uniqueness-Matrix benötigt (Grice, 2001b, S. 433). Daher müssen die Einträge in der Diagonalen dieser Matrix ungleich Null sein. Aus diesem Grund wurde die Zahl 1 für die zufällige Erzeugung der Faktorladungen exkludiert, damit die Matrix nicht-singulär und somit invertierbar ist.

2.2.2 Güte der Diagnostik

Die Diagnosen wurden auf Basis der höchsten wahren Faktorwerte auf beiden Faktoren wie auch basierend auf den Bartlett-Faktorwerten der beiden Faktoren aus den korrekten Modellen gebildet. Es wurden unterschiedliche Basisraten für die Diagnosegebung verwendet (die höchsten 97.5%, 95%, 92.5%, 90%, 70%, 50% und 30% auf beiden Faktoren). Es wurden unterschiedliche univariate Cut-Offs für die Vergabe der Diagnosen auf Basis der True Scores und der Bartlett-Faktorwerte aus den korrekten Modellen verwendet (siehe V. 1.3), die von der Korrelation der Faktoren in den Populationsmodellen abhingen. Für die Evaluation der Güte der dichotomen Klassifikationen wurden Sensitivität, Spezifität sowie Positiver und Negativer Prädiktionswert berechnet.

Basisraten und Faktorladungen

Zunächst ist zu erwähnen, dass die auf Basis der wahren Faktorwerte definierten bivariaten Basisraten (die 2.5%, 5%, 7.5%, 10%, 30%, 50% und 70% Fälle an der Gesamtpopulation mit den höchsten Faktorwerten auf beiden Faktoren) durch Anwendung der univariaten Cut-Offs auf die Bartlett-Faktorwerte, die aus den korrekten Modellen berechnet wurden, nicht reproduziert werden konnten (siehe Abbildungen 1a und b).

Die vier kleinen Basisraten in klinischen Größenordnungen wurden deutlich überschätzt (Überschätzung bis zu knapp 100 Prozentpunkte der definierten kleinsten Basisrate; siehe Abbildung 1a), die drei großen Basisraten wurden deutlich unterschätzt (bis zu 25 Prozentpunkte Unterschätzung der größten Basisrate; siehe Abbildung 1b). Für dieses Phänomen können die realistisch niedrig definierten Ladungen in den Populationsmodellen verantwortlich gemacht werden. Außerdem wurden bei den höheren Ladungen im Rahmen dieses Designs die vier kleinen Basisraten weniger überschätzt als bei den typischen Ladungen und die drei großen Basisraten weniger unterschätzt als bei den typischen Ladungen.

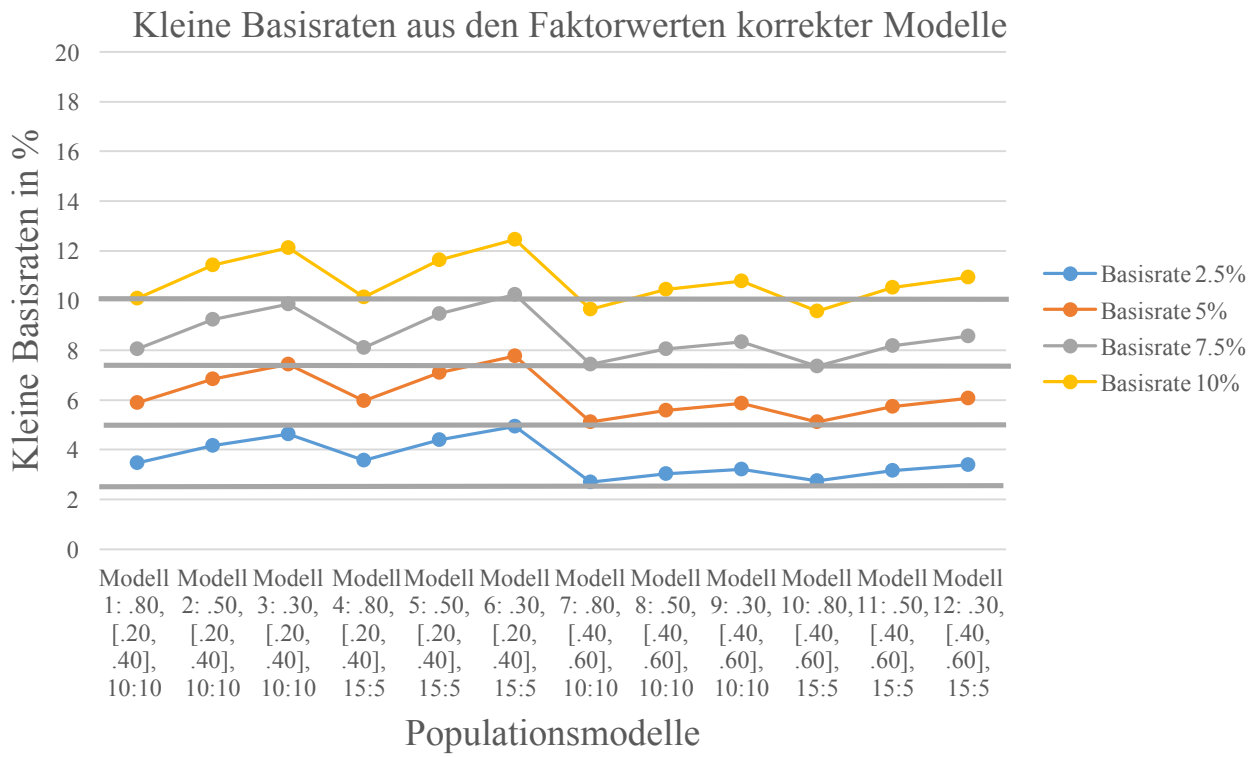

Abbildung 1a. Überschätzung der kleinen Basisraten basierend auf den aus den korrekten Modellen berechneten Bartlett-Faktorwerten, die vier breiten grauen Linien symbolisieren die Basisraten, wie sie in den Populationsmodellen definiert wurden; die Spalten der x-Achse sind mit den Beschreibungen der Populationsmodelle gekennzeichnet

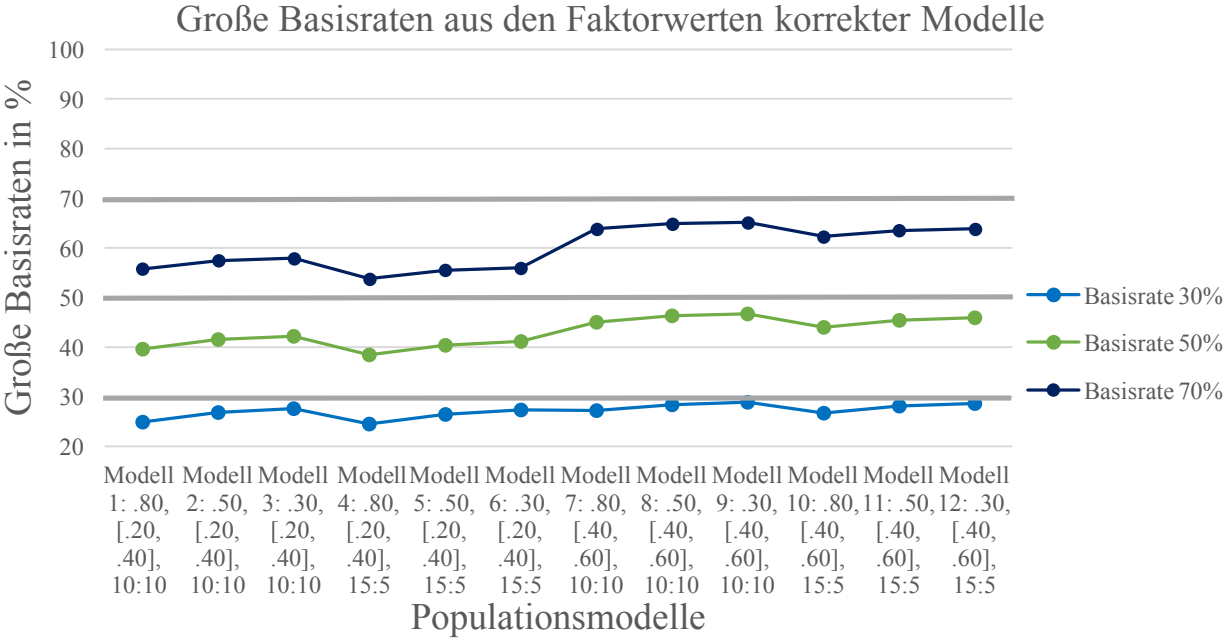

Abbildung 1b. Unterschätzung der großen Basisraten basierend auf den aus den korrekten Modellen berechneten Bartlett-Faktorwerten (für die Legende siehe Abbildung 1a)

Im Folgenden werden die Auswirkungen der über- und unterschätzten Basisraten auf die Güte der Diagnostik aus den Faktorwerten der korrekten Modelle geschildert.

Alle diagnostischen Kennwerte zeigten eine hohe Abhängigkeit von den Basisraten. Kleine Basisraten bzw. Basisraten unter 50% beeinträchtigten vor allem die Rate der korrekt als krank Erkannten (siehe Abbildung 14 im Anhang) im Vergleich zur Rate an korrekt als gesund Erkannten (siehe Abbildung 15 im Anhang). Die überschätzten kleinen und die unterschätzten großen Basisraten wirkten sich an erster Stelle auf den Positiven Prädiktionswert (siehe Abbildung 2) aus, an zweiter Stelle auf die Sensitivität (siehe Abbildung 3).

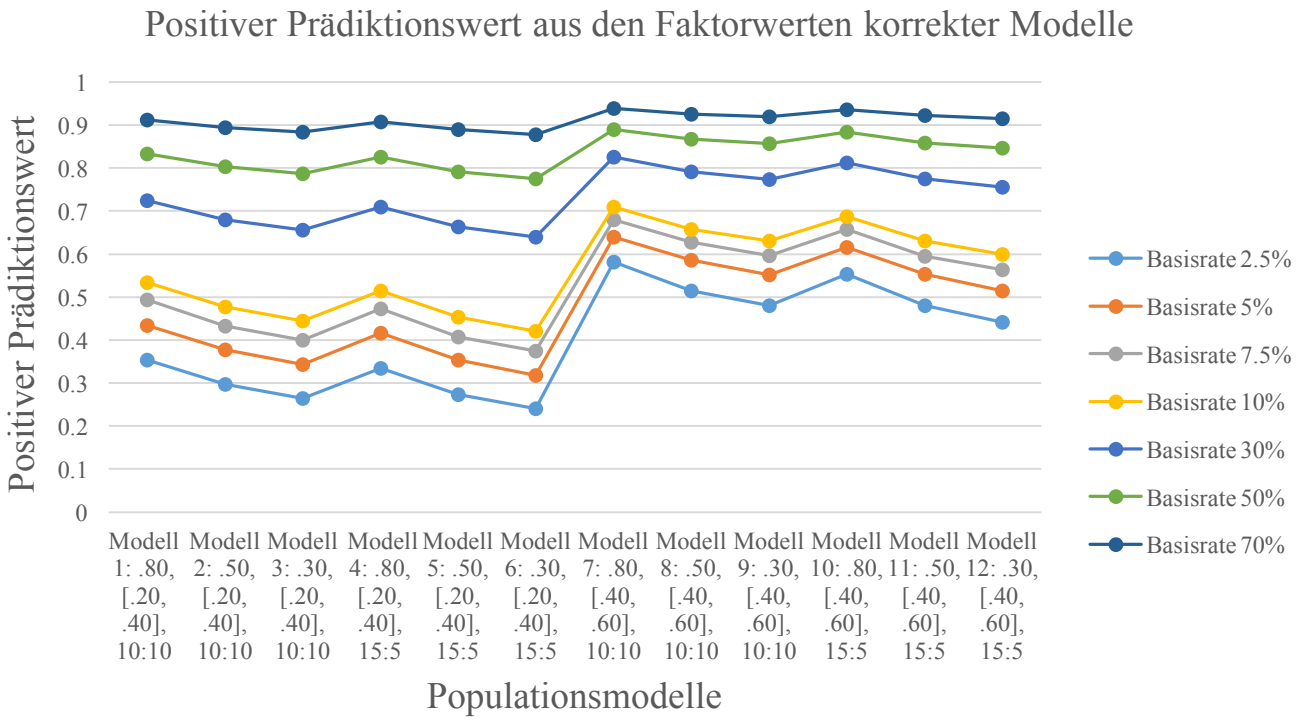

Abbildung 2. Positiver Prädiktionswert der Diagnostik basierend auf den Bartlett-Faktorwerten korrekter Modelle; die Spalten der x-Achse sind im Rahmen dieser und aller folgenden Abbildungen mit den Beschreibungen der Populationsmodelle gekennzeichnet

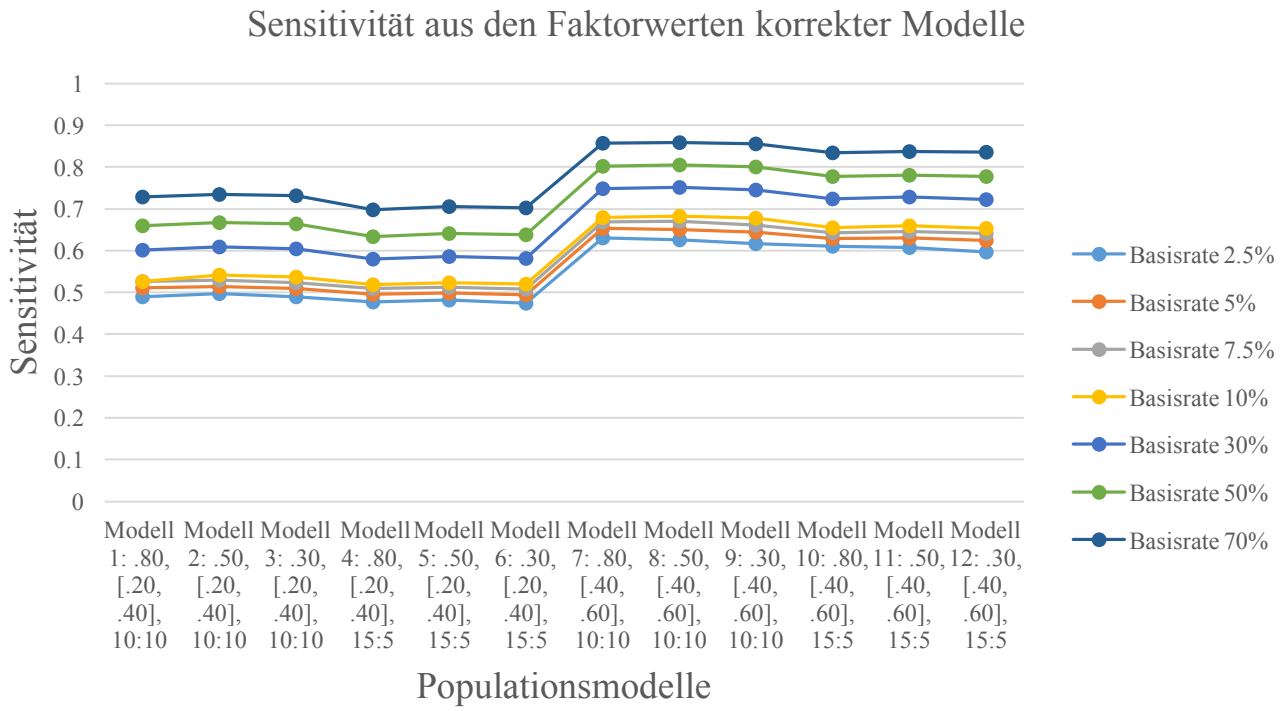

Abbildung 3. Sensitivität der Diagnostik basierend auf den Bartlett-Faktorwerten korrekter Modelle

Insgesamt fiel die Anzahl der False Positives (siehe Abbildung 16 im Anhang) aus den Bartlett-Faktorwerten korrekter Modelle zwischen den verschiedenen Populationsmodellbedingungen heterogener aus als die anderen Kennwerte der diagnostischen Konsistenzen (siehe die Abbildungen 14, 15 und 17 im Anhang), was sich primär auf den Positiven Prädiktionswert (siehe Abbildung 2) auswirkte (Erklärung folgt unter „Exkurs: Populationsmodelle mit extrem hohen Ladungen“ am Ende des Unterkapitels).

Dass der Positive Prädiktionswert (siehe Abbildung 2) bei kleinen Basisraten am niedrigsten von allen diagnostischen Kennwerten ausfiel, ist darauf zurückzuführen, dass bei kleinen Basisraten, die anhand des korrekten Modells überschätzt wurden, einerseits prozentual an der Gesamtpopulation pro Bedingung gesehen weniger Korrekt Positive auftraten als Korrekt Negative, und andererseits, dass bei kleinen Basisraten aufgrund deren Überschätzung sehr viele Falsch Positive auftraten (siehe Abbildung 16 im Anhang) und weniger Falsch Negative (siehe Abbildung 17 im Anhang; Unterschied zwischen Falsch Positiven und Falsch Negativen bis zu einem Prozentpunkt an allen Diagnosen insgesamt auf Basis der Faktorwerte des korrekten Modells). Auf Grund dieser Befunde fiel die Sensitivität (siehe Abbildung 3) bei kleinen Basisraten auch niedrig aus, aber höher als der Positive Prädiktionswert. Dadurch, dass bei großen Basisraten aufgrund deren Unterschätzung mehr Falsch Negative als Falsch Positive

auftraten (Unterschied bis zu 10 Prozentpunkte an allen Diagnosen in der Gesamtpopulation auf Basis der Faktorwerte des korrekten Modells), fiel die Sensitivität bei großen Basisraten niedriger aus als der Positive Prädiktionswert.

Weniger als auf Positiven Prädiktionswert und Sensitivität wirkten sich die überschätzten kleinen und unterschätzten großen Basisraten auf den Negativen Prädiktionswert (siehe Abbildung 4) und am wenigsten auf die Spezifität (siehe Abbildung 5) aus. Durch die Unterschätzung der großen Basisraten gab es bei der Diagnostik durch die Faktorwerte des korrekten Modells bei den großen Basisraten deutlich mehr Falsch Negative (siehe Abbildung 17 im Anhang) als Falsch Positive (siehe Abbildung 16 im Anhang; Unterschied bis zu 10 Prozentpunkte an allen Diagnosen der Gesamtpopulation auf Basis der Faktorwerte des korrekten Modells), sodass sich dieser Umstand vor allem auf den Negativen Prädiktionswert (siehe Abbildung 4) niederschlug und weniger auf die Spezifität (siehe Abbildung 5). Dadurch, dass die Rate der Korrekt Negativen absolut gesehen bei kleinen Basisraten sehr hoch war, fielen Negativer Prädiktionswert und Spezifität bei kleinen Basisraten ähnlich aus, da die kleinen Unterschiede in der Anzahl der Falsch Positiven und der Falsch Negativen hinsichtlich dieser beiden diagnostischen Kennwerte weniger ins Gewicht fielen.

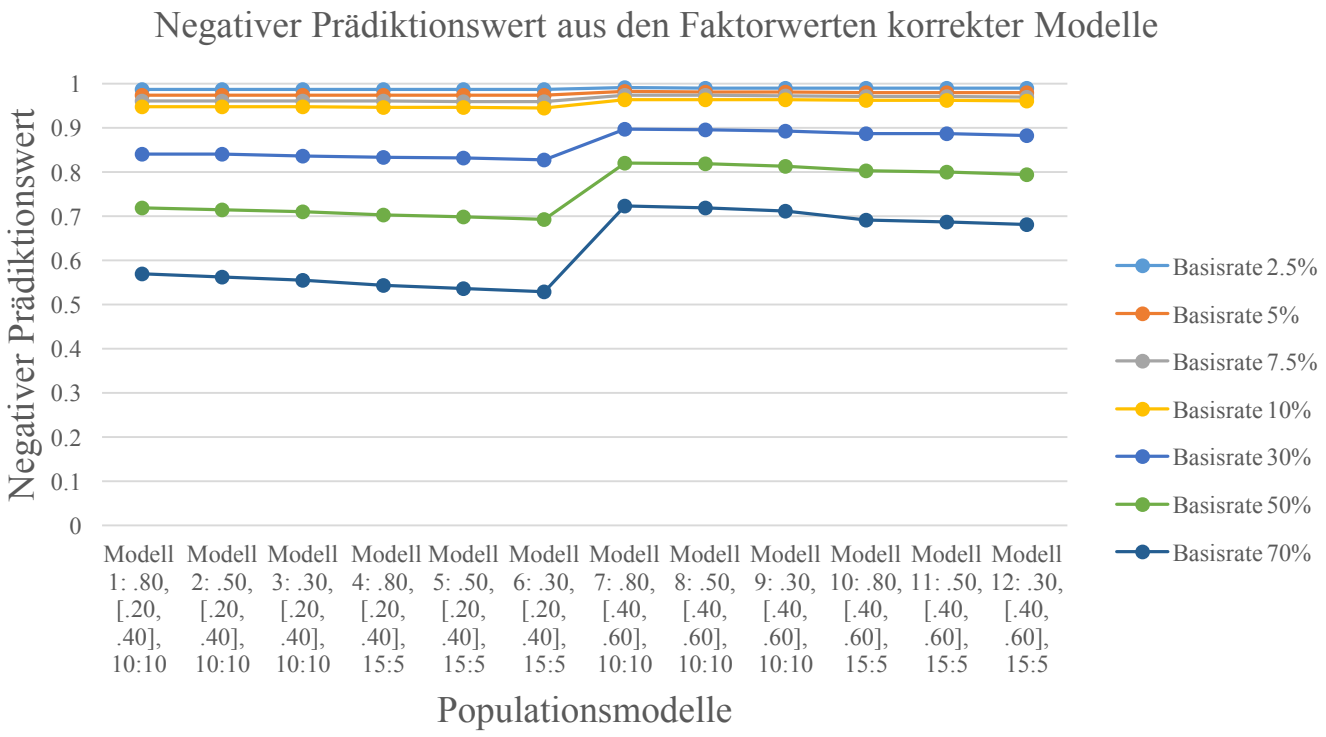

Abbildung 4. Negativer Prädiktionswert der Diagnostik basierend auf den Bartlett-Faktorwerten korrekter Modelle

Über alle diagnostischen Kennwerte hinweg (siehe die Abbildungen 2 bis 5) führte die ausgewogene Indikatorenaufteilung im Populationsmodell im Vergleich zu unausgewogener zu marginal besserer Diagnostik auf Basis der Faktorwerte der korrekten Modelle. Der Befund wird unter 3.2 im Kontext der Befunde zum missspezifizierten Modell diskutiert.

Die Abbildungen 2 bis 5 zeigen die beiden Hauptbefunde hinsichtlich der Diagnostik auf Basis der Bartlett-Faktorwerte der korrekten Modelle: Die Güte der Diagnostik hing vor allem von der Höhe der Basisrate ab, aber auch von der Höhe der Faktorladungen und von der Interaktion aus beidem. Hinsichtlich der Korrektheit eines Krankheitszustands und Korrektheit einer positiven Diagnose[41] zeigten sich kleine Basisraten als problematisch. Sofern die Korrektheit eines Gesundheitszustands und die Korrektheit einer negativen Diagnose[42] Ziel der Diagnostik ist, waren hohe Basisraten für das Vorliegen einer positiven Diagnose ungünstig. Mit höheren Faktorladungen in den Populationsmodellen stieg sowohl die Güte der Diagnostik anhand aller vier Kennwerte als auch wurde der Einfluss der unterschiedlichen Basisraten auf die Güte der Diagnostik mit höheren Ladungen geringer (siehe Modelle 7 bis 12 in den Abbildungen 2 bis 5).

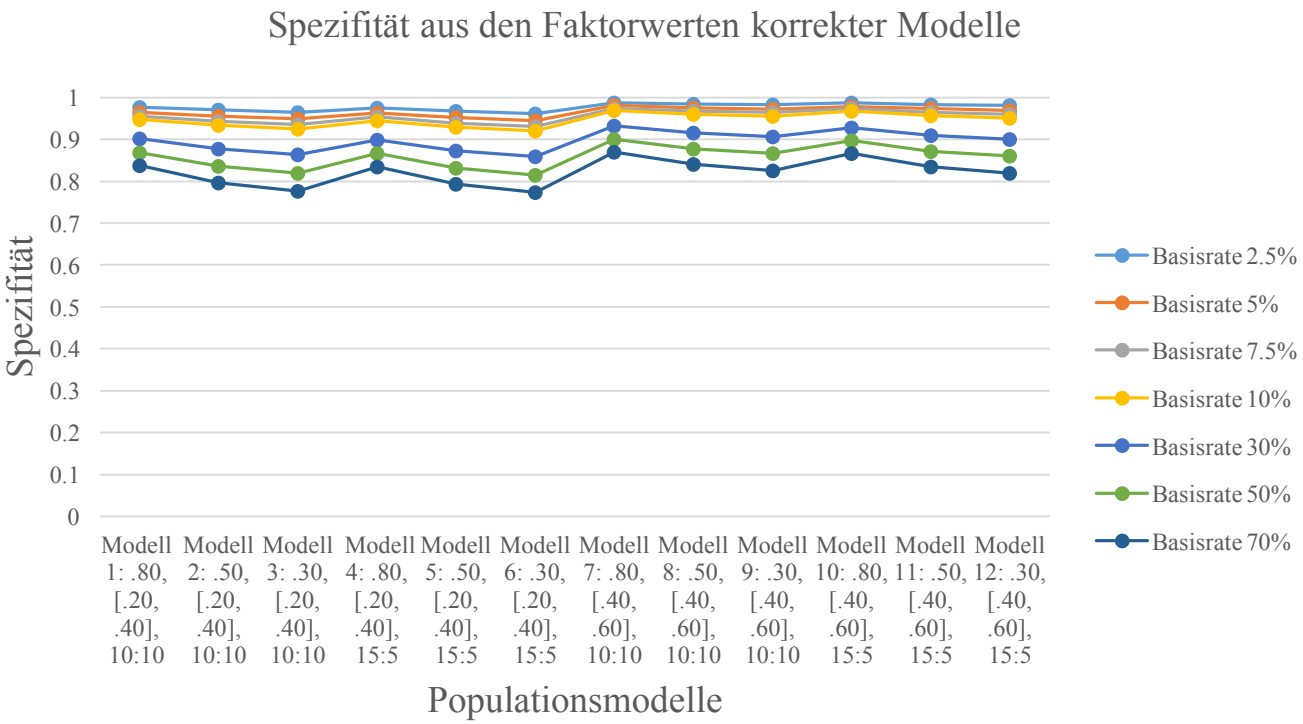

Abbildung 5. Spezifität der Diagnostik basierend auf den Bartlett-Faktorwerten korrekter Modelle

[41]Sensitivität und Positiven Prädiktionswert betreffend

[42]respektive Spezifität und Negativem Prädiktionswert

Exkurs: Populationsmodelle mit extrem hohen Ladungen

Aus Vergleichsgründen wurden Populationsdaten basierend auf Populationsmodellen mit unrealistisch hohen Ladungen aus dem Bereich [.80, 1.00[erzeugt. Diese Populationsmodelle entsprachen mit Ausnahme der Höhe der Ladungen den unter V. 1.2 beschriebenen Modellbedingungen.

Wie bereits unter V. 2.2.1 erwähnt, stieg die Validität der berechneten Faktorwerte aus den korrekten Modellen mit der Höhe der Faktorladungen. In Konsequenz dessen wurden bei Verwendung dieser extrem hohen Ladungen auch die im Rahmen der Populationsmodelle definierten Basisraten reproduziert (Unterschiede zwischen den Basisraten bei den True Scores und den Basisraten aus den Faktorwerten korrekter Modelle unter 0.1 Prozentpunkten).

Die Befunde zur Diagnostik basierend auf den Bartlett-Faktorwerten korrekter Modelle, die auf die Populationsdaten angewandt wurden, die aus Modellen mit extrem hohen Faktorladungen erzeugt wurden, zeigten, dass sich die Verteilungseigenschaften der Faktorwerte aus diesen Modellen denen der True Scores annäherten (*SD* der wahren Faktorwerte 0.999-1.000 versus *SD* der Faktorwerte aus korrekten Modellen 1.003-1.006). Bei den realistisch hoch definierten Ladungen im Rahmen dieses Designs wichen die Verteilungseigenschaften der Faktorwerte aus den korrekten Modellen hingegen stärker von denen der Populationsfaktorwerte ab (*SD* der Faktorwerte aus den korrekten Modellen mit realistisch hohen Ladungen über die Modellbedingungen hinweg 1.218-1.492).

Wurden diese extrem hohen Ladungen zur Generierung der Populationsmodelle verwendet, sanken Sensitivität und Positiver Prädiktionswert selbst bei kleinen Basisraten nicht unter .895. Mit extrem hohen Ladungen schwankte die Anzahl der False Positives weniger mit den Populationsmodellbedingungen (Unterschiede innerhalb einer Bedingung maximal 1 Prozentpunkt über die Basisraten hinweg an allen Diagnosen) und insofern variierte auch der Positive Prädiktionswert weniger (maximaler Range .895-.981 innerhalb einer Bedingung) als bei den Ladungen, die für das vorliegende Design verwendet wurden. Negativer Prädiktionswert und Spezifität sanken bei großen Basisraten nicht unter .951. Das heißt, mit sehr hohen Faktorladungen konnte der Einfluss der Höhe der Basisrate auf die Güte der Diagnostik weitestgehend ausgeglichen werden.

Eine Einschränkung dieser Befunde zu den extrem hohen Ladungen stellt allerdings dar, dass die Modellschätzung mittels Maximum-Likelihood-Algorithmus bei drei der korrekten Modelle nicht konvergierte. Allerdings lagen die frei zu schätzenden Modellparameter,

insbesondere die Faktorladungen dieser Modelle, welche primär für die Schätzung der Bartlett-Faktorwerte verantwortlich sind, sehr nahe an den in den Populationsmodellen definierten Parametern (Unterschiede < .0015 bei standardisierten Ladungen), weshalb diese Befunde trotz dieser Einschränkung als erwähnenswert gelten.

2.2 Faktorwerte missspezifizierter Modelle

2.3.1 Korrelationen

Auch bei Anwendung des missspezifizierten Modells auf die Populationsdaten wurde zunächst die Validität (Grice, 2001a, 2001b) der Faktorwerte selbst bestimmt, die aus diesem missspezifizierten Modell geschätzt wurden, bevor die Validität der Diagnosen aus den Faktorwerten des missspezifizierten Modells bestimmt wurde. Dazu wurden die wahren Faktorwerte der beiden definierten obliquen Faktoren jeweils mit den Bartlett-Faktorwerten des einen Faktors aus dem missspezifizierten Modell korreliert.

Es zeigte sich, dass die Höhe der Ladungen in den Populationsmodellen einen Einfluss auf die Korrelationen der wahren Faktorwerte mit den Bartlett-Faktorwerten aus dem missspezifizierten Modell hatte (siehe Tabelle 13). Hohe Faktorladungen führten zu höheren Korrelationen (Unterschiede zu typischen Faktorladungen auf der ersten Dezimalstelle). Die unausgewogene Indikatorenaufteilung im Populationsmodell (geringerer Grad an Missspezifikation) führte im Vergleich zur gleichmäßigen Aufteilung zu geringfügig höheren Korrelationen der Faktorwerte der ersten Faktoren (überrepräsentiert durch 15 Indikatoren im Populationsmodell) mit den Faktorwerten aus dem eindimensionalen missspezifizierten Modell mit den 20 Indikatoren (Unterschiede auf der zweiten Dezimalstelle). Umgekehrt ergaben die Korrelationen der Faktorwerte der zweiten Faktoren (5 Indikatoren im Populationsmodell) mit den Faktorwerten aus dem missspezifizierten Modell mit den 20 Indikatoren im Vergleich zur gleichmäßigen Indikatorenaufteilung auf die Faktoren im Populationsmodell geringfügig niedrigere Korrelationen (siehe Tabelle 13; Unterschiede auf der zweiten Dezimalstelle). Für diesen Befund können zwei alternative Erklärungen angeführt werden. Die erste Erklärung betrifft den Grad der Missspezifikation operationalisiert an der Indikatorenaufteilung in den Populationsmodellen. Das missspezifizierte Modell lag mit den 20 Indikatoren näher am Populationsmodell mit unausgewogener Itemaufteilung, dies könnte sich positiv auf die Korrelation zwischen den Faktorwerten des missspezifizierten Modells und dem ersten wahren

Faktor ausgewirkt haben und negativ auf die Korrelation der Faktorwerte des missspezifizierten Modells und dem zweiten wahren Faktor. Als zweite mögliche Erklärung wird die Indikatorenanzahl in Betracht gezogen. Mehr Indikatoren pro Faktor führen zu valideren Faktorwerteschätzungen als weniger Indikatoren (Estabrook & Neale, 2013). Außerdem nähern sich die Verteilungseigenschaften der geschätzten Faktorwerte mehr denen der True Scores an, wenn die Indikatorenanzahl höher ist im Vergleich zu niedriger (Lawley & Maxwell, 1971).

Tabelle 13

Korrelationen der wahren Faktorwerte und der Faktorwerte aus dem missspezifizierten Modell

	Modell 1: .80 [.20, .40] 10:10	Modell 2: .50 [.20, .40] 10:10	Modell 3: .30 [.20, .40] 10:10	Modell 4: .80 [.20, .40] 15:5	Modell 5: .50 [.20, .40] 15:5	Modell 6: .30 [.20, .40] 15:5
$r_{W1\text{-}M}$	.779***	.698***	.651***	.816***	.801***	.797***
$r_{W2\text{-}M}$	.774***	.674***	.589***	.716***	.495***	.312***
	Modell 7: .80 [.40, .60] 10:10	Modell 8: .50 [.40, .60] 10:10	Modell 9: .30 [.40, .60] 10:10	Modell 10: .80 [.40, .60] 15:5	Modell 11: .50 [.40, .60] 15:5	Modell 12: .30 [.40, .60] 15:5
$r_{W1\text{-}M}$	.887***	.807***	.780***	.927***	.920***	.920***
$r_{W2\text{-}M}$	.882***	.784***	.680***	.811***	.555***	.346***

Anmerkungen. $r_{W1\text{-}M}$ = Korrelation der wahren Faktorwerte des ersten definierten Faktors mit den Bartlett-Faktorwerten des Faktors des missspezifizierten Modells, $r_{W2\text{-}M}$ = Korrelation der wahren Faktorwerte des zweiten definierten Faktors mit den Bartlett-Faktorwerten des Faktors des missspezifizierten Modells, *** = höchst signifikanter Zusammenhang. Die Zellen der Beschreibung der Populationsmodelle enthalten als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

Höhere Faktorkorrelationen in den Populationsmodellen (geringere Missspezifikation) führten im Vergleich zu geringen und mittleren Faktorkorrelationen (höhere Missspezifikation) zu höheren Korrelationen der wahren Faktorwerte mit den Faktorwerten aus dem missspezifizierten Modell, wobei sich dieser Unterschied bei der unausgewogenen Indikatorenaufteilung nur auf der zweiten und dritten Dezimalstelle zeigte und demnach nicht überinterpretiert werden sollte (siehe Tabelle 13).

Insgesamt führte die Missspezifikation zu ähnlich hohen Korrelationen der Populationsfaktorwerte und der aus dem missspezifizierten Modell geschätzten Bartlett-

Faktorwerte wie die Anwendung des korrekten Modells auf die Populationsdaten (vgl. die Tabellen 11 und 13).

Dass die Faktorladungen großen Einfluss auf die Berechnung der Bartlett-Faktorwerte hatten und dementsprechend negative Auswirkungen auf die Güte der Diagnostik, wurde bereits im vorherigen Unterkapitel anhand der Anwendung korrekter Modelle illustriert. Daher sind die Auswirkungen der niedrigen Ladungen auf die Faktorenunbestimmtheit ebenso von Interesse, wenn zusätzlich noch eine Missspezifikation vorliegt.

Tabelle 14

Unbestimmtheit der Faktorwerte im missspezifizierten Modell

	Modell 1: .80 [.20, .40] 10:10	Modell 2: .50 [.20, .40] 10:10	Modell 3: .30 [.20, .40] 10:10	Modell 4: .80 [.20, .40] 15:5	Modell 5: .50 [.20, .40] 15:5	Modell 6: .30 [.20, .40] 15:5
ρ^2_{MF}	$.666^{***}$	$.616^{***}$	$.575^{***}$	$.675^{***}$	$.650^{***}$	$.639^{***}$
ρ^*_{MF}	$.331^{***}$	$.232^{***}$	$.151^{***}$	$.349^{***}$	$.301^{***}$	$.278^{***}$
	Modell 7: .80 [.40, .60] 10:10	Modell 8: .50 [.40, .60] 10:10	Modell 9: .30 [.40, .60] 10:10	Modell 10: .80 [.40, .60] 15:5	Modell 11: .50 [.40, .60] 15:5	Modell 12: .30 [.40, .60] 15:5
ρ^2_{MF}	$.863^{***}$	$.830^{***}$	$.802^{***}$	$.869^{***}$	$.856^{***}$	$.851^{***}$
ρ^*_{MF}	$.726^{***}$	$.660^{***}$	$.605^{***}$	$.739^{***}$	$.712^{***}$	$.701^{***}$

Anmerkungen. ρ^2 = Maß für die Höhe der Faktorenunbestimmtheit (Guttman, 1955, S. 73), ρ^* = Maß für die maximal mögliche Unbestimmtheit der Faktorwerte (Guttman, S. 73), MF = Faktor des missspezifizierten Modells, *** = höchst signifikanter Zusammenhang. Die Zellen der Beschreibung der Populationsmodelle enthalten als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

Die Faktorenunbestimmtheit war für den einen Faktor im missspezifizierten Modell bei hohen Ladungen niedriger als bei typischen Ladungen (siehe Tabelle 14). Die Bestimmtheit des Faktors im missspezifizierten Modell stieg geringfügig mit steigender Faktorkorrelation im Populationsmodell, also mit geringerer Missspezifikation, wobei der Unterschied zwischen diesen Bedingungen in elf der zwölf Bedingungen nur auf der zweiten oder dritten Dezimalstelle deutlich wurde (siehe Tabelle 14). Ebenso war die Bestimmtheit des Faktors bei

der unausgewogenen Indikatorenaufteilung (geringere Missspezifikation) geringfügig höher als bei der ausgewogenen Aufteilung (höhere Missspezifikation).

Exkurs: Populationsmodelle mit extrem hohen Ladungen

Wurden im Gegensatz zum Design dieser Studie mit realistischen Faktorladungen extrem hohe Faktorladungen aus dem Bereich [.80, 1.00[für die Generierung der Populationsmodelle verwendet, war die Faktorenbestimmtheit sehr hoch (Range ρ^2 .993-.995) bzw. die Faktorenunbestimmtheit sehr niedrig. Bei den Populationsmodellen mit extrem hohen Ladungen wurde ersichtlich, dass sich die Missspezifikation stärker negativ auf die Validität der Faktorwerte des missspezifizierten Modells auswirkte als bei der Verwendung realistisch hoher Ladungen: Die Differenz zwischen der Korrelation der Faktorwerte des ersten wahren Faktors und den Faktorwerten des missspezifizierten Modells und der Korrelation der Faktorwerte des zweiten wahren Faktors und den Faktorwerten des missspezifizierten Modells wurde größer (Unterschiede bis zu .691). Außerdem stieg die Differenz dieser Korrelationen mit steigendem Grad an Missspezifikation (von .129 bis zu .691 bei schwerwiegender Missspezifikation), operationalisiert an der Faktorkorrelation im Populationsmodell.

2.3.2 Güte der Diagnostik

Es wurde untersucht, wie die diagnostischen Kennwerte ausfallen würden, wenn die dichotomen Diagnosen auf Basis der Bartlett-Faktorwerte des einen Faktors im missspezifizierten Modell getroffen wurden. Für die Diagnosegebung wurden univariate Basisraten von 2.5%, 5%, 7.5%, 10%, 30%, 50% und 70% für die Klassifikation nach den höchsten Faktorwerten des einen Faktors aus dem missspezifizierten Modell verwendet.

Basisraten und Missspezifikation versus korrekte Spezifikation

Durch die univariaten Cut-Offs für die Basisraten im missspezifizierten Modell konnte sichergestellt werden, dass die Raten der Diagnosen dieselben waren wie die Raten an Diagnosen basierend auf den bivariaten Cut-Offs der Populationsmodelle. Dementsprechend verhielten sich die Befunde aus dem missspezifizierten Modell, wenn man sie mit den Ergebnissen aus dem korrekten Modell vergleicht. Dadurch, dass bei Anwendung des korrekten

Modells die kleinen Basisraten überschätzt und die großen unterschätzt wurden, resultierte die Diagnostik auf Basis der Faktorwerte des missspezifizierten Modells in weniger Korrekt Positiven bei kleinen Basisraten und in mehr Korrekt Positiven bei großen Basisraten im Vergleich zur Diagnostik auf Basis der Faktorwerte des korrekten Modells (siehe die Abbildungen 14 und 18 im Anhang). Dementsprechend führten die Faktorwerte des missspezifizierten Modells bei kleinen Basisraten zu mehr Korrekt Negativen und bei großen Basisraten zu weniger Korrekt Negativen als die Faktorwerte des korrekten Modells (siehe die Abbildungen 15 und 19 im Anhang).

Da die True Positives im Rahmen des Designs weniger mit den Basisraten schwankten als die True Negatives –letztere wiesen eine größere Spannweite auf (siehe die Abbildungen 14 und 15 im Anhang) – fielen die False Positives und die False Negatives (siehe die Abbildungen 16 und 17 im Anhang) in Relation zu den True Positives mehr ins Gewicht als in Relation zu den True Negatives. Dieser Befund zeigte sich auch bei der Diagnostik auf Basis der Faktorwerte des korrekten Modells. Dementsprechend beeinträchtigte die Diagnostik auf Basis der Faktorwerte des missspezifizierten Modells den Positiven Prädiktionswert (siehe Abbildung 6), aber auch die Sensitivität (siehe Abbildung 7). Da die Anzahl an False Positives (siehe Abbildung 20 im Anhang) und False Negatives (siehe Abbildung 21 im Anhang) durch das missspezifizierte Modell ähnlich hoch ausfielen, fielen die diagnostischen Kennwerte Positiver Prädiktionswert und Sensitivität sehr ähnlich aus (siehe Abbildungen 7 und 8).

Letzterer Befund, dass die Rate der False Positives und der False Negatives basierend auf den Faktorwerten des missspezifizierten Modells ähnlich hoch ausfiel, erklärt auch den Unterschied zum Positiven Prädiktionswert und zur Sensitivität basierend auf den Diagnosen durch die Faktorwerte des korrekten Modells (vgl. die Abbildungen 2 und 3 mit den Abbildungen 6 und 7): Bei den korrekten Modellen traten bei kleinen Basisraten mehr Falsch Positive auf als Falsch Negative (Unterschiede von bis zu 1 Prozentpunkt an allen Diagnosen je Populationsbedingung auf Basis der Faktorwerte korrekter Modelle) und die Falsch Positiven schwankten mehr mit den Modellbedingungen als die Falsch Negativen (Unterschiede unter 10 Prozentpunkten auf Basis der korrekten Modelle bei den Falsch Positiven versus unter 1 Prozentpunkt bei den Falsch Negativen bei allen Diagnosen in der Population). Bei den großen Basisraten verhielt sich dies umgekehrt (vgl. die Abbildungen 16 und 17 mit den Abbildungen 20 und 21 im Anhang).

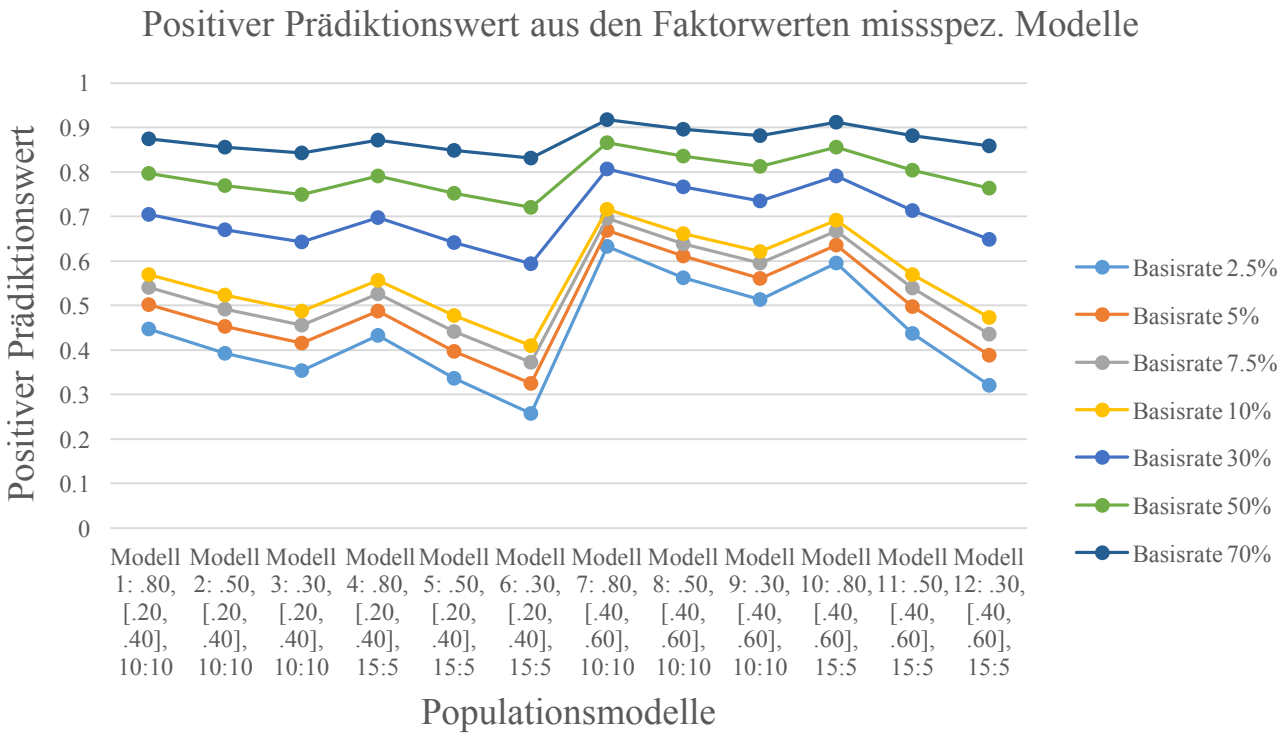

Abbildung 6. Positiver Prädiktionswert der Diagnostik basierend auf den Bartlett-Faktorwerten missspezifizierter Modelle

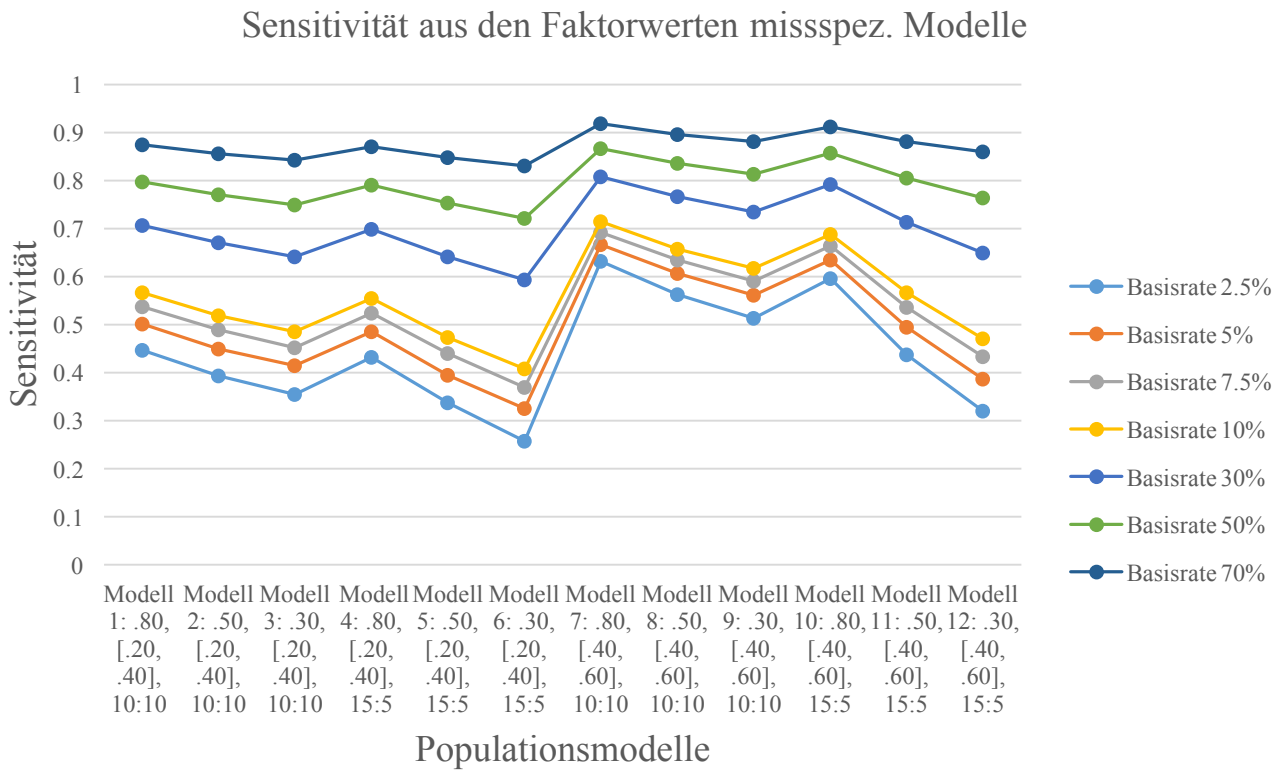

Abbildung 7. Sensitivität der Diagnostik basierend auf den Bartlett-Faktorwerten missspezifizierter Modelle

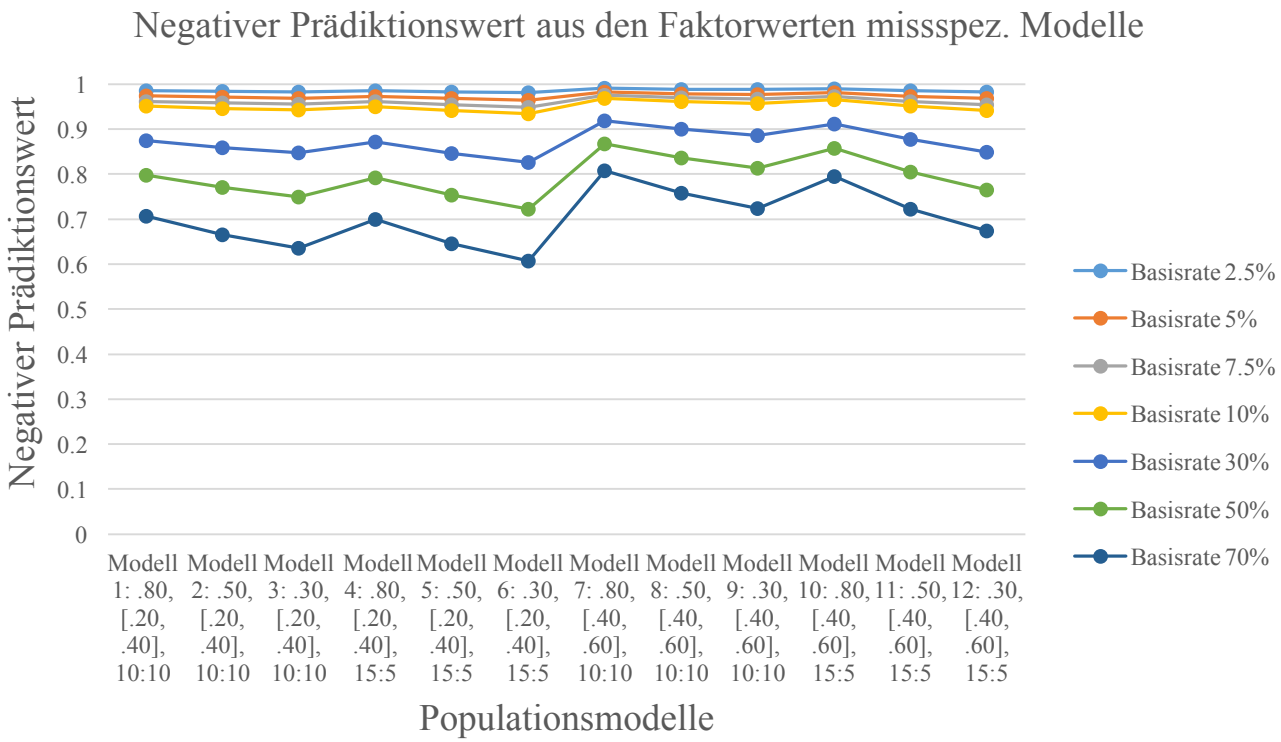

Abbildung 8. Negativer Prädiktionswert der Diagnostik basierend auf den Bartlett-Faktorwerten missspezifizierter Modelle

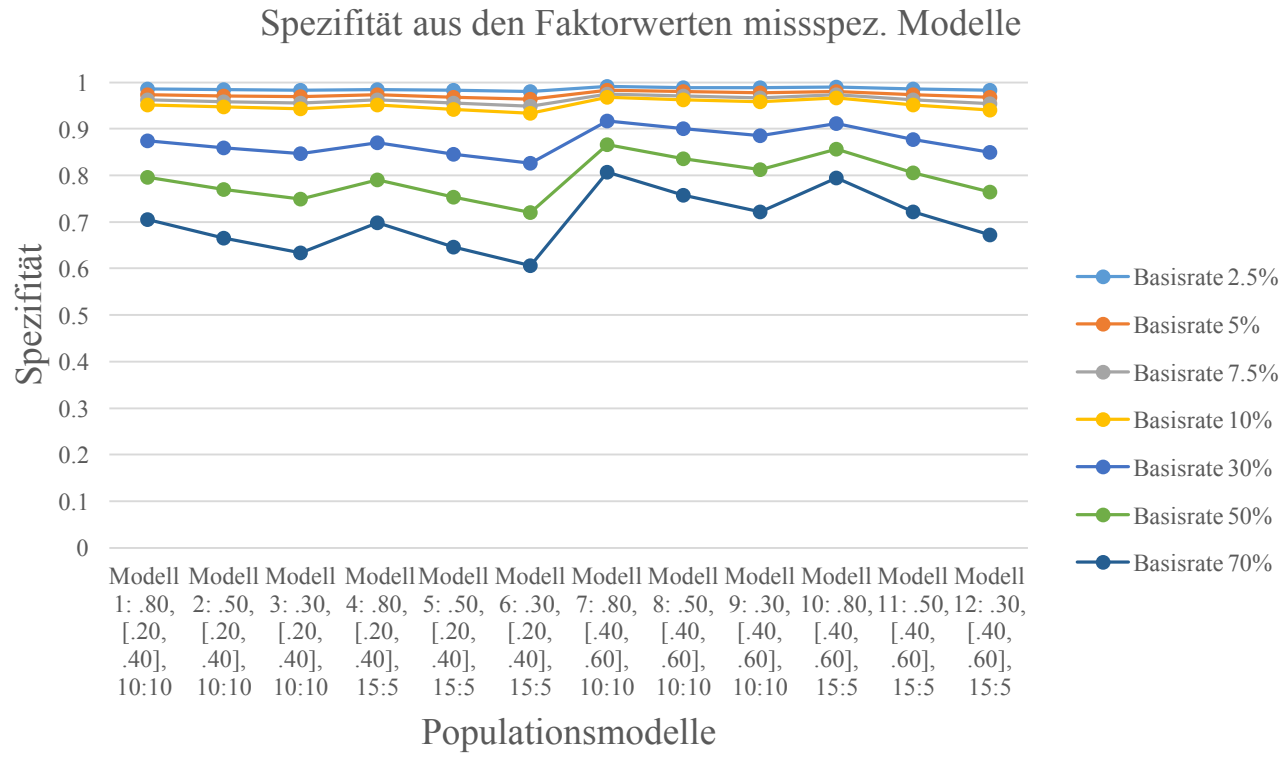

Abbildung 9. Spezifität der Diagnostik basierend auf den Bartlett-Faktorwerten missspezifizierter Modelle

Negativer Prädiktionswert (siehe Abbildung 8) und Spezifität (siehe Abbildung 9) schwankten weniger mit den Basisraten als Positiver Prädiktionswert und Sensitivität, da insbesondere bei den kleinen Basisraten die Rate der Richtig Negativen sehr hoch war (siehe Abbildung 19 im Anhang). Bei den großen Basisraten, bei denen sich die Anzahl der positiven und negativen Fälle an sich stärker annäherten als bei den kleinen Basisraten, schwankten Negativer Prädiktionswert und Spezifität mehr (siehe Abbildung 8 und 9).

Zusammenfassend ist zu sagen, dass sich die Missspezifikation im Vergleich zur korrekten Spezifikation des Modells vor allem negativ auf die Sensitivität bei kleinen Basisraten und auf die Spezifität bei großen Basisraten auswirkte. Die Diagnostik auf Basis der Faktorwerte des missspezifizierten Modells unterschied sich im Vergleich zur Diagnostik auf Basis der Faktorwerte des korrekten Modells im gravierendsten Fall um 28 Prozentpunkte bei der kleinsten Basisrate hinsichtlich der Sensitivität und um 17 Prozentpunkte bei der größten Basisrate hinsichtlich der Spezifität. Allerdings konnte die Missspezifikation im Vergleich zur korrekten Spezifikation des Modells – je nach betrachtetem diagnostischen Kennwert – durch eine geringe Basisrate (Negativer Prädiktionswert und Spezifität) oder hohe Basisrate (Positiver Prädiktionswert und Sensitivität) ausgeglichen werden.

Basisraten und Grade der Missspezifikation

Die psychometrischen Auswirkungen der unterschiedlichen Grade der Missspezifikation wie auch der unterschiedlichen Basisraten auf die Güte der Diagnostik auf Basis der Faktorwerte wurden an allen diagnostischen Kennwerten deutlich (siehe die Abbildungen 6 bis 9). Außerdem interagierte der Grad der Missspezifikation mit der Höhe der Basisrate. Beide Befunde werden im Folgenden näher beschrieben.

Mit sinkender Faktorkorrelation im Populationsmodell (höherer Grad an Missspezifikation im Vergleich zu höherer Faktorkorrelation) sanken alle diagnostischen Kennwerte. Außerdem zeigten alle vier diagnostischen Kennwerte, dass sich die verschiedenen Grade an Missspezifikation bei den vier kleinen Basisraten gravierender auswirkten als bei den drei großen Basisraten (siehe die Abbildungen 6 bis 9). Zwischen der Bedingung mit der höchsten Faktorkorrelation im Populationsmodell (geringe Missspezifikation) und der geringsten Faktorkorrelation im Populationsmodell (hohe Missspezifikation) gab es hinsichtlich der Güte der Diagnostik bei kleinen Basisraten Unterschiede von bis zu 28

Prozentpunkten, bei großen Basisraten von bis zu 12 Prozentpunkten in der Richtung, als dass bei höherem Grad der Missspezifikation (geringerer Faktorkorrelation) die diagnostischen Kennwerte sanken.

Die Diagnostik auf Basis einer unausgewogenen Indikatorenaufteilung führte im Vergleich zu einer ausgewogenen Indikatorenaufteilung im Populationsmodell zu schlechterer Diagnostik auf Basis der Faktorwerte des missspezifizierten Modells. Bei kleinen Basisraten war die Diagnostik auf Basis der Faktorwerte des missspezifizierten Modells bei einer unausgewogenen Indikatorenaufteilung im Populationsmodell um bis zu 19 Prozentpunkte schlechter als bei einer ausgewogenen Aufteilung, bei großen Basisraten um bis zu 5 Prozentpunkte schlechter. Als Begründung für diesen Befund kann angeführt werden, dass die geschätzten Faktorladungen der missspezifizierten Modelle, die auf Populationsdaten aus Modellen mit unausgewogener Indikatorenaufteilung angewandt wurden, breiter streuten (*SD* .063-.157 bei standardisierten Ladungen) als die geschätzten Faktorladungen der missspezifizierten Modelle, die auf Populationsdaten aus Modellen mit ausgewogener Indikatorenaufteilung angewandt wurden (*SD* .050-.059 bei standardisierten Ladungen). Die Streuung der Faktorladungen der missspezifizierten Modelle, denen Populationsdaten auf Basis einer unausgewogenen Aufteilung zugrunde lagen, wich außerdem stärker von der Streuung der definierten Faktorladungen in den Populationsmodellen ab (Unterschied maximal .10 bei standardisierten Ladungen) als die Faktorladungen der missspezifizierten Modelle, die auf die Daten mit einer ausgewogenen Indikatorenaufteilung im Populationsmodell angewandt wurden (Unterschied maximal .013 bei standardisierten Ladungen). Die Ladungen sind, wie bereits beschrieben, primär verantwortlich für die Berechnung der Bartlett-Faktorwerte. Der Befund wird unter 3.2 näher diskutiert.

Faktorladungen

Auch bei Anwendung des missspezifizierten Modells führten höheren Faktorladungen (Modelle 7 bis 12) zu höheren diagnostischen Kennwerten (siehe die Abbildungen 6 bis 9). Dass die Diagnostik auf Basis der Faktorwerte des unterschiedlich stark missspezifizierten Modells im Vergleich zur korrekten Spezifikation des Modells nicht zu noch größeren negativen Auswirkungen führte, lag wiederum an den realistisch niedrig gewählten Faktorladungen des Populationsmodells. Diese bewirkten bereits eine starke Beeinträchtigung der Güte der Diagnostik auf Basis der Bartlett-Faktorwerte des korrekten Modells (so zum

Beispiel im gravierendsten Fall einen Positiven Prädiktionswert von .240 bei der kleinsten Basisrate oder einen Negativen Prädiktionswert von .529 bei der größten Basisrate).

Exkurs: Populationsmodelle mit extrem hohen Ladungen

Einerseits zeigten die Ergebnisse, dass höhere Faktorladungen im Populationsmodell (Modelle 7 bis 12 in den Abbildungen) anhand aller vier diagnostischen Kennwerte zu besserer Diagnostik führten, und dies sowohl, wenn die Diagnostik auf Basis der Faktorwerte des korrekten Modells erfolgte (siehe die Abbildungen 2 bis 5), als auch basierend auf den Faktorwerten des missspezifizierten Modells (siehe die Abbildungen 6 bis 9). Andererseits wurden die negativen Auswirkungen der Missspezifikation auf die Diagnostik erst in vollem Umfang ersichtlich, sobald wiederum aus Vergleichsgründen extrem hohe Faktorladungen aus dem Bereich [.80, 1.00[zur Generierung der Populationsmodelle verwendet wurden. Die diagnostischen Kennwerte schwankten hier deutlicher mit dem Grad der Missspezifikation als bei den realistisch gewählten Faktorladungen und der Grad der Missspezifikation interagierte gleichzeitig wiederum mit der Höhe der Basisrate. Ein höherer Grad an Missspezifikation beeinträchtigte vor allem Sensitivität und Positiven Prädiktionswert, wenn die Basisrate gering war[43]. Der Grad der Missspezifikation wirkte sich auch auf den Negativen Prädiktionswert und die Spezifität aus[44], allerdings in geringerem Maße als auf die Sensitivität und den Positiven Prädiktionswert. Außerdem konnte die Missspezifikation im Vergleich zu einer korrekten Spezifikation auch bei den extrem hohen Ladungen in den Populationsmodellen in hohem Maße ausgeglichen werden, wenn die Basisrate, je nach Kennwert, entsprechend niedrig oder hoch war. Sensitivität und Positiver Prädiktionswert sanken bei der größten Basisrate von 70%

[43]Range der Sensitivität bei geringer Faktorkorrelation (hoher Missspezifikation) und extrem hohen Ladungen in den Populationsmodellen .299-.946, bei hoher Faktorkorrelation (geringe Missspezifikation) .737-.946; Range des Positiven Prädiktionswerts bei geringer Faktorkorrelation (hoher Missspezifikation) und extrem hohen Ladungen in den Populationsmodellen .299-.945, bei hoher Faktorkorrelation (geringer Missspezifikation) .736-.945 (zum Vergleich siehe die Abbildungen 6 und 7).

[44]Range des Negativen Prädiktionswerts bei geringer Faktorkorrelation (hoher Missspezifikation) und extrem hohen Ladungen in den Populationsmodellen .687-.993, bei hoher Faktorkorrelation (geringer Missspezifikation) .874-.993; Range der Spezifität bei geringer Faktorkorrelation (hoher Missspezifikation) und extrem hohen Ladungen in den Populationsmodellen .686-.993, bei hoher Faktorkorrelation (geringer Missspezifikation) .873-.993 (zum Vergleich siehe die Abbildungen 8 und 9).

unabhängig vom Grad der Missspezifikation nicht unter .865, Spezifität und Positiver Prädiktionswert sanken bei der kleinsten Basisrate von 2.5% nicht unter .982.

2.4 Gesamtsummenwerte

2.4.1 Korrelationen

Im Rahmen einer Nebenfragestellung wurden Diagnosen auf Basis der Gesamtsummenwerte gegeben, auf deren Basis, wie bereits unter III. 5.2 beschrieben, häufig psychologische Diagnosen vergeben werden.

Die Gesamtsummenwerte wurden über die 20 auf Basis der Populationsfaktorwerte erzeugten Indikatoren gebildet. Diese Diagnosen wurden ebenfalls mit den Diagnosen basierend auf den wahren Faktorwerten beider Populationsfaktoren anhand der diagnostischen Kennwerte verglichen. Die Anwendung der Gesamtsummenwerte zur Diagnostik unterschied sich insofern von der Anwendung der Faktorwerte eines missspezifizierten Modells, als dass alle Indikatoren bei der Aufsummierung die gleichen Gewichtungsfaktoren bekamen; die Gewichtungsfaktoren/Faktorladungen des missspezifizierten Modells, auf dessen Basis die Faktorwerte berechnet wurden, waren heterogen.

Zunächst wurden die Gesamtsummenwerte mit den wahren Faktorwerten der entsprechenden Populationsbedingungen korreliert (siehe Tabelle 15).

Auch bei der Korrelation der wahren Faktorwerte mit den Gesamtsummenwerten hatte die Höhe der Ladungen in den Populationsmodellen einen Einfluss auf die Güte der Diagnostik auf Basis der Gesamtsummenwerte (siehe Tabelle 15): Hohe Faktorladungen im Vergleich zu typischen Faktorladungen in den Populationsmodellen führten zu höheren Korrelationen der Gesamtsummenwerte mit den beiden Faktoren der Populationsmodelle. Außerdem führten höhere Faktorkorrelationen in den Populationsmodellen im Vergleich zu geringen und mittleren Faktorkorrelationen zu höheren Korrelationen der wahren Faktorwerte mit den Gesamtsummenwerten. Die unausgewogene Indikatorenaufteilung führte im Vergleich zur gleichmäßigen Aufteilung zu höheren Korrelationen der Faktorwerte der ersten Populationsfaktoren mit den Gesamtsummenwerten. Umgekehrt korrelierten die Faktorwerte der zweiten Populationsfaktoren niedriger mit den Gesamtsummenwerten als bei der ausgewogenen Itemaufteilung. Insgesamt ähnelten die Ergebnisse stark den unter 2.3.1

beschriebenen zur Diagnostik auf Basis der Faktorwerte des eindimensionalen missspezifizierten Modells.

Tabelle 15

Korrelationen der wahren Faktorwerte und der Gesamtsummenwerte

	Modell 1: .80 [.20, .40] 10:10	Modell 2: .50 [.20, .40] 10:10	Modell 3: .30 [.20, .40] 10:10	Modell 4: .80 [.20, .40] 15:5	Modell 5: .50 [.20, .40] 15:5	Modell 6: .30 [.20, .40] 15:5
r_{W1-S}	.770***	.683***	.620***	.805***	.776***	.755***
r_{W2-S}	.768***	.676***	.609***	.718***	.548***	.426***
	Modell 7: .80 [.40, .60] 10:10	Modell 8: .50 [.40, .60] 10:10	Modell 9: .30 [.40, .60] 10:10	Modell 10: .80 [.40, .60] 15:5	Modell 11: .50 [.40, .60] 15:5	Modell 12: .30 [.40, .60] 15:5
r_{W1-S}	.883***	.797***	.734***	.920***	.897***	.882***
r_{W2-S}	.881***	.792***	.727***	.822***	.636***	.502***

Anmerkungen. r_{W1-S} = Korrelation der wahren Faktorwerte des ersten definierten Faktors mit den Summenwerten, r_{W2-S} = Korrelation der wahren Faktorwerte des zweiten definierten Faktors mit den Summenwerten, *** = höchst signifikanter Zusammenhang. Die Zellen der Beschreibung der Populationsmodelle enthalten als ersten Wert die Höhe der Faktorkorrelation, der Bereich bezeichnet das Intervall, aus dem die Faktorladungen gezogen wurden und das Verhältnis gibt die Indikatorenaufteilung auf die Faktoren an.

Exkurs: Populationsmodelle mit extrem hohen Ladungen

Wurden wiederum die extrem hohen Faktorladungen aus dem Bereich [.80, 1.00[für die Generierung der Populationsmodelle verwendet, wird ersichtlich, dass auch die Korrelationen der Gesamtsummenwerte mit den Faktorwerten des ersten und zweiten Populationsfaktors im Vergleich zu den realistisch gewählten Faktorladungen des vorliegenden Designs (siehe Tabelle 13) deutlich stiegen[45]. Außerdem führten höhere Faktorkorrelationen in den Populationsmodellen im Vergleich zu geringen und mittleren Faktorkorrelationen zu

[45]Korrelationen der Gesamtsummenwerte mit den Faktorwerten des ersten Populationsfaktors .802-.983; Korrelationen der Gesamtsummenwerte mit den Faktorwerten des zweiten Populationsfaktors .546-.943

höheren Korrelationen der wahren Faktorwerte mit den Gesamtsummenwerten[46]. Die unausgewogene Indikatorenaufteilung führte im Vergleich zur ausgewogenen zu geringfügig höheren Korrelationen der Faktorwerte der ersten Populationsfaktoren mit den Gesamtsummenwerten; umgekehrt korrelierten die Faktorwerte der zweiten Populationsfaktoren geringfügig niedriger mit den Gesamtsummenwerten als bei der ausgewogenen Itemaufteilung[47].

2.4.2 Güte der Diagnostik

Es wurde untersucht, wie hoch die diagnostischen Kennwerte ausfallen, wenn dichotome Diagnosen, die auf Basis der höchsten wahren Faktorwerte zweier obliquer Faktoren in der Population gebildet wurden, auf Basis der höchsten Gesamtsummenwerte über alle Indikatoren vergeben wurden. Für die Diagnosegebung auf Basis der Gesamtsummenwerte wurden wiederum univariate Basisraten von 2.5%, 5%, 7.5%, 10%, 30%, 50% und 70% verwendet.

Basisraten

Dadurch, dass im korrekten Modell die kleinen Basisraten überschätzt und die großen unterschätzt wurden, resultierte die Diagnostik auf Basis der Gesamtsummenwerte genauso wie auf Basis der Faktorwerte des missspezifizierten Modells in weniger Korrekt Positiven (siehe Abbildung 22 im Anhang) bei kleinen Basisraten und in mehr Korrekt Positiven bei großen Basisraten im Vergleich zur Diagnostik auf Basis der Faktorwerte des korrekten Modells. Dementsprechend führten die Gesamtsummenwerte wie die Faktorwerte des missspezifizierten Modells bei kleinen Basisraten zu mehr Korrekt Negativen (siehe Abbildung 23 im Anhang)

[46]Die Korrelationen der zu .80 korrelierten True Scores mit den Gesamtsummenwerten lagen zwischen .878-.983, wohingegen sich die Korrelationen der zu .50 oder zu .30 korrelierten wahren Faktorwerte zwischen .546-.965 bewegten.

[47]Korrelation der Gesamtsummenwerte mit den True Scores ersten Faktors bei unausgewogener Indikatorenaufteilung im Populationsmodell .955-.983, bei ausgewogener .802-.944;
Korrelationen der Gesamtsummenwerte mit den True Scores des zweiten Faktors bei unausgewogener Indikatorenaufteilung im Populationsmodell .546-.878, bei ausgewogener .798-.943.

und bei großen Basisraten zu weniger Korrekt Negativen als die Faktorwerte des korrekten Modells.

Die Abbildungen 10 bis 13 zeigen die Güte der Diagnostik auf Basis der Gesamtsummenwerte. Die Güte der Diagnostik auf Basis der Gesamtsummenwerte stieg bei steigender Faktorkorrelation des Populationsmodells, aus dem die Daten erzeugt wurden. Eine unausgewogene Indikatorenaufteilung im Populationsmodell im Vergleich zu einer ausgewogenen führte in Interaktion mit einer kleinen Basisrate außerdem zu geringfügig niedrigeren Werten (Unterschiede maximal 7 Prozentpunkte) hinsichtlich der Sensitivität (siehe Abbildung 11) und des Positiven Prädiktionswerts (siehe Abbildung 10) und in Interaktion mit großen Basisraten zu maximal zwei Prozentpunkten Unterschied hinsichtlich Spezifität (siehe Abbildung 13) und Negativem Prädiktionswert (siehe Abbildung 12).

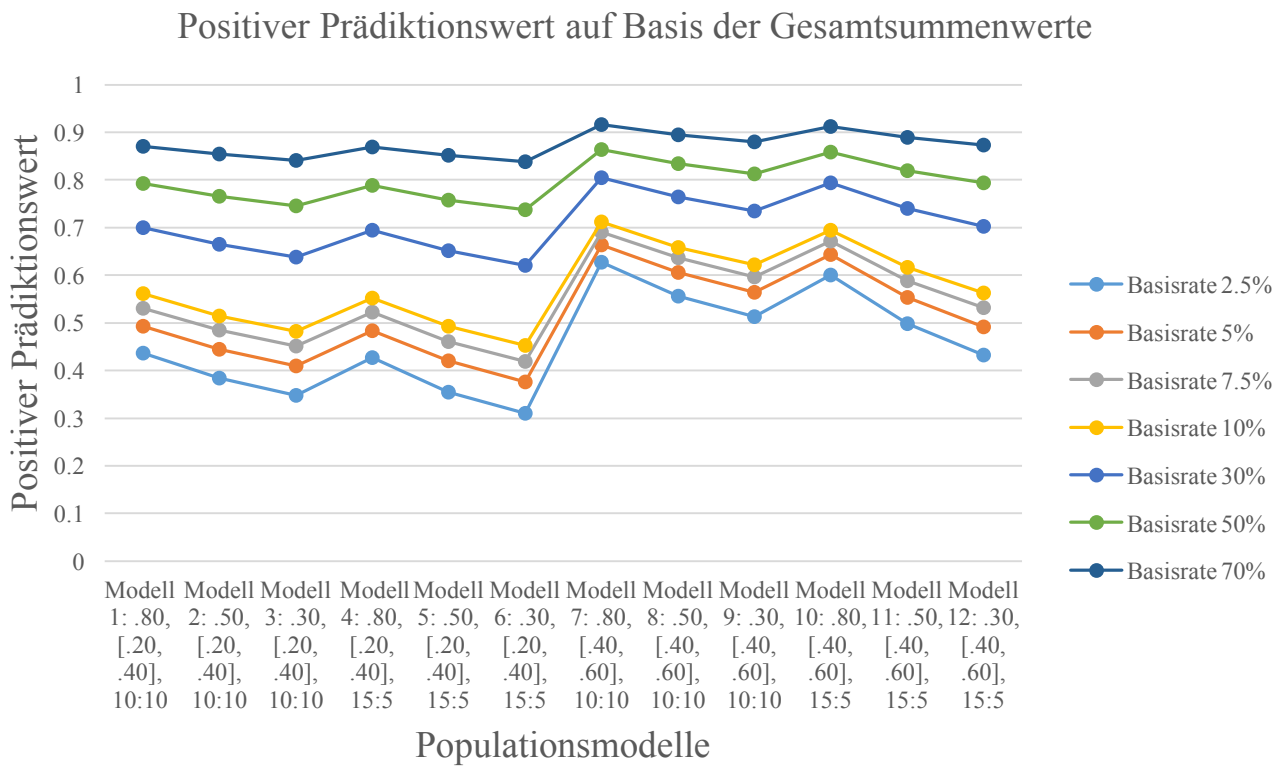

Abbildung 10. Positiver Prädiktionswert der Diagnostik basierend auf den Gesamtsummenwerten

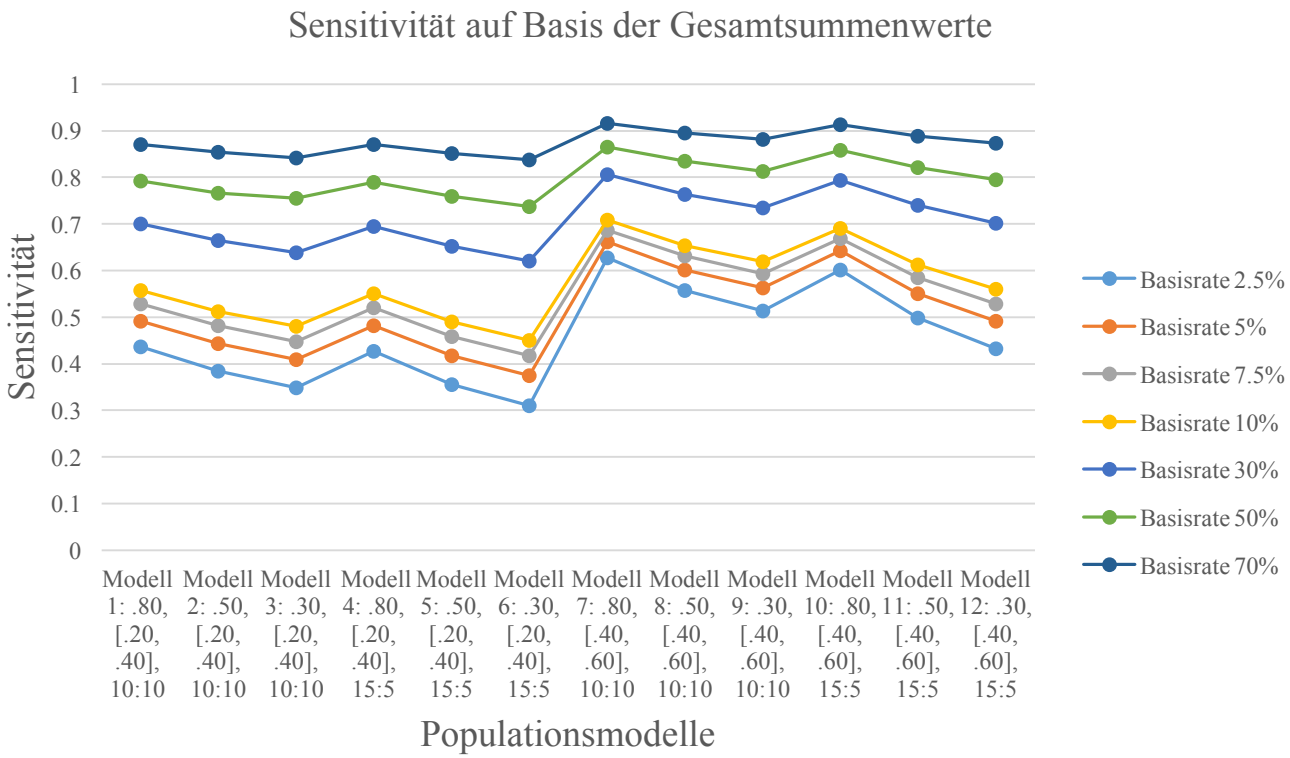

Abbildung 11. Sensitivität der Diagnostik basierend auf den Gesamtsummenwerten

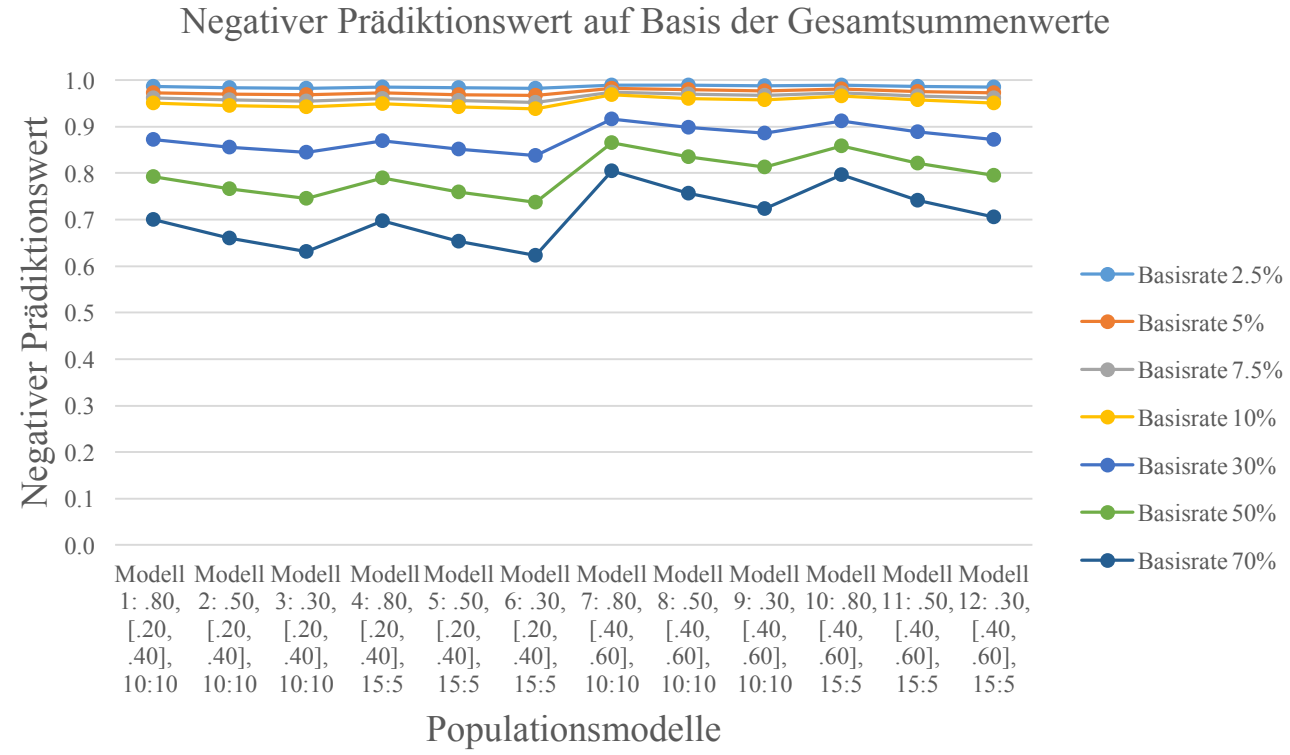

Abbildung 12. Negativer Prädiktionswert der Diagnostik basierend auf den Gesamtsummenwerten

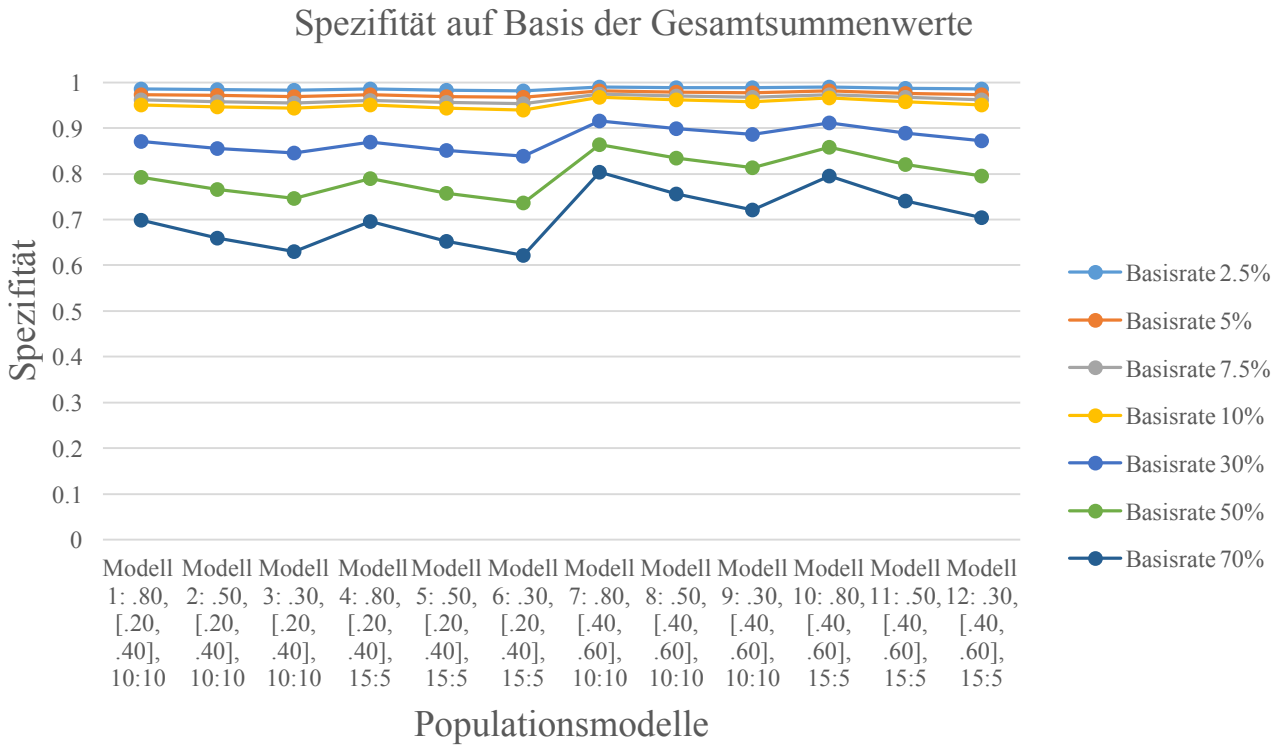

Abbildung 13. Spezifität der Diagnostik basierend auf den Gesamtsummenwerten

Basisraten und Gesamtsummenwerte versus Faktorwerte der korrekten Modelle

Insgesamt führte die Diagnostik auf Basis der Gesamtsummenwerte im Vergleich zur Diagnostik auf Basis der Bartlett-Faktorwerte des korrekten Modells vor allem zu Einbußen hinsichtlich der Sensitivität bei kleinen Basisraten (Unterschiede von bis zu knapp 17 Prozentpunkten zwischen der Diagnostik auf Basis der Faktorwerte des korrekten Modells und auf Basis der Gesamtsummenwerte). Umgekehrt beeinflusste die Diagnostik auf Basis der Gesamtsummenwerte aber auch die Spezifität bei großen Basisraten negativ (Unterschiede bis zu 9 Prozentpunkte zwischen der Diagnostik auf Basis der Faktorwerte des korrekten Modells und auf Basis der Gesamtsummenwerte). Dieses Muster in den Befunden zeigten sich auch, wenn die Diagnostik auf Basis der Faktorwerte des missspezifizierten Modells vorgenommen wurde und diese mit der Diagnostik auf Basis der Faktorwerte der korrekten Modelle verglichen wurde, wie im Folgenden berichtet wird.

Basisraten und Gesamtsummenwerte versus Faktorwerte des missspezifizierten Modells

Die Raten der Richtig Positiven, der Richtig Negativen, der Falsch Positiven und der Falsch Negativen auf Basis der Gesamtsummenwerte (siehe Abbildungen 22 bis 25 im Anhang) fielen im Vergleich zu den entsprechenden Raten auf Basis der Faktorwerte des missspezifizierten Modells sehr ähnlich aus (Unterschiede unter 1 Prozentpunkt an allen vergebenen Diagnosen in der jeweiligen Gesamtpopulation pro Bedingung). Verglichen mit der Diagnostik auf Basis der Faktorwerte des missspezifizierten Modells war die Diagnostik auf Basis der Gesamtsummenwerte marginal überlegen. Bei der kleinsten Basisrate führten die Gesamtsummenwerte zu maximal 8 Prozentpunkten besserer Diagnostik hinsichtlich Sensitivität und Positivem Prädiktionswert, wobei sich hinsichtlich der großen Basisraten diesbezüglich kein Unterschied mehr zeigte. Umgekehrt führten die Gesamtsummenwerte bei der größten Basisrate zu maximal 5 Prozentpunkten besserer Diagnostik hinsichtlich Negativem Prädiktionswert und Spezifität im Vergleich zu den Faktorwerten des missspezifizierten Modells, wobei sich dieser Unterschied bei kleinen Basisraten ausglich. Dass Positiver Prädiktionswert und Sensitivität bzw. Negativer Prädiktionswert und Spezifität auf Basis der Gesamtsummenwerte in Kombination mit einer kleinen bzw. großen Basisrate höher ausfielen als auf Basis der Faktorwerte des missspezifizierten Modells, kann zum Teil mit der Reduktion der Faktorladungen in den missspezifizierten Modellen im Vergleich zur Höhe der definierten Faktorladungen in den Populationsmodellen erklärt werden. Die standardisierten Faktorladungen des missspezifizierten Modells waren im Mittel um bis zu .106 im Vergleich zu den definierten Faktorladungen geringer. Ferner streuten die Ladungen der missspezifizierten Modelle im Mittel breiter als die definierten Faktorladungen (bis zu .10 Unterschied bei standardisierten Ladungen). Die Faktorladungen beeinflussen, wie bereits mehrfach erläutert wurde, wiederum die Schätzung der Bartlett-Faktorwerte. Dieser Befund wird unter 3.2 anhand weiterer Literatur diskutiert.

Faktorladungen

Wie bei der Diagnostik auf Basis der Faktorwerte der korrekten und missspezifizierten Modelle stieg auch die Güte der Diagnostik auf Basis der Gesamtsummenwerte anhand aller Kennwerte, wenn zur Generierung der Daten im Populationsmodell höhere Ladungen im

Vergleich zu typischen Ladungen verwendet wurden (vgl. die Modelle 1 bis 6 mit den Modellen 7 bis 12 in den Abbildungen 10 bis 14).

Exkurs: Populationsmodelle mit extrem hohen Ladungen

Sofern in den Populationsmodellen extrem hohe Ladungen aus dem Bereich [.80, 1.00[zur Generierung der Populationsdaten verwendet wurden, stieg auch die Güte der Diagnostik auf Basis der Gesamtsummenwerte[48]. Es wurde anhand der extrem hohen Ladungen deutlich, dass die Diagnostik auf Basis der Gesamtsummenwerte der Diagnostik auf Basis der Faktorwerte des korrekten Modells unterlegen war[49]. Die Befunde zum Vergleich der Diagnostik auf Basis der Gesamtsummenwerte und der Faktorwerte des missspezifizierten Modells mit extrem hohen Ladungen in den Populationsmodellen werden im Folgenden berichtet.

Vergleicht man die Diagnostik auf Basis der Faktorwerte des eindimensionalen missspezifizierten Modells mit der Diagnostik auf Basis der Gesamtsummenwerte, wenn die Daten basierend auf Populationsmodellen mit extrem hohen Ladungen erzeugt wurden, fällt die Diagnostik auf Basis der Gesamtsummenwerte bei den kleinen Basisraten besser aus als die Diagnostik auf Basis der Faktorwerte des missspezifizierten Modells[50]. Bei den großen Basisraten war die Diagnostik auf Basis der Gesamtsummenwerte zwar immer noch geringfügig besser als auf Basis der Faktorwerte des missspezifizierten Modells, jedoch glich sich die Güte der Diagnostik auf Basis der Gesamtsummenwerte mehr und mehr der diagnostischen Präzision auf Basis der Faktorwerte des missspezifizierten Modells an (Unterschiede maximal 5 Prozentpunkte bei großen Basisraten). Dieser Befund mag, wie bereits im Rahmen der realistisch hoch gewählten Faktorladungen des vorliegenden Designs beschrieben, einerseits an der Reduktion und andererseits an der zunehmenden Heterogenität der Faktorladungen des missspezifizierten Modells im Vergleich zu den definierten

[48]Der maximale Range der diagnostischen Kennwerte auf Basis der Gesamtsummenwerte bei extrem hohen Ladungen in den Populationsmodellen über alle Basisraten und Bedingungen hinweg betrug .515-.955.

[49]Der maximale Range der diagnostischen Kennwerte auf Basis der Faktorwerte des korrekten Modells mit extrem hohen Ladungen in den Populationsmodellen über alle Basisraten und Bedingungen hinweg betrug .895-.984.

[50]Der maximale Range der diagnostischen Kennwerte auf Basis der Faktorwerte des missspezifizierten Modells mit extrem hohen Ladungen in den Populationsmodellen betrug über alle Basisraten und Bedingungen hinweg .299-.946, im Gegensatz dazu bei den Gesamtsummenwerten .515-955.

Faktorladungen liegen. Im Rahmen der Populationsmodelle mit extrem hohen Faktorladungen im Vergleich zu den realistischen Faktorladungen sanken die Ladungen des eindimensionalen missspezifizierten Modells in höherem Maße (um bis zu .308 bei standardisierten Ladungen im Vergleich zu max. .11 bei den realistischen Ladungen). Wie aus den vorherigen Ergebnissen ersichtlich steigt die Validität (Grice, 2001a, 2001b) der Bartlett-Faktorwerte, aus denen die Diagnosen gebildet wurden, mit der Höhe der Faktorladungen. Die Ladungen des missspezifizierten Modells wurden außerdem umso geringer, desto geringer die Faktorkorrelation (desto höher die Missspezifikation) im Populationsmodell mit extrem hohen Ladungen war (Reduktion von .06 bei hoher Faktorkorrelation bis zu .309 bei geringer Faktorkorrelation im Populationsmodell). Desto geringer die Faktorkorrelation im Populationsmodell, desto größer wurde auch die Standardabweichung der Ladungen des missspezifizierten Modells (von .087 bei hoher Faktorkorrelation bis zu .324 bei geringer Faktorkorrelation im Populationsmodell); d.h., bei steigender Missspezifikation wichen die Verteilungseigenschaften der Faktorladungen im missspezifizierten Modell stärker von denen der definierten Faktorladungen in den Populationsmodellen ab.

3 Diskussion

3.1 Zusammenfassung der Ergebnisse

Im Rahmen einer Simulationsstudie auf Basis zweier definierter Faktorwerte der Individuen in einer Population wurde untersucht, inwieweit die Güte der Diagnostik beeinträchtigt wurde, sofern dichotome Diagnosen („krank“/positive Diagnose versus „gesund“/negative Diagnose), die auf Basis der höchsten Werte auf beiden definierten Faktorwerten vergeben wurden, fälschlicherweise auf Basis der höchsten Faktorwerte missspezifizierter Strukturgleichungsmodelle vergeben wurden. Diese Missspezifikation im Strukturmodell stellte ein missspezifiziertes einfaktorielles Modell im Vergleich zu einem obliquen zweifaktoriellen Populationsmodell dar. Im Rahmen einer Nebenfragestellung wurden außerdem dichotome Diagnosen, die auf Basis der höchsten Gesamtsummenwerte über alle Indikatoren hinweg gebildet wurden, hinsichtlich der diagnostischen Kennwerte mit den Diagnosen auf Basis der definierten Faktorwerte verglichen.

Die Ergebnisse zeigten, dass die unterschiedlich gewählten Basisraten für die Diagnosegebung im Rahmen des gewählten Designs den größten Effekt auf die Güte der Diagnostik auf Basis der Faktorwerte der korrekten wie auch der missspezifizierten Modelle und der Gesamtsummenwerte hatte. Einen ähnlich hohen Einfluss, wie sie die Faktorladungen auf die Diagnostik aus den Faktorwerten des korrekten und missspezifizierten Modells sowie aus den Gesamtsummenwerten hatte, hatte/n der Grad der Missspezifikation/die unterschiedlichen Populationsmodelle auf die Diagnostik aus den Faktorwerten des missspezifizierten Modells sowie auf die Diagnostik aus den Gesamtsummenwerten. Die genannten Parameter interagierten hinsichtlich des Einflusses auf die Diagnostik miteinander.

Die verschiedenen Basisraten hatten den höchsten Einfluss auf die Güte der Diagnostik sowohl anhand der Faktorwerte korrekter und missspezifizierter Modelle als auch anhand der Gesamtsummenwerte. Große Basisraten wirkten sich positiv auf die Rate der korrekt erkannten Kranken an allen Kranken (Sensitivität) sowie die Rate der korrekt erkannten Kranken an allen als krank Diagnostizierten (Positiver Prädiktionswert) aus. Umgekehrt waren kleine Basisraten hinsichtlich der Rate an korrekt als gesund erkannten Gesunden an allen Gesunden (Spezifität) sowie der korrekt erkannten Gesunden an allen als gesund Diagnostizierten (Negativer Prädiktionswert) vorteilhaft.

Die Befunde zeigten außerdem, dass höhere Faktorladungen in den Populationsmodellen zu besserer Diagnostik anhand aller Kennwerte führten. Die realistisch niedrig gewählten Faktorladungen im Rahmen des Studiendesigns beeinträchtigte die Güte der Diagnostik auf Basis der Faktorwerte selbst bei Anwendung korrekter Modelle stark. Diese Beeinträchtigung begann damit, dass die in den Populationsmodellen definierten Basisraten zur Diagnosegebung auf Basis der Faktorwerte durch die korrekten Modelle nicht reproduziert werden konnten. Die kleinen Basisraten wurden um bis zu 100 Prozentpunkte überschätzt, die großen um bis zu 25 Prozentpunkte unterschätzt. Diese Beeinträchtigung durch die realistischen Faktorladungen wirkte sich insbesondere bei kleinen Basisraten in klinischen Größenordnungen negativ auf die Rate der korrekt als krank diagnostizierten Fälle an allen positiven Diagnosen aus (Positiver Prädiktionswert; Worst-Case-Szenario 24%), aber auch auf die Rate der korrekt als krank erkannten Kranken an allen Kranken aus (Sensitivität; Worst-Case-Szenario 47%). Umgekehrt beeinträchtigten die realistisch hohen Faktorladungen in Kombination mit großen Basisraten bei Anwendung korrekter Modelle die Rate der korrekt als gesund diagnostizierten Fälle an allen negativen Diagnosen (Negativer Prädiktionswert; Worst-Case-Szenario 53%), aber auch die Rate der korrekt als gesund erkannten Gesunden an allen Gesunden (Spezifität; Worst-Case-Szenario 77%). Anhand des Exkurses konnte gezeigt werden, dass sich der Effekt der unterschiedlichen Basisraten auf die Güte der Diagnostik nivellierte, sobald unrealistisch hohe Ladungen für die Populationsmodelle verwendet wurden, aus denen die Daten generiert wurden.

Der Grad der Missspezifikation (operationalisiert an der Höhe der Faktorkorrelation im Populationsmodell versus eindimensionales missspezifiziertes Modell) im Vergleich zur korrekten Spezifikation hatte ebenfalls einen entscheidenden Einfluss auf die Güte der Diagnostik anhand der Faktorwerte des missspezifizierten Modells. Der Grad der Missspezifikation (Höhe der Faktorkorrelation) beeinträchtigte vor dem Hintergrund der niedrigen Faktorladungen in Kombination mit kleinen Basisraten die Rate der korrekt als krank erkannten Fälle an allen Kranken (Sensitivität), in Kombination mit großen Basisraten die Rate der korrekt als gesund erkannten Fälle an allen Gesunden (Spezifität). Eine ausgewogene Indikatorenaufteilung auf die Populationsfaktoren, welche mehr vom eindimensionalen missspezifizierten Modell abwich, führte zu besseren diagnostischen Kennwerten auf Basis der Faktorwerte missspezifizierter Modelle als eine unausgewogene Indikatorenaufteilung auf die beiden Populationsfaktoren. Dieser Befund wird unter 3.2. diskutiert.

Die Diagnostik auf Basis der Gesamtsummenwerte war der Diagnostik auf Basis der Faktorwerte der korrekten Modelle hinsichtlich der Rate der korrekt als krank erkannten Fälle an allen Kranken (Sensitivität) bei kleinen Basisraten und der Rate der korrekt als gesund erkannten Fälle an allen Gesunden (Spezifität) bei großen Basisraten unterlegen. Jedoch schnitt die Diagnostik auf Basis der Gesamtsummenwerte, welche sich vom missspezifizierten Modell nur durch die gleichen Gewichtungen der Indikatoren unterschied, marginal besser ab als die Diagnostik basierend auf den Faktorwerten des missspezifizierten Modells. Letzterer Befund wird unter 3.2 ausführlicher diskutiert.

3.2 Diskussion der Ergebnisse

Dass insbesondere die unter 3.1 genannten diagnostischen Kennwerte litten, erklärt sich, wie bereits unter 2.2.2 beschrieben, durch die Überschätzung kleiner Basisraten und die Unterschätzung großer Basisraten für die Diagnosen auf Basis der Faktorwerte korrekter Modelle aufgrund der realistisch niedrig gewählten Faktorladungen der Populationsmodelle. Als Konsequenz dieser Verzerrungen der Basisraten häuften sich bei kleinen Basisraten die falsch positiven Diagnosen und bei großen Basisraten die falsch negativen Diagnosen. Dementsprechend verhielten sich die diagnostischen Kennwerte. Bei der Diagnostik auf Basis der Faktorwerte missspezifizierter Modelle wurden die univariaten Basisraten genauso festgelegt wie die bivariaten Basisraten in den Populationsmodellen, sodass sich die Raten an falsch positiven und falsch negativen Diagnosen mehr ausglichen als bei den Faktorwerten der korrekten Modelle. Daraus ergaben sich einige der beschriebenen Unterschiede zwischen der Diagnostik auf Basis der Faktorwerte des missspezifizierten Modells und des korrekten Modells. Verstärkt wurden diese Unterschiede zusätzlich durch eine niedrigere Faktorkorrelation im Populationsmodell (höherer Grad an Missspezifikation) im Vergleich zu einer höheren.

Die Befunde zu den Basisraten können aufgrund des unterschiedlichen Designs und der unterschiedlichen Beschaffenheit der Indikatoren kaum mit den unter III. 4 beschriebenen Studien von Emons et al. (2007), Kruyen et al. (2012), Schönemann und Thompson (1996) sowie Taylor und Russell (1939) verglichen werden. Sowohl der vorliegenden Studie als auch den genannten Studien ist allerdings gemein, dass die Sensitivität eines diagnostischen Instruments unter realistischen Bedingungen bei kleinen Basisraten beeinträchtigt ist und bei großen Basisraten die Spezifität. Außerdem stimmen die Studien mit der vorliegenden insofern

überein, als dass eine geringe Reliabilität (niedrige Faktorladungen im Rahmen der vorliegenden Studie) oder niedrige Trennschärfen im Rahmen der drei erstgenannten Studien sowie eine geringe Validität (Missspezifikation im Strukturmodell als Verletzung der Konstruktvalidität bei der vorliegenden Studie und niedrige Kriteriumsvalidität bei Schönemann und Thompson sowie Taylor und Russell) hinsichtlich der Güte der Diagnostik problematisch sind. Im Rahmen des vorliegenden Designs interagierten die genannten Gütekriterien außerdem mit den Basisraten. Bei kleinen Basisraten wirkte sich eine geringe Reliabilität besonders negativ auf die Rate der korrekt als krank erkannten Kranken an allen als krank Diagnostizierten (Positiver Prädiktionswert) aus, eine geringe Konstruktvalidität vor allem auf die Rate der korrekt als krank erkannten Kranken an allen Kranken (Sensitivität). Umgekehrt beeinträchtigten große Basisraten in Kombination mit einer niedrigen Reliabilität vor allem die Rate der korrekt als gesund erkannt Gesunden an allen als gesund Diagnostizierten (Negativer Prädiktionswert), eine niedrige Konstruktvalidität vor allem die Rate der korrekt als gesund erkannten Gesunden (Spezifität) an allen Gesunden.

Eine unausgewogene Indikatorenaufteilung im Populationsmodell führte zu schlechterer Diagnostik auf Basis der Faktorwerte des eindimensionalen missspezifizierten Modells als eine ausgewogene Indikatorenaufteilung im Populationsmodell. Für diesen Befund können zwei Erklärungen angeführt werden. Die erste Erklärung betrifft die Faktorladungen. Little, Cunningham, Shahar, und Widaman (2002a) sowie Marsh und Hocevar (1988) konnten zeigen, dass eine unausgewogene Indikatorenaufteilung auf die Faktoren im Rahmen konfirmatorischer Faktorenanalysen zu instabileren Faktorlösungen wie auch zu höheren Standardabweichungen und Standardfehlern bei den Modellparameterschätzungen im Vergleich zu einer ausgewogenen Indikatorenaufteilung führten. Ersteres Muster zeigte sich auch an den unter 2.1.1 beschriebenen Eigenwerten der Kovarianzmatrix der Indikatoren, welches auch dazu führte, dass die Diagnostik bei Anwendung korrekter Modelle auf Populationsdaten, denen ein Populationsmodell mit ungleicher Indikatorenaufteilung anstatt ein Populationsmodell mit gleicher Indikatorenaufteilung zugrunde lag, marginal schlechter war[51]. Mit letzterem Argument von Little et al. und Marsh und Hocevar zu den Standardabweichungen der Faktorladungen kann der Befund zur Diagnostik bei Anwendung missspezifizierter Modelle erklärt werden. Der Befund, dass eine unausgewogene Indikatorenaufteilung im Vergleich zu

[51]Die Eigenwerte und Eigenvektoren der Matrix der Indikatoren bestimmt die Schätzung der Faktorladungen, welche wiederum die Schätzung der Bartlett-Faktorwerte bestimmt (Mulaik, 2009).

einer ausgewogenen zu höheren Standardabweichungen der Faktorladungen führt, trat auch bei der vorliegenden Studie bei den missspezifizierten Modellen auf (der Standardfehler der Faktorladungen hingegen ist aufgrund der Populationssimulation kein Argument). Von den Faktorladungen abhängig ist wiederum die Validität der Faktorwerte (für den Begriff siehe Grice [2001a, 2001b]). Hinsichtlich der Verteilung der aus den missspezifizierten berechneten Faktorwerte selbst zeigte sich kein Unterschied zwischen der unausgewogenen und der ausgewogenen Indikatorenaufteilung im Populationsmodell. Zu einem Vergleich der Schätzung der Faktorwerte bei unausgewogener versus ausgewogener Indikatorenaufteilung gibt es nach Kenntnisstand der Autorin noch keine Befunde. Als zweite Erklärung für das oben genannte Ergebnis zur Diagnostik auf Basis der Faktorwerte eines missspezifizierten Modells bei ungleicher Indikatorenaufteilung auf die Faktoren im Populationsmodell kann ein Befund zum Raschmodell angeführt werden. Stelzl (1979) stellte anhand von simulierten Stichproben fest, dass der Likelihood-Quotiententest von Andersen (1973) nicht sensitiv dafür war, Modellgültigkeit des Raschmodells bei einem Test abzulehnen, der aus zwei Itemgruppen bestand. Diese Itemgruppen erfassten jeweils eine eigene latente Dimension, die Items waren also heterogen. Dem eindimensionalen missspezifizierten Modell im Rahmen des vorliegenden Designs lagen ebenfalls zwei Populationsdimensionen zugrunde, weshalb die Faktorladungen der Items umso heterogener ausfielen, desto niedriger die Korrelation der Populationsfaktoren war (desto höher der Grad der Missspezifikation). Die beiden Gruppen, die aus den unterschiedlich hohen Faktorwerten der Individuen des missspezifizierten Modells gebildet wurden, stellten die Gruppen mit positiven und negativen Diagnosen dar. Formann (1981) sowie Heene, Kyngdon, und Sckopke (2016) erklärten das Versagen des Andersen-Tests so, dass sich in jeder der beiden Itemgruppen Personengruppen mit unterschiedlichen Parameterkombinationen hinsichtlich der zwei Dimensionen befanden. Es gab eine Personengruppe, die hinsichtlich ihres Antwortverhaltens auf einer Dimension einen höheren Wert besaß als auf der anderen Dimension, bei einer zweiten Personengruppe verhielt es sich umgekehrt und die dritte Personengruppe besaß gleich hohe Werte auf beiden Dimensionen. Dadurch, dass in beiden Itemgruppen Personen mit allen drei Kombinationen waren, kam es zu einer Kompensation der Heterogenität der Items. Die Personengruppe mit gleich hohen Werten auf beiden Dimensionen homogenisierte zusätzlich die Heterogenität der Items. Dieser Kompensationseffekt wurde größer, wenn jede Dimension von gleich vielen Indikatoren gemessen wurde; hingegen wurde der Kompensationseffekt bei unterschiedlich vielen Items pro Dimension geringer (Heene et al., 2016). Vor diesem Hintergrund, dass die unausgewogene

Indikatorenaufteilung im Populationsmodell im Vergleich zur ausgewogenen zu schlechterer Diagnostik auf Basis der Faktorwerte des missspezifizierten Modells führte trotz dass das eindimensionale missspezifizierte Modell näher am Populationsmodell mit unausgewogener Indikatorenaufteilung als mit ausgewogener Indikatorenaufteilung lag, ist der Befund zu den Fit-Indizes in Studie 1 alarmierend: Alle drei untersuchten Fit-Indizes zeigten bei Anwendung des eindimensionalen missspezifizierten Modells auf die Stichprobendaten eine bessere Modellpassung an, wenn das Populationsmodell eine unausgewogene Indikatorenaufteilung hatte im Vergleich zu einer ausgewogenen. Hinsichtlich der Diagnostik auf Basis der Faktorwerte zeigte sich jedoch genau das Gegenteil, die ausgewogene Indikatorenaufteilung führte zu besserer Diagnostik auf Basis der Faktorwerte des missspezifizierten Modells als die unausgewogene.

Diese Heterogenität der Faktorladungen des missspezifizierten Modells leitet zur Erklärung des Befundes über, nach dem die Gesamtsummenwerte zu geringfügig besserer Diagnostik führten als die Faktorwerte des missspezifizierten Modells. Im missspezifizierten Modell sanken die Faktorladungen einerseits im Vergleich zu den festgelegten Faktorladungen in den Populationsmodellen, andererseits wurden die Ladungen mit dem Grad der Missspezifikation (Höhe der Faktorkorrelation im Populationsmodell) heterogener. Da die Ladungen primär verantwortlich sind für die Schätzung der Bartlett-Faktorwerte (Grice, 2001b), kann vermutet werden, dass letztere beide Befunde zu niedrigerer Validität der geschätzten Faktorwerte eines Modells führten. Die Ergebnisse des verwendeten Designs legen die Schlussfolgerung nahe, dass die Gesamtsummenwerte zu besserer Diagnostik führten als die Faktorwerte eines eindimensionalen missspezifizierten Modells. Der Befund steht in Einklang mit DiStefano et al. (2009), Dobie, McFarland, und Long (1986) sowie Kukuk und Baty (1979), nach denen die Summenwerte zu besseren Trefferquoten führten als die Faktorwerte, sofern es sich um ein Konstrukt mit verschiedenen Attributen handelte, also letztendlich eine mehrdimensionale Skala, die jedoch eindimensional erfasst wurde. Dieser Befund kann allerdings nicht verallgemeinert werden, da er aus der Nicht-Spezifizierung der Mehrdimensionalität im Strukturmodell und der damit einhergehenden Reduktion und größeren Streuung der Faktorladungen im missspezifizierten Modell einherging.

Hinzu kommt, dass die Bartlett-Faktorwerte, die im Rahmen der Studie verwendet wurden, zwar, im Gegensatz zu beispielsweise den Regressionsfaktorwerten, erwartungstreue Schätzer für die True Scores sind (für die Herleitung siehe Lawley & Maxwell, 1971), die Bartlett-Faktorwerte bei obliquen Faktoren aber noch abhängiger von den Faktorladungen sind

bzw. mehr mit den Ladungen variieren als die Regressionsfaktorwerte (Beauducel, 2005). Dies liegt daran, dass sich die Ähnlichkeit zwischen der Mustermatrix der Faktorladungen und der Matrix der True Scores in den Bartlett-Faktorwerten widerspiegelt. In den Regressionsfaktorwerten spiegelt sich hingegen die Ähnlichkeit zwischen der Strukturmatrix der Faktorladungen und der Matrix der True Scores wider (S. 147).

3.3 Implikationen

3.3.1 Implikationen für weitere Forschungen

Vor dem Hintergrund der verschiedenen Möglichkeiten, Faktorwerte zu berechnen, wäre von Interesse, wie die Güte der Diagnostik ausfallen würde, wenn Diagnosen basierend auf den Regressionsfaktorwerten im Kontext von Strukturgleichungsmodellen oder konfirmatorischen Faktorenanalysen getroffen werden. Im Rahmen dieser Studie wurden die Diagnosen basierend auf den Bartlett-Faktorwerten getroffen; diese sind, wie bereits erläutert, erwartungstreue Schätzer für die wahren Faktorwerte (Lawley & Maxwell, 1971). In der angewandten Forschung werden aber, wie bereits unter II. 3 erwähnt, meist die Regressionsfaktorwerte zur Bestimmung der individuellen Ausprägungen auf den latenten Variablen herangezogen (Grice, 2001b). Aufgrund der Tatsache, dass die Regressionsfaktorwerte nicht erwartungstreu sind (Lawley & Maxwell), ist zu vermuten, dass die Diagnostik auf Basis der Regressionsfaktorwerte schlechter ausfallen würde als auf Basis der Bartlett-Faktorwerte. Andererseits werden die Regressionsfaktorwerte, wie oben beschrieben, bei obliquen Faktoren weniger durch die Ladungen und somit die Reliabilität beeinflusst als die Bartlett-Faktorwerte, was wiederum ein Pluspunkt für die Diagnostik auf Basis der Regressionsfaktorwerte im Rahmen des vorliegenden Designs sein könnte, der zu untersuchen ist.

Im Zuge weiterer Forschungen wäre außerdem zu untersuchen, inwieweit sich die Diagnostik auf Basis der Faktorwerte und auf Basis der Gesamtsummenwerte verschlechtert, wenn anstatt einer Populationssimulation die Methode des Resampling (Stichprobenziehung; vgl. Carsey & Harden, 2014) verwendet werden würde, zumal in der angewandten Forschung einerseits eine eingeschränkte Anzahl von Fällen vorliegt, andererseits ein Testverfahren in der Praxis einer geringen Stichprobengröße standhalten muss.

Im Rahmen des vorliegenden Designs wurde Heterogenität der Faktorladungen in den Populationsmodellen simuliert, wobei für alle simulierten Fälle die gleichen Faktorladungen verwendet wurden. Diese Heterogenität der Faktorladungen vergrößerte sich noch zusätzlich durch die Nicht-Spezifizierung der Mehrdimensionalität im Strukturmodell. Dass – durch den Grad der Missspezifikation induziert – heterogenere Faktorladungen zu schlechteren Schätzungen der Faktorwerte führten als homogene, wurde bereits unter 3.2 diskutiert. Im Rahmen weiterer Simulationsforschung ist daher die Fragestellung von großem Interesse, ob und inwieweit heterogene Faktorladungen für sich und nicht durch eine Missspezifikation verursacht, die Validität der Faktorwerte der Individuen und die Validität diagnostischer Entscheidungen basierend auf den Faktorwerten der Individuen beeinträchtigen würden. Von Kelderman und Molenaar (2007) wurde bereits analytisch und anhand einer Simulationsstudie gezeigt, dass individuell heterogene Faktorladungen die Validität der individuellen Faktorwerte beeinträchtigen.

Die Studien von Emons et al. (2007) sowie Kruyen et al. (2012) konnten zeigen, dass die Güte der Diagnostik insbesondere leidet, wenn Kurzskalen im Vergleich zu Langskalen (mindestens 20 Indikatoren) verwendet wurden. Estabrook und Neale (2013) konnten im Rahmen einer Simulationsstudie mit mehrdimensionalem Populationsmodell zeigen, dass im Rahmen des Designs der Autoren die Itemanzahl sogar noch wichtiger für die Validität der Faktorwerte war als die Höhe der Faktorladungen. Lawley und Maxwell (1971) zeigten formal-analytisch, dass sich die Verteilungseigenschaften der geschätzten Faktorwerte mehr und mehr denen der True Scores annähern, wenn viele und reliable Indikatoren verwendet werden. Anknüpfend an das Design der vorliegenden Studie mit niedriger Reliabilität (niedrigen Faktorladungen) und niedriger Konstruktvalidität (durch Nicht-Spezifikation der Mehrdimensionalität in der Faktorenstruktur) wäre von Interesse, zu untersuchen, ab wie vielen Indikatoren pro Faktor und insgesamt sich die Güte der Diagnostik in klinisch bedeutsamem Ausmaß verbessern würde.

Die Ergebnisse zeigten außerdem, dass die Faktorenunbestimmtheit als Konsequenz der realistisch niedrig gewählten Faktorladungen hoch war. Dadurch werden zwei Implikationen deutlich. Erstens drängt sich die Frage auf, ob die Faktorenunbestimmtheit im Kontext klassischer Verfahren bisher genug Beachtung fand. Zweitens schrieben Grice (2001b), Maraun (1996a), Schönemann und Steiger (1978) sowie Steiger (1979), dass die Unbestimmtheit der Faktoren im Kontext klassischer Verfahren vor allem problematisch würde, sobald kriterielle Aussagen aus den Faktorwerten getroffen werden würden. Daher wäre eine nächste

Forschungsfrage im Kontext einer Simulationsstudie, inwieweit Diagnosen, die auf Basis einer oder mehrerer exogener latenter Variablen getroffen werden, eine endogene latente Variable oder sogar Diagnosen basierend auf einer endogenen latenten Variablen vorhersagen können. Eine derartige angewandte Forschungsfrage beträfe zum Beispiel die Untersuchung der Passung der Diagnosen auf Basis verschiedener Testverfahren, unterschiedlicher methodischer Herangehensweisen oder auch die Überprüfung der Veränderung von individuellen Diagnosen über die Zeit[52].

3.3.2 Empfehlungen für die Testkonstruktion

Die Ergebnisse der Simulationsstudie insgesamt (insbesondere die Ergebnisse aus dem Exkurs zu den Populationsmodellen mit extrem hohen Ladungen) zeigen aber auch eine wichtige gute Nachricht: Eine gute Testkonstruktion im Sinne der Gütekriterien führt zu diagnostischem Erfolg. Die Ergebnisse zu den missspezifizierten Modellen zeigen außerdem im Speziellen, wie wichtig eine gute Theorie im Sinne einer präzisen und trennscharfen Begriffsbestimmung bzw. ein empirisch gut abgesichertes Konstrukt ist, bevor überhaupt damit begonnen wird, ein diagnostisches Verfahren zu konstruieren. Sofern ein zu messendes Konstrukt aus verschiedenen Attributen besteht, die aus Gründen der Inhaltsvalidität niemals vernachlässigt werden sollten, kann als Empfehlung für die Testkonstruktion ausgesprochen werden, den Messgegenstand in möglichst homogene Teilkonstrukte aufzuteilen[53]. Diese homogenen Teilkonstrukte sollten dann von statistisch eindimensionalen und möglichst reliablen Items gemessen werden. Höhere Kommunalitäten erhöhen außerdem die Stabilität der Faktorladungen und somit auch der Faktoren selbst (Cliff & Hamburger, 1967; Cliff & Pennell, 1967; Pennell, 1968) und damit verbessert sich wiederum die Faktorwerteschätzung, wie Estabrook und Neale (2013) zeigen konnten. Basierend auf einer hohen Reliabilität sollte dann die faktorielle Struktur der Teilkonstrukte zueinander möglichst valide spezifiziert werden.

Die Ergebnisse der vorliegenden Simulationsstudie legen außerdem umfassendere Untersuchungen zur Validität neu entwickelter, aber auch bestehender Testverfahren

[52]Letzteres wurde bereits von Kruyen, Emons, und Sijtsma (2014) vor dem Hintergrund von Kurzskalen anhand von Raschmodellen untersucht.

[53]Dies entspricht dem unter III. 2.2 genannten Konzept der Mehrdimensionalität zwischen Items, die an der betreffenden Stelle von der Mehrdimensionalität innerhalb der Items (Verletzung der Einfachstruktur) abgegrenzt wurde.

insbesondere im Kontext kleiner Basisraten nahe. Korrelationen der beobachteten Variablen zwischen zwei verschiedenen Testverfahren sind nicht ausreichend, vor allem, wenn es um die Schätzung der individuellen Ausprägungen auf den latenten Variablen anhand der Faktorwerte geht. Da Simulationsstudien nicht möglich sind, können als Kriterien für die Evaluation der Diagnostik auf Basis von Faktorwerten die Faktorwerte eines anderen Testverfahrens oder ein Expertenurteil dienen. Weiters zeigen die Ergebnisse insbesondere zu den Graden der Missspezifikation auf, wie wichtig nicht nur gute faktorielle Validierungen hinsichtlich der Faktorwerte im Kontext neu entwickelter Testverfahren und Fragebögen sind, sondern auch andere Arten der Validierung, zum Beispiel kriterielle Validierungen. Außerdem verbessern Methoden, die über faktorielle und kriterielle Validierungen hinausgehen, die Güte der Diagnostik, so zum Beispiel Kreuzvalidierungen (vgl. Carsey & Harden, 2014, S. 255).

Es ist offensichtlich, dass sich die Basisraten unter den untersuchten Einflussfaktoren auf die Güte der Diagnostik in der Praxis nicht verändern lassen. Eine mögliche Lösung aus dem Dilemma mit den kleinen Basisraten stellt möglicherweise dar, was bereits Emons et al. (2007) vorschlugen, nämlich, die Diagnostik in kleinere Teile an Information aufzuteilen. Möglicherweise kommt dafür die Beurteilung nach der Präsenz oder Absenz eines Symptoms, das für das Vorliegen einer psychischen Störung notwendig ist, in Frage. Ein derartiges Vorgehen würde die Basisraten für die diagnostischen Entscheidungen vergrößern und somit die Trefferquoten erhöhen, wenn es um möglichst hohe Raten an korrekt Positiven an allen Positiven (Sensitivität) und an korrekt Positiven an allen als positiv Diagnostizierten (Positiver Prädiktionswert) geht. Gleichzeitig ist ein Symptom an sich homogener als der Symptompool, der zur Diagnose führt, und kann somit auch homogener und reliabler erfasst werden als das Syndromkomplex. Dies stellt einen Vorschlag dar, dessen diagnostische Sinnhaftigkeit und Praktikabilität im Rahmen weiterer Forschung zu überprüfen ist.

3.4 Limitationen

Die Tatsache, dass im Rahmen der vorliegenden Simulationsstudie einerseits die Voraussetzung der multivariaten Normalverteilung für die erzeugten Indikatoren erfüllt war, andererseits die Indikatoren auch dem Intervallskalenniveau genügten, findet sich im Rahmen der angewandten Forschung selten wieder. Daher wäre von Interesse für die angewandte Forschung, wie die Güte der Diagnostik ausfällt, wenn fälschlicherweise von multivariater Normalverteilung und Intervallskalenniveau ausgegangen wird, zumal die Items vieler

Fragebögen meist nur dem Ordinalskalenniveau genügen. Sofern bekannt, kann das Ordinalskalenniveau in den beobachteten Variablen jedoch durch eine andere Schätzmethode kompensiert werden (siehe beispielsweise Browne, 2011) und eine Verletzung der multivariaten Normalverteilung durch robuste Korrekturen der Schätzmethoden (vgl. Finney & DiStefano, 2006; Satorra & Bentler, 1994; Yuan & Bentler, 1998, 2000).

Die Missspezifikation im Rahmen dieser Simulationsstudie stellte eine Unterparametrisierung des Modells dar. Im Zuge weiterer Studien mit unterschiedlichen Basisraten wäre zu überprüfen, wie die diagnostischen Kennwerte ausfallen, wenn basierend auf den Faktorwerten eines fälschlicherweise überparametrisierten Modells diagnostiziert wird.

Außerdem stellte die Missspezifikation im Strukturmodell im Rahmen dieser Studie eine schwerwiegende Missspezifikation dar. Es wäre von Interesse, zu überprüfen, inwieweit sich Diagnosen, die auf Basis von Faktorwerten mit Missspezifikationen im Messmodell getroffen werden, auf die diagnostischen Kennwerte auswirkt. Die sehr häufige Missspezifikation in Form von nicht-spezifizierten korrelierten Messfehlern (Heene et al., 2012; Savalei, 2012; siehe III. 2.2) könnte dafür ein geeignetes Design darstellen.

Ferner wurden im Rahmen dieses Designs Diagnosen auf Basis des Top-Down-Prinzips vergeben. Im Zuge weiterer Forschung zu Diagnosen aus Strukturgleichungsmodellen heraus wäre von Interesse, ob Diagnosen basierend auf einem Cut-Off zu anderen Ergebnissen hinsichtlich der diagnostischen Kennwerte führen als auf Basis einer Top-Down-Entscheidung. Im Rahmen des probabilistischen Designs von Kruyen et al. (2012) zeigten sich kaum Unterschiede hinsichtlich der Güte der Klassifikation zwischen Cut-Off-basierten und Top-Down-Klassifikationen.

Weiters wurde im Rahmen des beschriebenen Designs nur eine positive Diagnose vergeben, wenn das Individuum auf beiden Populationsfaktoren (bzw. auf beiden Faktoren im Rahmen des korrekten Modells) innerhalb der höchsten Faktorwerte je nach untersuchter Basisrate rangierte. Dies entspricht einer konjunktiven Entscheidungsstrategie (Amelang & Schmidt-Atzert, 2006, S. 399). Es wäre zu überprüfen, wie die Diagnostik im Vergleich zum vorliegenden Design ausfallen würde, wenn ein hoher Wert auf einem der beiden Populationsfaktoren für eine positive Diagnose ausreichen würde, also auf Basis einer kompensatorischen Entscheidungsstrategie (S. 399) diagnostiziert werden würde. Eine kompensatorische Entscheidungsstrategie wurde im Rahmen des vorliegenden Designs nicht untersucht, da diese hinsichtlich der Güte der Diagnostik mit der Modellbedingung der unausgewogenen Indikatorenaufteilung konfundiert gewesen wäre.

Die konjunktive Diagnosestrategie (Amelang & Schmidt-Atzert, 2006, S. 399) führt zu einer weiteren Limitation der Studie hinsichtlich der unterschiedlichen Cut-Off-Werte für die Diagnosevergabe. Die unterschiedlich hoch definierten Faktorkorrelationen zwischen den True Scores zwischen den einzelnen Bedingungen führten dazu, dass mit steigender Korrelation zwischen den True Scores auch die Anzahl der Fälle, die hohe Werte auf beiden Faktoren hatte, stieg. Das heißt, mit höherer Faktorkorrelation stieg die bivariate Basisrate für die Diagnosegebung. Diese Tatsache bot zwei Möglichkeiten: entweder die gleichen univariaten Cut-Offs pro Faktor zu verwenden und damit die Vergleichbarkeit der Basisraten über die Modellbedingungen hinweg einzuschränken, oder, unterschiedliche univariate Cut-Off-Werte je nach Korrelation der True Scores zu definieren, um unabhängig von der Faktorkorrelation dieselben bivariaten Basisraten für die Diagnosen zu erhalten. Aufgrund dessen, dass bereits mehrere Studien den Einfluss der unterschiedlichen Basisraten auf die Güte von diagnostischen Entscheidungen bestätigt hatten (Emons et al., 2007; Kruyen et al., 2012; Schönemann, 1997; Schönemann & Thompson, 1996; siehe III. 4), wurde die Konstanthaltung der Basisraten über die Bedingungen hinweg als wichtiger für die Studie erachtet als die Konstanthaltung der univariaten Cut-Off-Werte.

Die Basisraten für die Diagnosen wurden aus Gründen der Verfügbarkeit verlässlicher Zahlen nach den 12-Monats-Prävalenzen einer EU-Studie konstruiert (Wittchen et al., 2011). Eine ähnlich gute und so groß angelegte Studie fand sich für die Punktprävalenzen der psychischen Störungen nicht. Jedoch wären aufgrund des querschnittlichen Designs Punktprävalenzen zur Vergabe der Diagnosen angemessener gewesen als 12-Monats-Prävalenzen.

Des Weiteren war bei der Konstruktion des Untersuchungsdesigns geplant, die großen Basisraten nicht basierend auf den Prävalenzen für Komorbiditäten und Lebenszeitprävalenzen in der Allgemeinbevölkerung oder Basisraten der Eignungsdiagnostik, sondern basierend auf den Diagnose-Raten psychologischer Beratungsstellen zu vergeben. Es ist zu vermuten, dass die Rate an positiven Diagnosen im Rahmen der Diagnostik durch beispielsweise Beratungsstellen oder auch andere Erstanlaufstellen deutlich höher ausfällt als die Rate an positiven Diagnosen in der Allgemeinbevölkerung. Dies hat den Grund, dass eine hohe Rate tatsächlich kranker Menschen oder auch subklinisch kranker Menschen derartige Anlaufstellen ansteuern, um sich Hilfe zu suchen, wohingegen gesunde Menschen derartige Beratungsstellen kaum aufsuchen. Der Versuch, die großen Basisraten an den Prävalenzen von Erstanlaufstellen zu orientieren, scheiterte daran, dass kaum verlässliche Zahlen dazu aufzufinden waren, die

dokumentiert hätten, wie viele der Personen, die eine derartige Beratungsstelle aufsuchten, auch tatsächlich eine positive Diagnose bekamen. Sofern die Vermutung naheliegt, dass die betreffenden Personen tatsächlich erkrankt sind, werden sie meist an Psychologische Psychotherapeutinnen und Psychologen Psychotherapeuten oder Psychiaterinnen und Psychiatern weiter verwiesen, die dann auch Diagnosen erstellen, sofern eine Psychotherapie indiziert ist. Aufgrund dieses Schritts ist dann allerdings nicht mehr nachvollziehbar, wie viele der anfänglich Hilfe-Suchenden sich tatsächlich in Behandlung begeben hat und wie viele nicht. Wittchen und Jacobi (2001) schätzen auf Basis ihrer Studie zur deutschen Allgemeinbevölkerung, dass nur etwa ein Drittel aller tatsächlich Kranken mindestens einmal im Verlauf des Lebens behandelt wird.

Eine weitere Limitation der vorliegenden Studie stellt die Möglichkeit der Evaluierung von Diagnosen anhand der klassischen Methoden generell in Frage. Diagnosen stellen formative[54] Modelle dar, da die Präsenz von bestimmten Symptomen zur Diagnose führt bzw. die Abwesenheit dieser Symptomen zur Einordnung in die Gruppe der Gesunden führt. Insbesondere kann sogar die Präsenz unterschiedlicher Symptomatik zur selben Diagnose führen (man beachte beispielsweise die 256 Kombinationen an Symptomen, die nach dem ICD-10 [WHO, 1993] zur Diagnose „Borderline-Störung“ führen; Anm. der Autorin). Die klassische Testtheorie fußt allerdings auf reflexiven Messmodellen, nach denen die individuelle Ausprägung auf der latenten Variablen für die Ausprägungen auf den beobachteten Variablen verantwortlich ist (Bühner, 2011, S. 21). Demnach ändert eine Änderung auf einem Indikator oder das Weglassen oder Hinzufügen eines Indikators die Ausprägung auf der latenten Variablen nicht, da alle Indikatoren das gleiche messen und positiv miteinander korrelieren. Diese positive Korrelation wird Null, sobald die latente Variable spezifiziert wurde, da Unterschiede der Personen auf dieser Variable die Ursache für die Korrelationen der Items sind. Im Rahmen formativer Messmodelle jedoch müssen die Indikatoren nicht (positiv) miteinander korrelieren (Bühner, 2011; Reinecke, 2014). Außerdem führt eine Änderung auf der Ebene der manifesten Variablen zu einer Änderung hinsichtlich der Ausprägung auf dem Konstrukt und kann im klinischen Fall über das Vorliegen oder Nicht-Vorliegen einer Diagnose entscheiden. Die Tatsache, dass die Kriterien für einige Störungen, wie die „Borderline-Störung“ oder das „AD(H)S“-Syndrom, ungenau definiert sind, erschwert die Diagnostik erheblich. Eine weiterführende Fragestellung, die von hoher Relevanz für die klinische Forschung und

[54]Formative Modelle wurden erstmals von Curtis und Jackson (1962) beschrieben.

Anwendung ist, stellt daher dar, welche Auswirkungen es auf die diagnostischen Kennwerte hätte, wenn Diagnosen auf Basis reflektiver Modelle gegeben werden, die auf Basis formativer Modelle konstruiert wurden. Das reflektive Modell stellt in diesem Kontext eine Missspezifikation dar. MacKenzie, Podsakoff, und Jarvis (2005) untersuchten bereits den Einfluss von Missspezifikationen in Form von reflektiven Messmodellen im Gegensatz zu formativen Messmodellen auf die Schätzung der Modellparameter, also auf die Reliabilität eines Modells. Die Autoren zeigten, dass diese Form der Missspezifikation zu fälschlicherweise signifikanten Parameterschätzungen und zur Überschätzung der Varianz der latenten Variablen führte. Studien zu den Auswirkungen dieser Form der Missspezifikation auf die Validität eines Modells, zum Beispiel auf die Güte von diagnostischen Entscheidungen aus den Faktorwerten heraus, liegen nach Kenntnisstand der Autorin noch keine vor.

Im folgenden Kapitel VI werden die Hauptergebnisse der vorliegenden Dissertation noch einmal zusammengefasst, bevor über die Arbeit reflektiert wird und deren Relevanz für Wissenschaft und Praxis herausgearbeitet wird.

VI ALLGEMEINE DISKUSSION

1 Zusammenfassung der Ergebnisse

Das Ziel dieser Arbeit war, die Auswirkungen von Missspezifikationen im Rahmen der linearen Strukturgleichungsmodellierung in Form einer nicht-spezifizierten zweiten latenten Dimension zu untersuchen. Die untersuchten Auswirkungen bezogen sich auf die Sensitivität der Fit-Indizes für diese Art und die unterschiedlichen Grade der Missspezifikation sowie auf die psychometrischen Auswirkungen hinsichtlich von Diagnosen, die auf Basis der Bartlett-Faktorwerte eines missspezifizierten Modells getroffen wurden. Das Design der nicht-spezifizierten Mehrdimensionalität im Strukturmodell wurde gewählt, da Mehrdimensionalität ein allgegenwärtiges Problem der Psychometrie darstellt (Heene et al., 2011; Little et al., 2002b; Savalei, 2012; siehe III. 2), welche auf eindimensionalen Konstrukten fußt, und die lineare Strukturgleichungsmodellierung als Datenanalysemethode sowie als Validierungsmethode für Testverfahren und Fragebögen sehr beliebt ist ("Datenbanksegment PSYNDEX Tests," 2013; MacCallum & Austin, 2000; Reinecke, 2014; Tremblay & Gardner, 1996; siehe I. und III. 4). Der Schweregrad der Missspezifikation wurde durch die Höhe der Faktorkorrelation und die (Un-)Ausgewogenheit der Indikatoren pro latenter Variable im Populationsmodell variiert. Es wurde zum einen überprüft, ob die weit verbreiteten Fit-Indizes CFI, RMSEA und SRMR diese Art und die unterschiedlichen Schweregrade der Missspezifikation anhand der Cut-Off-Kriterien nach Hu und Bentler (1998, 1999) bei Anwendung auf aus den Populationsmodellen gezogene Stichproben anzeigen würden. Zum anderen wurde geprüft, inwieweit die Güte der Diagnostik darunter leiden würde, wenn dichotome Diagnosen auf Basis der Bartlett-Faktorwerte aus dem missspezifizierten einfaktoriellen Modell vergeben wurden, die wahren Diagnosen jedoch auf Basis der True Scores zweier Populationsfaktoren gegeben wurden, aus denen heraus die Populationsdaten simuliert wurden. Die Diagnostik erfolgte auf Basis der Bartlett-Faktorwerte (Bartlett, 1937) nach einer hinsichtlich der beiden Populationsfaktoren konjunktiven Entscheidungsstrategie (Amelang & Schmidt-Atzert, 2006, S. 399) und anhand des Top-Down-Prinzips (Gatewood et al., 2016, S. 662). Für die Diagnosen wurden unterschiedliche Basisraten verwendet.

Anhand der ersten Simulationsstudie konnte gezeigt werden, dass die Fit-Indizes CFI, RMSEA und SRMR korrekte Modelle anhand der Daumenregeln nach Hu und Bentler (1998, 1999) als korrekt indizierten. Allerdings zeigten die drei Fit-Indizes hinsichtlich der missspezifizierten Modelle ein heterogeneres Bild. Die Fit-Indizes RMSEA und SRMR zeigten die Missspezifikation anhand der Cut-Offs nicht an, und dies selbst bei im Rahmen der angewandten Forschung hohen Faktorladungen. Der CFI zeigte die Modellabweichung mit Ausnahme der Bedingungen mit der hohen Faktorkorrelation im Populationsmodell (geringe Missspezifikation) an. Allerdings zeigte der CFI im Rahmen dieses Designs ein anderes Muster als das von Heene et al. (2011) beschriebene: Bei hohen Faktorladungen zeigten die Werte des CFI im Mittel bei der vorliegenden Studie eine höhere Sensitivität gegenüber der Missspezifikation als bei typischen Ladungen. Als Erklärung für diesen Unterschied wird vermutet, dass die nonlineare Teststatistik des CFI auch abhängig von der Art der Missspezifikation ist, wie von Curran et al. (2002) angenommen wurde. Die „two-index strategy“ funktionierte nur in Kombination mit dem CFI und dies auch nur bei mittlerem und hohem Grad an Missspezifikation. Im Gegensatz zu den Fit-Indizes konnte der χ^2-Test jeglichen Grad an Missspezifikation als Modellabweichung identifizieren.

Im Rahmen der zweiten Simulationsstudie wurde gezeigt, dass die unterschiedlichen Basisraten, die Reliabilität (Höhe der Faktorladungen) und die Validität (korrekte Spezifikation) eines Strukturgleichungsmodells sowie Interaktion dieser Parameter für die Güte der Diagnostik auf Basis der Faktorwerte einen bedeutenden Einfluss hatten, wobei die Basisraten im Rahmen des verwendeten Designs den größten Effekt hatten. Die Güte der Diagnostik auf Basis der Bartlett-Faktorwerte wurde selbst bei Anwendung eines korrekten Modells durch eine niedrige Reliabilität stark beeinträchtigt. Diese Beeinträchtigung verstärkte sich bei kleinen Basisraten in klinischen Größenordnungen und wirkte sich insbesondere hinsichtlich der Erkennung von Krankheit (Sensitivität) und der Korrektheit der Diagnose (Positiver Prädiktionswert) negativ aus. Wenn gleichzeitig auch die Konstruktvalidität nicht gegeben war (unterschiedliche Schweregrade der Missspezifikation), beeinträchtigte dies die Wahrscheinlichkeit, Krankheit zu erkennen (Sensitivität) bei kleinen Basisraten und die Wahrscheinlichkeit, Gesundheit zu erkennen (Spezifität) bei großen Basisraten zusätzlich.

Gleichzeitig zeigten die Ergebnisse aber auch, dass ein hinsichtlich der Gütekriterien sehr gut konstruierter Test zu validen diagnostischen Entscheidungen führt. Wenn ein hoch reliables und valides Modell für die diagnostischen Entscheidungen aus den Faktorwerten

verwendet wurde, hing die Güte der Diagnostik anhand aller Kennwerte kaum noch von den unterschiedlichen Basisraten ab. Die Gesamtsummenwerte waren den Faktorwerten des korrekt spezifizierten/konstruktvaliden Modells unterlegen, wenn die Reliabilität sehr hoch war. Wenn das Modell nicht konstruktvalide bzw. missspezifiziert war, führten die Gesamtsummenwerte zu besserer Diagnostik als die Faktorwerte, was auf die Art der Missspezifikation zurückzuführen ist.

2 Kritische Reflexion der eigenen Arbeit

2.1 Stärken der Arbeit

Die vorliegende Dissertation hat drei große Stärken, welche im Folgenden herausgearbeitet werden. Erstens besticht das Design der Simulationsstudien durch seine in jeder Hinsicht äußerst realistische Konstruktion. Dies betrifft einerseits die Modellbedingungen, andererseits den Prozess der Datengenerierung. Als Konsequenz dieses Forschungsdesigns, welches sich über beide Studien zieht, zeichnet sich die Arbeit durch ihre hohe Struktur und ihren roten Faden aus. Der Vergleich des Ist-Zustandes der gängigen Forschungspraxis mit dem Idealzustand erlaubt konkrete Handlungsempfehlungen für die Testkonstruktion.

Neben der Tatsache, dass psychologische Phänomene in ihrer Mehrdimensionalität oft nicht erfasst werden (können) und sich diese Arbeit den Konsequenzen dieses Problems annimmt, wurde als Datenanalysemethode für die berichteten Studien ein Verfahren gewählt, das zunehmend an Popularität in Grundlagen-, wie in angewandten Forschungsbereichen der Psychologie und verwandter Disziplinen gewinnt (MacCallum & Austin, 2000; Reinecke, 2014; Tremblay & Gardner, 1996; siehe I). Außerdem wurden – im Vergleich zu den meisten Studien, die die Auswirkungen von Modellabweichungen austesteten – realistisch hohe (Peterson, 2000) und realistisch heterogene (Buzick, 2010) Faktorladungen als inzidentelle Modellparameter gewählt (siehe III. 3 und 5 sowie IV. 1 und V. 1). Ebenso ist die Indikatorenanzahl typisch (Peterson, 2000; Shrout & Yager, 1989) und eine oblique Faktorenstruktur in der angewandten psychologischen Forschung häufiger als eine orthogonale Faktorenstruktur (vgl. Steel et al., 2008; siehe IV. 1 und V. 1). Die beiden latenten Variablen im Populationsmodell korrelierten in Größenordnungen, wie sie in der Psychologie vorkommen (z.B. Rost, 2009; Shrout & Yager, 1989), und determinierten gleichzeitig den Grad der Missspezifikation im Strukturmodell (siehe IV. 1 und V. 1). Um die praktische Relevanz der Arbeit zu erhöhen, wurden diagnostische Entscheidungen auf klinische Diagnosen übertragen, wobei verschiedene Basisraten in klinischen Größenordnungen (Wittchen & Jacobi, 2001; Wittchen et al., 2011) zur dichotomen Diagnostik verwendet wurden. Der Einbezug der Gesamtsummenwerte zur Diagnostik komplettiert das realitätsnahe Design im Sinne der gängigen Praxis der Testkonstruktion und Testanwendung (DiStefano et al., 2009; Estabrook & Neale, 2013; siehe III. 5.2).

Im Rahmen der ersten Simulationsstudie wurde eine Simulation mit anschließendem Resampling gewählt, da die Güte der Modellevaluation anhand der Fit-Indizes hinsichtlich der beschriebenen Missspezifikation ausgetestet werden sollte. Dies entspricht einer Robustheitsstudie hinsichtlich des Datenanalyseverfahrens an sich. Im Rahmen der zweiten Studie wurde die Praxis der klinischen Einzelfalldiagnostik simuliert. Das Ziel der Psychometrie ist es, die Ausprägungen der Individuen auf den latenten Variablen möglichst valide erfassen zu können. Daher wurde eine Simulationsmethode gewählt, die einerseits eine Populationssimulation darstellte, andererseits von den Populationsfaktorwerten ausging, also den wahren Ausprägungen der Individuen auf den latenten Variablen. Dies hatte den Vorteil, dass nicht nur die Güte der Diagnostik auf Basis der Faktorwerte eines missspezifizierten Modells berechnet werden konnte, sondern als Referenz zu dieser auch die Güte der Diagnostik, die die Faktorwerte eines korrekten Modells unter realistischen Bedingungen überhaupt leisten konnten.

Das genannte Forschungsdesign führt zur zweiten Stärke dieser Arbeit. Im Rahmen beider Simulationsstudien wurde das gleiche Design für die Populationsmodelle verwendet, sodass der Aufbau der Forschungsarbeit klar strukturiert und demnach die Aussagekraft hoch war: Zunächst wurden Fit-Indizes, welche in der angewandten psychologischen Forschung sehr oft für die Evaluation der Modellpassung herangezogen werden (Beauducel & Wittmann, 2005; Marsh et al., 2013; McDonald & Ho, 2002; Savalei, 2012; siehe II. 2 und IV. 1), auf ihre Sensitivität hinsichtlich einer realistischen Art und realistisch hohen Graden an Missspezifikation untersucht. In einem zweiten Schritt wurde dann am Beispiel von Diagnosen gezeigt, was diese Missspezifikation hinsichtlich der der substanziellen Aussagen, die aus den Faktorwerten dieser Modelle getroffen werden, für die Individuen überhaupt bedeuten würde.

Die dritte Stärke dieser Dissertation liegt darin, dass nicht nur aufgezeigt wurde, welche substanziellen negativen Konsequenzen für die Individuen im Rahmen der psychologischen Einzelfalldiagnostik entstehen, wenn bei der Testkonstruktion, wie auch bei der Auswertung von Daten, nicht nach dem Gold-Standard vorgegangen wird. Es wurde auch dargestellt, dass eine äußerst sorgfältige Test- oder Fragebogenkonstruktion nach den Gütekriterien zu sehr valider Diagnostik auf Basis der Bartlett-Faktorwerte führt. Basierend auf diesem Vergleich konnten einerseits Ratschläge für die Beurteilung der Modellpassung im Rahmen der Auswertung von Daten mittels Strukturgleichungsmodellen gegeben werden (siehe IV. 3.2), andererseits konkrete Handlungsempfehlungen für die Test- und Fragebogenkonstruktion

(siehe V. 3.3.2). Diese werden unter VI. 3.1 noch einmal aufeinander aufbauend zusammengefasst.

2.2 Grenzen der Arbeit

Trotz eines Forschungsdesigns, das sehr nahe an der Realität konzipiert wurde, konnte der diagnostische Entscheidungsprozess im Falle der klinischen Diagnosen dennoch nicht vollständig nachgestellt werden und nur beispielhaft vor dem Hintergrund einer untersuchten Art der Modellmissspezifikation skizziert werden.

An dieser Stelle sei neben den weiteren unter V. 3.3 genannten Limitationen nochmals die Tatsache erwähnt, dass es sich im Rahmen klinischer Diagnosen um formative Messmodelle handelt (Bühner, 2011; Curtis & Jackson, 1962; Reinecke, 2014; siehe V. 3.4), im Rahmen des Designs dieser Arbeit jedoch lediglich reflexive Messmodelle untersucht wurden.

Außerdem wurden die Diagnosen im Rahmen dieser Arbeit rein auf Basis einer konjunktiven Entscheidungsstrategie (Amelang & Schmidt-Atzert, 2006, S. 339) nach den höchsten Faktorwerten auf beiden Populationsfaktoren vergeben. Zumeist führen in der psychologischen Einzelfalldiagnostik aber auch hohe Ausprägungen auf den Indikatoren des einen Faktors kombiniert mit niedrigen Ausprägungen auf den Indikatoren eines zweiten Faktors insgesamt zu einer Diagnose, zumal in der gängigen diagnostischen Praxis meist Rohsummenwerte zu einer Diagnose führen (z.B. Beck et al., 2006; Estabrook & Neale, 2013). Letztere Entscheidungsstrategie wäre allerdings mit der unausgewogenen Indikatorenaufteilung hinsichtlich der Güte der Diagnostik basierend auf den Faktorwerten konfundiert gewesen.

Ferner waren im Rahmen des simulierten Fragebogen- bzw. Testdesigns das erforderliche Skalenniveau der Indikatoren, als auch die Voraussetzung der multivariaten Normalverteilung erfüllt, da der Fokus der Arbeit auf den Auswirkungen der unterschiedlichen Grade der Missspezifikation lag. Oft sind jedoch in der angewandten Forschung genau diese beiden genannten Annahmen verletzt (Kuzon, Urbanchek, & McCabe, 1996; Norman, 2010; Von Eye & Bogar, 2004), ohne dass dafür, z.B. anhand von robusten Maximum-Likelihood-Schätzmethoden bei Verletzung der multivariaten Normalverteilung (Finney & DiStefano, 2006; Satorra & Bentler, 1994; Yuan & Bentler, 1998, 2000) oder durch ein anderes

Schätzverfahren bei Vorliegen von ordinalskalierten Daten (Browne, 2011) korrigiert werden würde.

Zum Abschluss soll die Relevanz der Dissertation für Wissenschaft und Praxis herausgestellt werden und Implikationen aus den Befunden der Dissertation für die jeweiligen Bereiche abgeleitet werden.

3 Relevanz der Arbeit für Wissenschaft und Praxis

3.1 Wissenschaft

> „The difficulty faced by psychologists in measuring is not mathematical or empirical in nature, but is instead that the concepts they wish to have enter into their measurement operations are typically of the common-or-garden variety."
>
> (Maraun, 1998, S. 436)

Dieses Kommentar Marauns (1998) auf Wittgensteins eingangs erwähnte Worte betrifft eine der Hauptimplikationen bei der Testkonstruktion. Diese Implikation ist keineswegs neu, doch angesichts der aktuellen Replikationskrise der Psychologie von großer Bedeutung. Bei der Konstruktion von Tests, Fragebögen und Skalen bedarf es einer guten Theorie im Sinne einer präzisen und trennscharfen Konstruktdefinition. Dies stellt den ersten Schritt auf dem Weg zu erfolgreicher Psychometrie dar.

Die erste konkrete Empfehlung für die Konstruktion von Tests, Fragebögen oder Skalen, die sich aus der vorliegenden Dissertation ableiten lässt, betrifft den Prozess, den zu erfassenden Messgegenstand, der in den meisten Fällen psychologischer Forschung mehrdimensional sein wird (Little et al., 2002b), in ein empirisch überprüfbares Konstrukt zu übersetzen. Die Empfehlung lautet, das mehrdimensionale Konstrukt in homogene Teilbereiche aufzuteilen und diese möglichst reliabel zu erfassen und deren Faktorenstruktur zueinander möglichst valide zu spezifizieren[55]. Dabei ist eine äußerst sorgfältige Indikatoren-Konstruktion zur Erfassung dieses Konstrukts notwendig, insbesondere, wenn es sich um kleine Basisraten handelt. So wurde im Rahmen der Dissertation gezeigt, dass hohe Faktorladungen, also eine hohe Reliabilität des Tests bzw. Fragebogens oder der Skala, sowohl die Sensitivität des χ^2-Tests und der Fit-Indizes für Missspezifikationen erhöhte, als auch die Güte diagnostischer Entscheidungen auf Basis der Faktorwerte eines Modells verbesserte. Eine hohe Reliabilität zeigte sich insbesondere vor dem Hintergrund kleiner Basisraten, wie sie in der klinischen

[55]Dies ist beispielsweise bei der Konstruktion des „I-S-T 2000 R" (Liepmann et al., 2007) sehr gelungen.

Psychologie vorkommen, von großer Bedeutung für die diagnostische Präzision. Das Gleiche wie für die Reliabilität gilt für die Korrektheit der faktoriellen Struktur, also für die Konstruktvalidität. Es wurde gezeigt, dass eine hohe Konstruktvalidität im Sinne der Absenz einer Missspezifikation im Strukturmodell in Kombination mit einer sehr hohen Reliabilität zu valider Diagnostik auf Basis der Faktorwerte führte. Desto höher der Grad der Missspezifikation (desto niedriger die Konstruktvalidität), desto mehr litt die Güte der diagnostischen Entscheidungen beim Vorliegen kleiner Basisraten, sofern das Ziel der Diagnostik das Erkennen eines Krankheitszustands war. Umgekehrt beeinträchtigte eine niedrige Konstruktvalidität bei großen Basisraten die Wahrscheinlichkeit, Gesunde als gesund zu erkennen. Ferner wurde im Rahmen des Exkurses gezeigt, dass die Diagnostik auf Basis der Faktorwerte eines sorgfältig konstruierten (hoch reliablen und konstruktvaliden) Testinstruments der bloßen Aufsummierung der Antworten auf den Indikatoren vorzuziehen ist.

Sofern die Voraussetzungen einer guten Theorie sowie deren reliable und valide Erfassung gesichert sind und der nächste Schritt die Auswertung der (Norm-)Daten ist, ist in jedem Fall davon abzuraten, sich bei der Modellevaluation im Rahmen der Strukturgleichungsmodellierung nur auf die Fit-Indizes zu verlassen. Wie in der vorliegenden Dissertation gezeigt wurde, konnten die gängigsten Fit-Indizes die Dimensionalitätsverletzung größtenteils anhand der Cut-Off-Kriterien nach Hu und Bentler (1998, 1999) nicht als modellabweichend identifizieren, der χ^2-Test allerdings schon. Sofern die Fit-Indizes Modellpassung indizieren, der χ^2-Test aber nicht, ist insofern Vorsicht geboten. Da aus der Signifikanz des χ^2-Tests insbesondere bei großen Stichproben jedoch nicht geschlussfolgert werden kann, wie groß die Modellverletzung ist (Saris et al., 1987), ist die gleichzeitige Betrachtung lokaler Maße der Modellpassung unbedingt erforderlich.

Der letzte Schritt der Testkonstruktion betrifft die Wahl eines geeigneten Cut-Off-Werts, der – im Rahmen eines dichotomen Beispiels, wie es auch in der vorliegenden Dissertation verwendet wurde – die Individuen mit der Diagnose „krank“ möglichst trennscharf von den Individuen mit der Diagnose „gesund“ differenziert. Hinsichtlich dieses Cut-Offs ist eine äußerst kritische Abwägung hinsichtlich Sensitivität und Spezifität anhand der ROC-Kurve unter Beachtung des Testzwecks (Screening, Beurteilung des Schweregrades einer Erkrankung, etc.) zu treffen (Amelang & Schmidt-Atzert, 2006). Wird der Testtrennwert erhöht, sodass die Sensitivität steigt und die Rate an falsch negativen Diagnosen geringer wird,

sinkt jedoch die Spezifität und die Rate an falsch positiven Diagnosen wird höher (Amelang & Schmidt-Atzert, 2006; Ziegler & Bühner, 2012). Wie die vorliegende Dissertation zeigte, spielt die Reliabilität (Faktorladungen) auch hinsichtlich des Cut-Offs bei mehrdimensionalen Testverfahren mit unterschiedlich hohen Korrelationen zwischen den Faktoren eine Rolle. Die realistisch niedrige Reliabilität im Rahmen des vorliegenden Studiendesigns führte dazu, dass der gewählte univariate Cut-Off für die Diagnosen in einer Verzerrung der Basisraten bei Anwendung des korrekten Modells resultierte.

Schwieriger zu erreichen als eine reliable und valide Testkonstruktion ist ein geeigneter Umgang mit kleinen Basisraten, da diese nicht veränderbar sind. Basisraten können insbesondere im klinischen Bereich oft sehr klein ausfallen, sogar deutlich kleiner als im Rahmen des untersuchten Designs. So liegen die Punktprävalenzen für psychische Störungen wie Anorexie, Schizophrenie oder die Borderline-Störung beispielsweise unter 1% (Wittchen et al., 2011). Das aus den kleinen Basisraten resultierende Dilemma hinsichtlich der Sensitivität und des Prädiktionswerts eines diagnostischen Instruments wurde sowohl in der vorliegenden Arbeit als auch anhand der unter III. 4 beschriebenen Studien deutlich. Dieser Befund impliziert besondere Aufmerksamkeit vonseiten der Diagnostikerin/des Diagnostikers bei der Verwendung von Screening-Verfahren.

Ein Ansatz, der die Problematik, die aus den kleinen Basisraten resultiert, möglicherweise entschärfen könnte, stellt dar, die Diagnose durch den Einsatz verschiedener Testverfahren mehrfach abzusichern. Dies macht allerdings bei Testpersonen, die durch ihre Störung stark kognitiv beeinträchtigt sind, wenig Sinn. Ein anderer Ansatz, der ökonomischer und zumutbarer für die getesteten Individuen wäre, könnte darstellen, die Diagnostik in Teile aufzuteilen (vgl. Emons et al., 2007), die zusammen die End-Diagnose ergeben sollen. Dies könnte, wie bereits unter V. 3.3.2 beschrieben, durch die Beurteilung der Präsenz oder Absenz eines Symptoms erreicht werden. Dieser Vorschlag würde sich zusammen mit dem bekannten Vorschlag, das zu erfassende Konstrukt in möglichst homogene Teilkonstrukte aufzuteilen, um es anschließend hinsichtlich der Faktorenstruktur möglichst valide zu spezifizieren, sehr gut in Einklang bringen und umsetzen lassen.

3.2 Praxis

Neben den genannten Implikationen für die Testkonstruktion macht diese Dissertation auch deutlich, wie wichtig nicht nur eine gute wissenschaftliche Ausbildung für Diagnostikerinnen und Diagnostiker ist, sondern auch, wie wichtig diagnostische Erfahrung in den jeweiligen Bereichen und gewissenhaftes Vorgehen bei der Diagnostik ist, zumal die diagnostische Erfahrung höchstwahrscheinlich zusätzlich von der Höhe der Basisrate abhängt.

Im Rahmen dieser Dissertation und einem nachsimulierten klinischen Design wurden nur unterschiedliche Basisraten untersucht, nicht auch unterschiedliche Selektionsraten. In der Eignungsdiagnostik kommt jedoch in den meisten Fällen eine von den Basisraten verschiedene Selektionsrate hinzu. Sofern sich Basis- und Selektionsrate nicht gleichen, kann eine im Vergleich zur Basisrate kleinere Selektionsrate die Trefferquote erhöhen (Schönemann & Thompson, 1996; Taylor & Russell, 1939). Das heißt, mehr Bewerberinnen und Bewerber erhöhen die Chance, eine geeignete Kandidatin/einen geeigneten Kandidaten für eine vakante Position auszuwählen.

Für die Festlegung von geeigneten Testtrennwerten gibt es keine eindeutige Lösung (Amelang & Schmidt-Atzert, 2006, S. 432). Gleichzeitig geht eine diagnostische Entscheidung aber mit einer qualitativen Wertung einher, die hinsichtlich individueller und gesellschaftlicher Konsequenzen äußerst sorgfältig erwägt werden sollte (Amelang & Schmidt-Atzert; Wieczerkowski & Oeveste, 1982). Die praktische Bedeutsamkeit mangelhafter diagnostischer Instrumente für die Einzelfalldiagnostik wird im Folgenden ausgeführt.

Falsche Diagnosen können weitreichende Konsequenzen nach sich ziehen. Falsch positive Diagnosen verursachen nicht nur unnötige Behandlungskosten für Individuum und Gesundheitswesen und damit wiederum für die Gesellschaft, sondern können für die Betroffenen auch zu negativen psychischen und physischen Konsequenzen führen. Derartige negative Folgen führen von der Stigmatisierung, welche wiederum erst psychische Beeinträchtigung verursachen kann, bis hin zu unerwünschten Nebenwirkungen durch psychologische Psychotherapie und/oder Pharmakotherapie (Angermeyer, 2003; Berk & Parker, 2009; Brüggemann, 2007; Holzinger, Beck, Munk, Weithaas, & Angermeyer, 2003; Linden, 2013; Rüsch, 2010). Falsch negative Diagnosen können die Unterlassung notwendiger Interventionsmethoden nach sich ziehen. Eine hohe Rate an falsch negativen Diagnosen ist insbesondere im Rahmen von Screenings problematisch, da Patienten mit einer negativen

Diagnose nicht weiter untersucht werden. Dieser Umstand kann sich, verglichen mit einer frühzeitigen korrekten Diagnosestellung, in einer verringerten Chance auf Heilungserfolg äußern (Milos, Spindler, Schnyder, & Fairburn, 2005; Von Holle et al., 2008).

Vor dem Hintergrund der Relevanz dieser Dissertation für Wissenschaft und Praxis erscheinen Wittgensteins eingangs erwähnte Worte zur Bedeutung der Wissenschaftlichkeit ebenso wesentlich wie 1953 (Maraun, 1998).

Literaturverzeichnis

Amelang, M., & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4th ed.). Heidelberg: Springer.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140.

Anderson, R. D., & Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability* (Vol. 5, pp. 111–150).

Angermeyer, M. C. (2003). Das Stigma psychisch Kranker aus der Sicht der Patienten - Ein Überblick. *Psychiatrische Praxis*, *30*(7), 358–366.

Arminger, G. (1979). Faktorenanalyse. In E. K. Scheuch & H. Sahner (Eds.), *Studienskripten zur Soziologie - Statistik für Soziologen, Bd. 3*. Stuttgart: Teubner Studienskripten.

Bandalos, D. L. (2002). The Effects of Item Parceling on Goodness-of-Fit and Parameter Estimate Bias in Structural Equation Modeling. *Structural Equation Modeling*, *9*(1), 37–41. http://doi.org/10.1207/S15328007SEM0901

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology*, *28*, 97–104.

Beauducel, A. (2005). How to describe the difference between factors and corresponding factor-score estimates. *Methodology*, *1*(4), 143–158. http://doi.org/10.1027/1614-2241.1.4.143

Beauducel, A. (2007). In spite of Indeterminacy many common factor score estimates yield an

identical reproduced covariance matrix. *Psychometrika*, *72*(3), 437–441.

Beauducel, A., & Wittmann, W. W. (2005). Simulation Study on Fit Indexes in CFA Based on Data With Slightly Distorted Simple Structure. *Structural Equation Modeling*, *12*(1), 41–75.

Beck, A. T., Steer, R. A., & Brown, G. K. (2013). *Beck Depressions-Inventar - Fast Screen [Beck Depression Inventory - Fast Screen]*. Frankfurt am Main: Pearson PsychCorp.

Beck, A. T., Steer, R. A., Brown, G. K., Hautzinger, M., Keller, F., & Kühner, M. (2006). *Beck-Depressionsinventar (BDI-II)*. Göttingen: Hogrefe.

Bem, D. J. (2011). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology*, Advance Online Publication. http://doi.org/10.1037/a0021524

Bentler, P. M. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin*, *107*(2), 238–246.

Bentler, P. M. (1992). *EQS: Structural Equations Programm Manual*. Los Angeles, CA: BMDP Statistical Software.

Bentler, P. M. (1995). EQS structural equations program manual. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606.

Bentler, P. M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: test statistics. *Multivariate Behavioral Research*, *34*(2), 181–197.

Berk, M., & Parker, G. (2009). The elephant on the couch: side-effects of psychotherapy.

Australian and New Zealand Journal of Psychiatry, *43*(9), 787–794. http://doi.org/10.1080/00048670903107559

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality*. Unpublished PhD thesis: University of Groningen.

Boomsma, A., & Hoogland, J. J. (2001). The robustness of maximum likelihood estimations in structural equation models. In R. Cudeck, S.-H. C. du Toit, & D. Sörbom (Eds.), *Structural modeling by example: Present and future. A festschrift in honor of Carl Jöreskog* (pp. 139–168). Lincolnwood: Scientific Software International.

Bortz, J., & Döring, N. (n.d.). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (3. Auflage). Berlin, Heidelberg: Springer.

Breivik, E., & Olsson, U. H. (2001). Adding variables to improve model fit: The effect of model size on fit assessment in LISREL. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 169–194). Lincolnwood, IL: Scientific Software International.

Brown, T. A., & Moore, M. T. (2012). Confirmatory Factor Analysis. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (1st ed., pp. 361–379). New York: The Guilford Press.

Browne, M. W. (2011). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. L. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury

Park, CA: Sage.

Browne, M. W., MacCallum, R. C., Kim, C.-T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, *7*(4), 403–21. http://doi.org/10.1037/1082-989X.7.4.403

Brüggemann, B. R. (2007). Ethische Aspekte in der Frühintervention und Akutbehandlung. *Ethik in Der Medizin*, *19*, 91–102.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3rd ed.). München: Pearson.

Buzick, H. M. (2010). Testing for heterogeneous factor loadings using mixtures of confirmatory factor analysis models. *Frontiers in Psychology*, *1*, 1–9. http://doi.org/10.3389/fpsyg.2010.00165

Byrne, B. M. (1998). *Structural Equation Modeling with LISREL, PRELIS and SIMPLIS: Basic Concepts, Applications and Programming*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Byrne, B. M. (2009). *Structural Equation Modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York, NY: Routledge, Taylor & Francis Group.

Carsey, T. M., & Harden, J. J. (2014). *Monte Carlo Simulation and Resampling Methods for Social Science*. Los Angeles: Sage Publications.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504.

Cliff, N., & Hamburger, C. D. (1967). The Study of Sampling Errors in Factor Analysis by

Means of Artificial Experiments. *Psychological Bulletin*, *68*(6), 430–445. http://doi.org/10.1037/h0025178

Cliff, N., & Pennell, R. (1967). The Influence of Communality, Factor Strength, and Loading Size on the Sampling Characteristics of Factor Loadings. *Psychological Science*, *32*(3), 209–326.

Cohen, L., & Manion, L. (1980). *Research Methods in Education*. London: Croom Helm.

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The Noncentral Chi-Square Distribution ion Missspecified Structural Equation Models: Finite Sample Results from a Monte Carlo Simulation. *Structural Equation Modeling*, *37*(1), 1–36.

Curtis, R. F., & Jackson, E. F. (1962). Multiple Indicators in Survey Research. *American Journal of Sociology*, *68*(195–204).

Datenbanksegment PSYNDEX Tests. (2013). Retrieved March 30, 2016, from http://www.zpid.de/index.php?wahl=products&uwahl=fee&uuwahl=ptinfo#klassifikation

Dillon, W. R., Kumar, A., & Mulani, N. (1987). Offending Estimates in Covariance Structure Analysis: Comments on the Causes of and Solutions to Heywood Cases. *Psychological Bulletin*, *101*(1), 126–135.

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling*, *2*(2), 119–143. http://doi.org/10.1080/10705519509540000

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Assessment, Research &*

Evaluation, *14*(20), 1–11.

Dobie, T., McFarland, K., & Long, N. (1986). Raw Score and Factor Score Multiple Regression: An Evaluative Comparison. *Educational and Psychological Measurement*, *46*, 337–347.

Dozois, D. J. A., Dobson, K. S., & Ahnberg, J. L. (1998). A Psychometric Evaluation of the Beck Depression Inventory-II, *10*(2), 83–89.

Eid, M., Gollwitzer, M., & Schmitt, M. (2013). *Statistik und Forschungsmethoden* (3rd ed.). Weihnheim, Basel: Beltz.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*(1), 105–20. http://doi.org/10.1037/1082-989X.12.1.105

Erdfelder, E., Faul, F., Buchner, A., & Cüpper, L. (2010). Effektgröße und Teststärke. In H. Holling & B. Schmitz (Eds.), *Handbuch Statistik, Methoden und Evaluation*. Göttingen: Hogrefe.

Estabrook, R., & Neale, M. (2013). A Comparison of Factor Score Estimation Methods in the Presence of Missing Data: Reliability and an Application to Nicotine Dependence. *Multivariate Behavioral Research*, *48*(1), 1–27. http://doi.org/10.1080/00273171.2012.730072

Fahrmeier, L., Hamerle, A., & Tutz, G. (1996). *Appendix A in Multivariate statistische Verfahren* (2nd ed.). Berlin: de Gryter.

Fan, X., & Sivo, S. A. (2005). Sensitivity of Fit Indexes to Misspecified Structural or Measurement Model Components: Rationale of Two-Index Strategy Revisited. *Structural Equation Modeling*, *12*(3), 343–367.

http://doi.org/10.1207/s15328007sem1203_1

Fan, X., Thompson, B., & Wang, L. (2009). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, *6*(1), 56–83.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwich: Information Age Publishing.

Formann, A. (1981). Über die Verwendung von Items als Teilungskriterium für Modellkontrollen im Modell von Rasch. *Zeitschrift Für Experimentelle Und Angewandte Psychologie*, *28*(4), 541–560.

Gardner, R. C., & Neufeld, R. W. J. (2013). What the correlation coefficient really tells us about the individual. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, *45*(4), 313–319. http://doi.org/10.1037/a0033342

Gatewood, R. G., Feild, H. S., & Barrick, M. R. (2016). *human resource selection* (8th ed.). Boston: Cengage Learning.

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In Bollen Scott J. & K. A. . Long (Eds.), *Testing structural equation models* (pp. 40–65). Sage.

Green, D. M., & Swets, J. A. (1966). *Signal Detection and Psychophysics*. New York: Wiley.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling*, *7*(2), 251–270.

Grice, J. W. (2001a). A Comparison of Factor Scores Under Conditions of Factor Obliquity.

Psychological Methods, *6*(1), 67–83.

Grice, J. W. (2001b). Computing and Evaluating Factor Scores. *Psychological Methods*, *6*(4), 430–450.

Grice, J. W., & Harris, R. J. (1998). A Comparison of Regression and Loading Weights for the Computation of Factor Scores. *Multivariate Behavioral Research*, *33*(2), 221–247. http://doi.org/10.1207/s15327906mbr3302

Gu, F., Little, T. D., & Kingston, N. M. (2013). Misestimation of Reliability Using Coefficient Alpha and Structural Equation Modeling When Assumptions of Tau-Equivalence and Uncorrelated Errors Are Violated. *Methodology*, *9*(1), 30–40.

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical Psychology*, *8*(2), 65–81.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking Misfit in Confirmatory Factor Analysis by Increasing Unique Variances: A Cautionary Note on the Usefulness of Cutoff Values of Fit Indices. *Psychological Methods*, *16*(3), 319–336.

Heene, M., Hilbert, S., Freudenthaler, H., & Bühner, M. (2012). Sensitivity of SEM Fit Indexes with Respect to Violations of uncorrelated Errors. *Structural Equation Modeling*, *6*(1), 56–83.

Heene, M., Kyngdon, A., & Sckopke, P. (2016). Detecting Violations of Unidimensionality by Order-Restricted Inference Methods. *Frontiers in Applied Mathematics and Statistics*, *2*(3). http://doi.org/10.3389/fams.2016.00003

Hershberger, S. L. (2003). The Growth of Structural Equation Modeling: 1994-2001. *Structural Equation Modeling*, *10*(1), 35–46.

http://doi.org/10.1207/S15328007SEM1001_2

Holzinger, A., Beck, M., Munk, I., Weithaas, S., & Angermeyer, M. C. (2003). Das Stigma psychischer Krankheit aus der Sicht schizophren und depressiv Erkrankter. *Psychiatrische Praxis*, *30*(7), 395–401.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural Equation Modelling : Guidelines for Determining Model Fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. http://doi.org/10.1037//1082-989X.3.4.424

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. http://doi.org/10.1080/10705519909540118

Janssen, J., & Laatz, W. (2013). *Statistische Datenanalyse mit SPSS - Eine anwendungsorientierte Einführung in das Basissystem und das Modul Exakte Tests* (8th ed.). Berlin, Heidelberg: Springer.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 83–112). New York: Seminar Press.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods*. Uppsala, Sweden.

Kaplan, D. (1988). The Impact of Specification Error on the Estimation, Testing, and Improvement of Structural Equation Models. *Multivariate Behavioral Research*, *23*, 37–41.

Kauermann, G., & Küchenhoff, H. (2011). *Stichproben - Methoden und praktische Umsetzung in R*. Berlin, Heidelberg: Springer.

Kelderman, H., & Molenaar, P. C. M. (2007). The effect of individual difference in factor loadings on the standard factor model. *Multivariate Behavioral Research*, *42*(3), 435–456. http://doi.org/10.1080/00273170701382997

Kenny, D. a, & McCoach, D. B. (2009). Effect of the Number of Variables on Measures of Fit in Structural Equation Modeling. *Structural Equation Modeling*, *10*(3), 333–351. http://doi.org/10.1207/S15328007SEM1003

Kline, R. B. (2001). *Principles and Practice of Structural Equation Modeling* (3rd ed.). New York: NY: Guilford.

Kolenikov, S., & Bollen, K. A. (2012). Testing Negative Error Variances: Is a Heywood Case a Symptom of Misspecification? *Sociological Methods & Research*, *41*(1), 124–167. http://doi.org/10.1177/0049124112442138

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test Length and Decision Quality in Personnel Selection : When Is Short Too Short ? *International Journal of Testing*, *12*, 321–344. http://doi.org/10.1080/15305058.2011.643517

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2014). Assessing Individual Change Using Short Tests and Questionnaires. *Applied Psychological Measurement*, *38*(3), 201–216. http://doi.org/10.1177/0146621613510061

Kukuk, C. R., & Baty, C. F. (1979). The Misuse of Multiple Regression with Compositve

Scales obtained from Factor Scores. *Educational and Psychological Measurement*, *39*, 277–290.

Kuzon, W. M., Urbanchek, M. G., & McCabe, S. (1996). The seven deadly sins of statistical analyses. *Annals of Plastic Surgery*, *37*, 265–272.

lavaan Google Groups. (2015). Retrieved July 27, 2015, from https://groups.google.com/forum/#!forum/lavaan

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). London: Butterworth & Co (Publishers) Ltd.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)* (2nd ed.). Göttingen: Hogrefe.

Linden, M. (2013). How to Define, Find and Classify Side Effects in Psychotherapy: From Unwanted Events to Adverse Treatment Reactions. *Clinical Psychology & Psychotherapy*, *20*(4), 286–296. http://doi.org/10.1002/cpp.1765

Little, T. D., Cunningham, W. a., Shahar, G., & Widaman, K. F. (2002a). To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits. *Structural Equation Modeling*, *9*(2), 151–173. http://doi.org/10.1207/S15328007SEM0902_1

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002b). Evaluating Goodness-of- Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. http://doi.org/10.1207/S15328007SEM0902

Maathuis, M. (2008). Factor analysis. Retrieved April 10, 2016, from https://stat.ethz.ch/~maathuis/teaching/fall08/Notes5.pdf

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201–226.

MacCallum, R. C., & Austin, J. T. (2000). Applications Of Structural Equation Modeling in Psychological Research. *Annual Review of Psychology*, 201–226. http://doi.org/10.1146/annurev.psych.51.1.201

MacKenzie, S. B., Podsakoff, N. P., & Jarvis, C. (2005). The Problem of Measurement Model Missspecification in Behavioral and Organizational Research and some Recommended Solutions. *Journal of Applied Psychology*, *90*(4), 710–730.

Mahler, C. (2011). *The effects of misspecification type and nuisance variables on the behaviors of population fit indices used in structural equation modeling*. Unpublished master's thesis: University of British Columbia.

Maraun, M. D. (1996a). Metaphor Taken as Math : Indeterminancy in the Factor Analysis Model. *Multivariate Behavioral Research*, *31*(4), 517–538. http://doi.org/10.1207/s15327906mbr3104

Maraun, M. D. (1996b). The Claims of Factor Analysis. *Multivariate Behavioral Research*, *31*(4), 673–689. http://doi.org/10.1207/s15327906mbr3104

Maraun, M. D. (1998). Measurement as a Normative Practice. *Theory & Psychology*, *8*(4), 435–461. http://doi.org/0803973233

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis. *Psychological Bulletin*, *103*, 391–410.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2013). Goodness of Fit in Structural Equation Models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary Psychometrics* (3rd ed., pp. 275–339). New York and Hove: Psychology Press.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cut-off values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341. http://doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., & Hocevar, D. (1988). A New, More Powerful Approach to Multitrait Multimethod Analyses - Application of a 2nd-Order Confirmatory Factor Analysis. *Journal of Applied Psychology*, *73*, 107–117.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*(3), 471–491. http://doi.org/10.1037/a0019227

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: two wrongs do not make a right--camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, *18*(3), 257–84. http://doi.org/10.1037/a0032773

Maxwell, A. E. (1968). The Effect Of Correlated Errors On Estimates Of Reliability Coefficients. *Educational and Psychological Measurement*, *28*, 803–811.

McDonald, R. P., & Burr, E. J. (1967). A comparison of four methods of constructing factor scores. *Psychometrika*, *32*(4), 381–401.

McDonald, R. P., & Ho, M.-H. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns or cutting scores. *Psychological Bulletin*, *52*(3), 194–216.

Meyer, C., Rumpf, H.-J., Hapke, U., Dilling, H., & John, U. (2000). Lebenszeitprävalenz psychischer Störungen in der erwachsenen Allgemeinbevölkerung. *Nervenarzt*, *71*, 535–542.

Milos, G., Spindler, A., Schnyder, U., & Fairburn, C. G. (2005). Instability of eating disorder diagnoses: prospective study. *The British Journal of Psychiatry*, *187*, 573–578.

Mulaik, S. A. (2009). *Foundations of Factor Analysis* (2nd ed.). Boca Raton, FL: CRC Press.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. http://doi.org/10.1007/s10459-010-9222-y

Oberski, D., & Satorra, A. (2013). Measurement error models with uncertainty about the error variance. *Structural Equation Modeling*, *20*(3), 409–428.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The Performance of ML, GLS, and WLS Estimation in Structural Equation Modeling Under Conditions of Misspecification and Nonnormality. *Structural Equation Modeling*, *7*(4), 557–595. http://doi.org/10.1207/S15328007SEM0704_3

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. http://doi.org/10.1126/science.aac4716

Osburn, H. (2000). Coefficient Alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*, 343–355.

Paxton, P., Curran, P. J., Bollen, K. a., Kirby, J., & Chen, F. (2001). Monte Carlo Experiments: Design and Implementation. *Structural Equation Modeling*, *8*(2), 287–312. http://doi.org/10.1207/S15328007SEM0802_7

Pennell, R. (1968). The Influence of Communality and N on the Sampling Distributions of Factor Loadings. *Psychometrika*, *33*(4), 423–439.

Peterson, R. A. (2000). A meta analysis of variance accounted for and factor loadings in exploratory factor analysis. *Marketing Letters*, *11*, 261–275. http://doi.org/10.1023/A:1008191211004

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *The Journal of Applied Psychology*, *88*(5), 879–903. http://doi.org/10.1037/0021-9010.88.5.879

Pornprasertmanit, S., Miller, P., & Schoemann, A. (2015). simsem: SIMulated Structural Equation Modeling. Retrieved from http://cran.r-projetct.org/package=simsem

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.r-project.org/

Raykov, T. (2001). Bias of Coefficient Alpha for Fixed Congeneric Measures with Correlated Errors. *Applied Psychological Measurement*, *25*(1), 69–76. http://doi.org/10.1177/01466216010251005

Reinecke, J. (2014). *Strukturgleichungsmodelle in den Sozialwissenschaften*. München: De Gruyter Oldenbourg Wissenschaftsverlag GmbH.

Revelle, W. (2015). psych: Procedures for Personality and Psychological Research. Evanston, Illinois, USA: Northwestern University. Retrieved from http://cran.r-project.org/package=psych

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling*, *3*(4), 369–379.

http://doi.org/10.1080/10705519609540052

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Rost, J. (2009). *Intelligenz – Fakten und Mythen*. Weinheim, Basel: Beltz.

Rüsch, N. (2010). Reaktionen auf das Stigma psychischer Erkrankungen - Sozialpsychologische Modelle und empirische Befunde. *Zeitschrift Für Psychiatrie, Psychologie Und Psychotherapie*, *58*(4), 287–297.

Saris, W. E., Satorra, A., & Sorbom, D. (1987). The Detection and Correction of Specification Errors in Structural Equation Models. *Sociological Methodology*, *17*, 105–129.

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling*, *16*(4), 561–582. http://doi.org/10.1080/10705510903203433

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. Von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Savalei, V. (2012). The Relationship Between Root Mean Square Error of Approximation and Model Misspecification in Confirmatory Factor Analysis Models. *Educational and Psychological Measurement*, *72*(6), 910–932. http://doi.org/10.1177/0013164412452564

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models : Tests of Significance and Descriptive Goodness-of-Fit Measures, *8*(2), 23–74.

Schimmack, U. (2012). The Ironic Effect of Significant Results in the Credibility of Multiple-Study Articles. *Psychological Methods*, *17*(4), 551–566.

Schimmack, U. (2016). The Replicability-Index: Quantifying Statistical Research Integrity. Retrieved January 1, 2016, from https://wordpress.com/post/replication-index.wordpress.com/920

Schönemann, P. H. (1981). Power as a function of communality in factor analysis. *Bulletin of the Psychonomic Society*, *17*, 57–60.

Schönemann, P. H. (1996). The Psychopathology of Factor Indeterminacy. *Multivariate Behavioral Research*, *31*(4), 571–577.

Schönemann, P. H. (1997). Some new results on hit rates and base rates in mental testing. *Chinese Journal of Psychology*, *39*(2), 173–192.

Schönemann, P. H., & Steiger, J. H. (1978). On the validity of indeterminate factor scores. *Bulletin of the Psychonomic Society*, *12*(4), 287–290.

Schönemann, P. H., & Thompson, W. W. (1996). Hit rate bias in mental testing. *Cahiers de Psychologie/Current Psychology of Cognition*, *15*, 3–28.

Schuler, H. (2014). *Psychologische Personalauswahl- Eignungsdiagnostik für Personalentscheidungen und Berufsberatung* (4th ed.). Göttingen: Hogrefe.

SCImago. (2007). SJR - SCImago Journal & Country Rank. Retrieved March 30, 2016, from http://www.scimagojr.com

Shrout, P. E., & Yager, T. J. (1989). Reliability and validity of screening scales: effect of reducing scale length. *Journal of Clinical Epidemiology*, *42*(1), 69–78. http://doi.org/10.1016/0895-4356(89)90027-9

Skrondal, A., & Rabe-Hesketh, S. (2014). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall.

Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *American Journal of Psychology*, *15*, 201–293.

Spearman, C. (1927). *The abilities of man*. New York: Macmillan.

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin*, *134*(1), 138–161.

Steiger, J. H. (1979). The relationship between external variables and common factors. *Psychometrika*, *44*(1), 93–97.

Steiger, J. H. (1990). Structural model evaluation and modification. *Multivariate Behavioral Research*, *58*(1), 935–943.

Steiger, J. H. (1996). Coming Full Circle in the History of Factor Indeterminancy. *Multivariate Behavioral Research*, *31*(4), 617–630. http://doi.org/10.1207/s15327906mbr3104

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*(2), 893–898.

Steiger, J. H., & Lind, J. C. (1980). Statistically-based tests for the number of common factors. In *Paper presented at the annual Spring Meeting of the Psychometric Society*. Iowa City.

Steiger, J. H., & Schönemann, P. H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis* (pp. 136–178). Chicago: University of Chicago Press.

Stelzl, I. (1979). Ist der Modelltest des Rasch-Modells geeignet, Homogenitätshypothesen zu prüfen? Ein Bericht über Simulationsstudien mit inhomogenen Daten. *Zeitschrift Für Experimentelle Und Angewandte Psychologie*, *26*, 652–672.

Stout, W. F., Habing, B., Douglas, J., Kim, H. R., Roussos, L. A., & Zhang, I. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*, 331–354.

Structural Equation Modeling: A Multidisciplinary Journal. (2015). Retrieved January 1, 2015, from http://www.tandfonline.com/loi/hsem20#.VYfjSWCInJF

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology*, *23*(5), 565–578. http://doi.org/10.1037/h0057079

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: a historical overview and some guidelines. *Educational and Psychological Measurement*, *56*(2), 928197–208.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, *1*, 31–65. http://doi.org/10.1146/annurev.clinpsy.1.102803.144239

Tremblay, P. F., & Gardner, R. C. (1996). On the growth of structural equation modeling in psychological journals. *Structural Equation Modeling*, *3*(2), 93–104. http://doi.org/10.1080/10705519609540035

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. http://doi.org/10.1007/BF02291170

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New

York: Springer.

Von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution Rasch models: Externsions and applications*. New York: Springer.

Von Eye, A., & Bogar, A. G. (2004). Testing the assumption of multivariate normality. *Psychology Science*, *2*, 243–258.

Von Holle, A., Poyastro Pinheiro, A., Thornton, L. M., Klump, K. L., Berrettini, W. H., Brandt, H., & et al. (2008). Temporal patterns of recovery across eating disorder subtypes. *Australian and New Zealand Journal of Psychiatry*, *42*(2), 108–117. http://doi.org/10.1080/00048670701787610

Wagenmakers, E.-J., Wetzels, R., & Borsboom, D.van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. http://doi.org/http://dx.doi.org/10.1037/a0022790

Wei, H. (2008). *Multidimensionality in the NAEP Science Assessment: Substantive Perspectives, Psychometric Models, and Task Design*. Unpublished Doctoral Dissertation: University of Maryland.

Weiber, R., & Mühlhaus, D. (2010). *Strukturgleichungsmodellierung. Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS*. Berlin, Heidelberg: Springer.

Wieczerkowski, W., & Oeveste, H. Z. (1982). Zuordnungs- und Entscheidungsstrategien. In K. J. Klauer (Ed.), *Handbuch der Pädagogischen Diagnostik (Bd. 2, Studienausgabe)* (pp. 919–951). Düsseldorf: Schwann.

Wilson, E. B. (1928). On Hierarchical Correlation Systems. *Proceedings of the National*

Academy of Sciences, *14*, 283–291.

Wittchen, H., & Jacobi, F. (2001). Die Versorgungssituation psychischer Störungen in Deutschland, *44*, 993–1000. http://doi.org/10.1007/s001030100269

Wittchen, H. U., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jönsson, B., & et al. (2011). The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology*, *21*(9), 655–679. http://doi.org/10.1016/j.euroneuro.2011.07.018

Wittgenstein, L. J. J. (1953). *Philosophische Untersuchungen [Philosophical Investigations]*. Oxford: Blackwell.

World Health Organization. (1993). Chapter V (F): mental and behavioral disorders. Clinical descriptions and Diagnostic Guidelines. In *Tenth revision of the international classification of diseases*. Geneva: World Health Organization.

Yang-Wallentin, F., & Jöreskog, K. G. (2001). Robust standard errors and chi squares for interaction models. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 159–171). Mahwah, NJ: Erlbaum.

Yang, Y., & Green, S. B. (2010). A Note on Structural Equation Modeling Estimates of Reliability. *Structural Equation Modeling*, *17*(1), 66–81. http://doi.org/10.1080/10705510903438963

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, *51*(2), 289–309.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis. *Sociological Methodology*, *30*(1), 165–200.

Ziegler, M., & Bühner, M. (2012). *Grundlagen der Psychologischen Diagnostik*. Wiesbaden: VS Verlag fuer Sozialwissenschaften.

Zimmermann, D. W., & Williams, R. H. (1977a). The Theory of Test Validity and Correlated Errors of Measurement. *Journal of Mathematical Psychology*, *16*, 135–152.

Zimmermann, D. W., & Williams, R. H. (1977b). Validity Coefficients And Correlated Errors In Test Theory. *Journal of Experimental Education*, *45*(3), 4–9.

Anhang

1 Diagnostische Konsistenzen korrekte Modelle

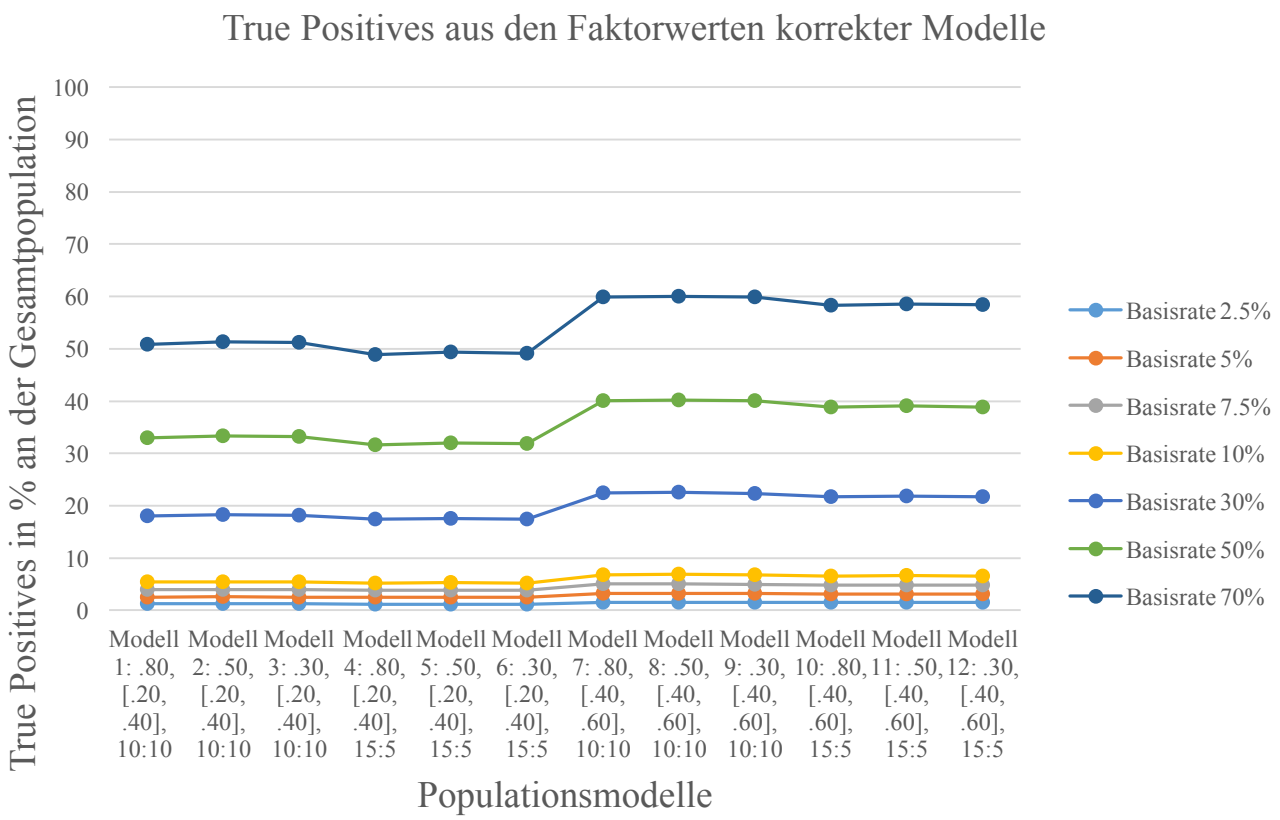

Abbildung 14. Durch die Bartlett-Faktorwerte korrekter Modelle korrekt erkannte Positive

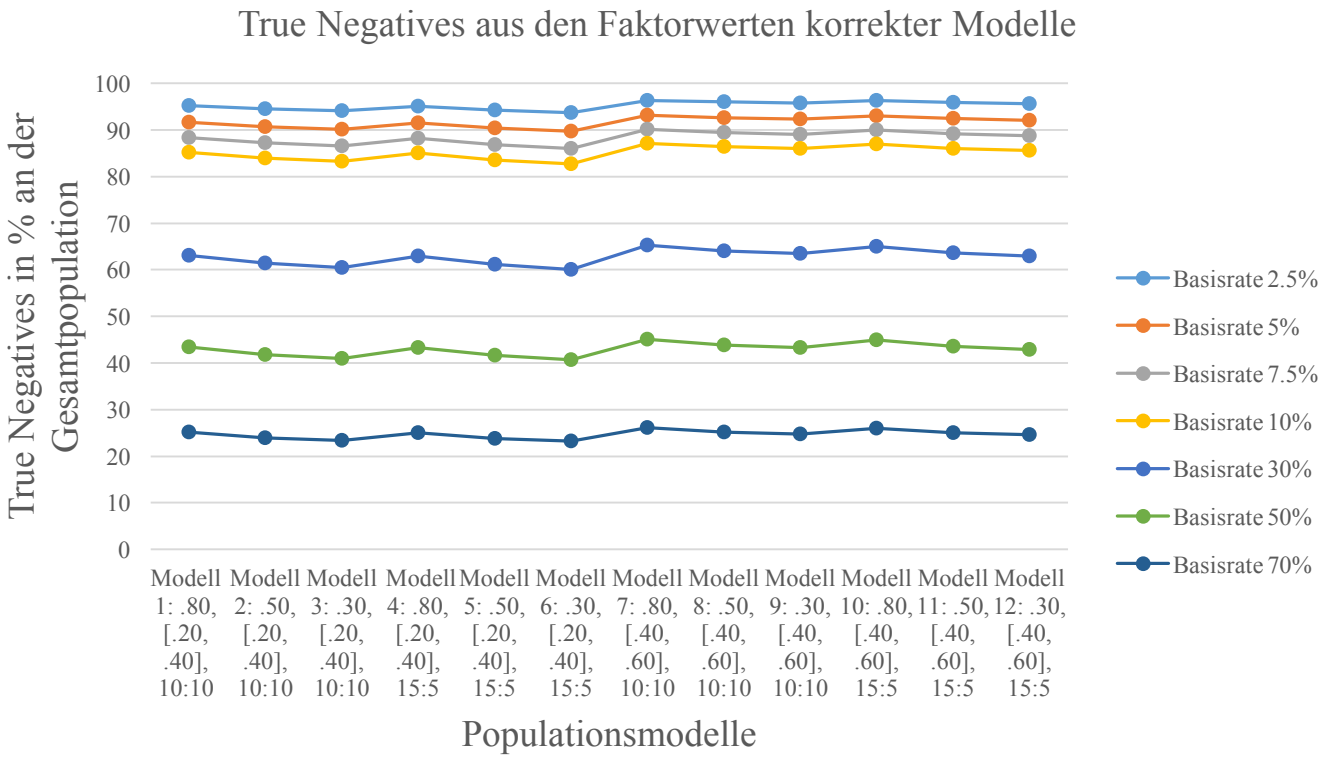

Abbildung 15. Durch die Bartlett-Faktorwerte korrekter Modelle korrekt erkannte Negative

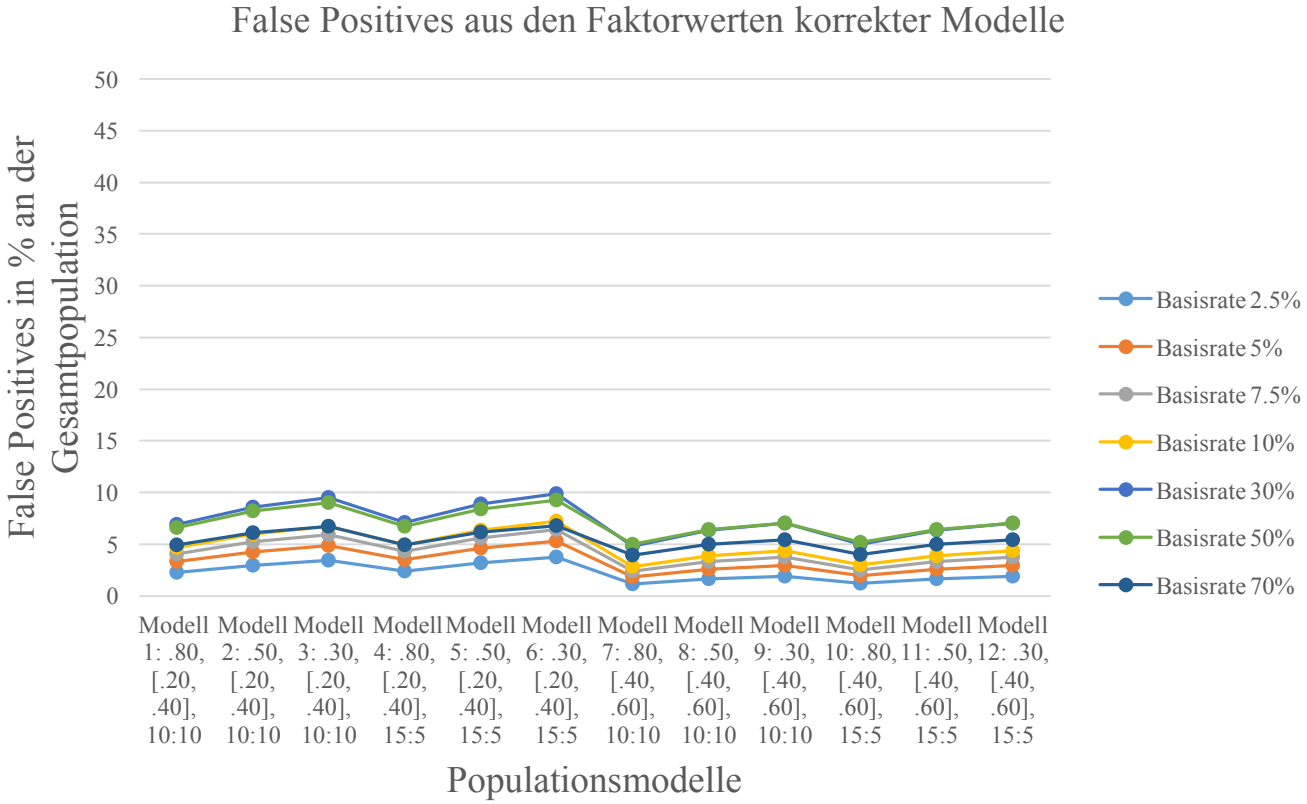

Abbildung 16. Durch die Bartlett-Faktorwerte korrekter Modelle als falsch-positiv diagnostizierte Fälle

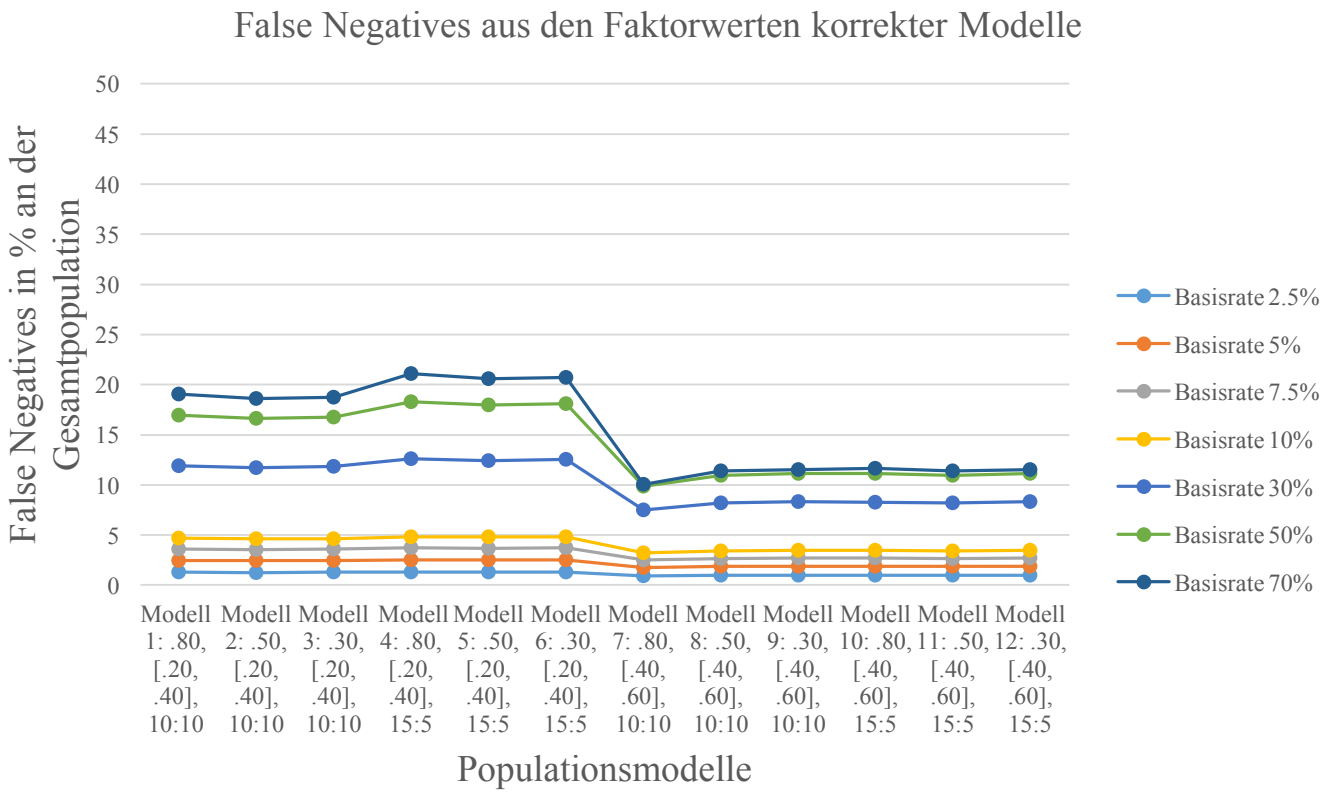

Abbildung 17. Durch die Bartlett-Faktorwerte korrekter Modelle als falsch-negativ diagnostizierte Fälle

2 Diagnostische Konsistenzen missspezifizierte Modelle

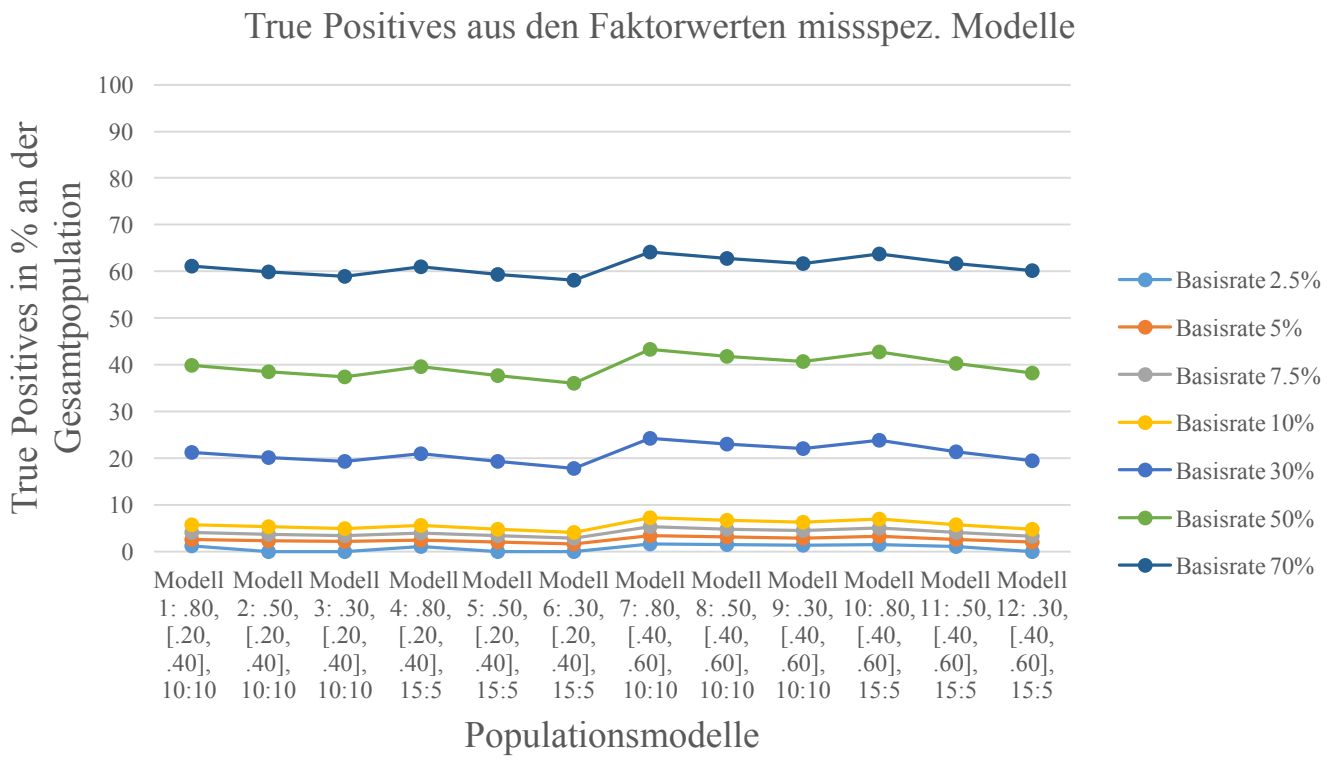

Abbildung 18. Durch die Bartlett-Faktorwerte missspezifizierter Modelle korrekt erkannte Positive

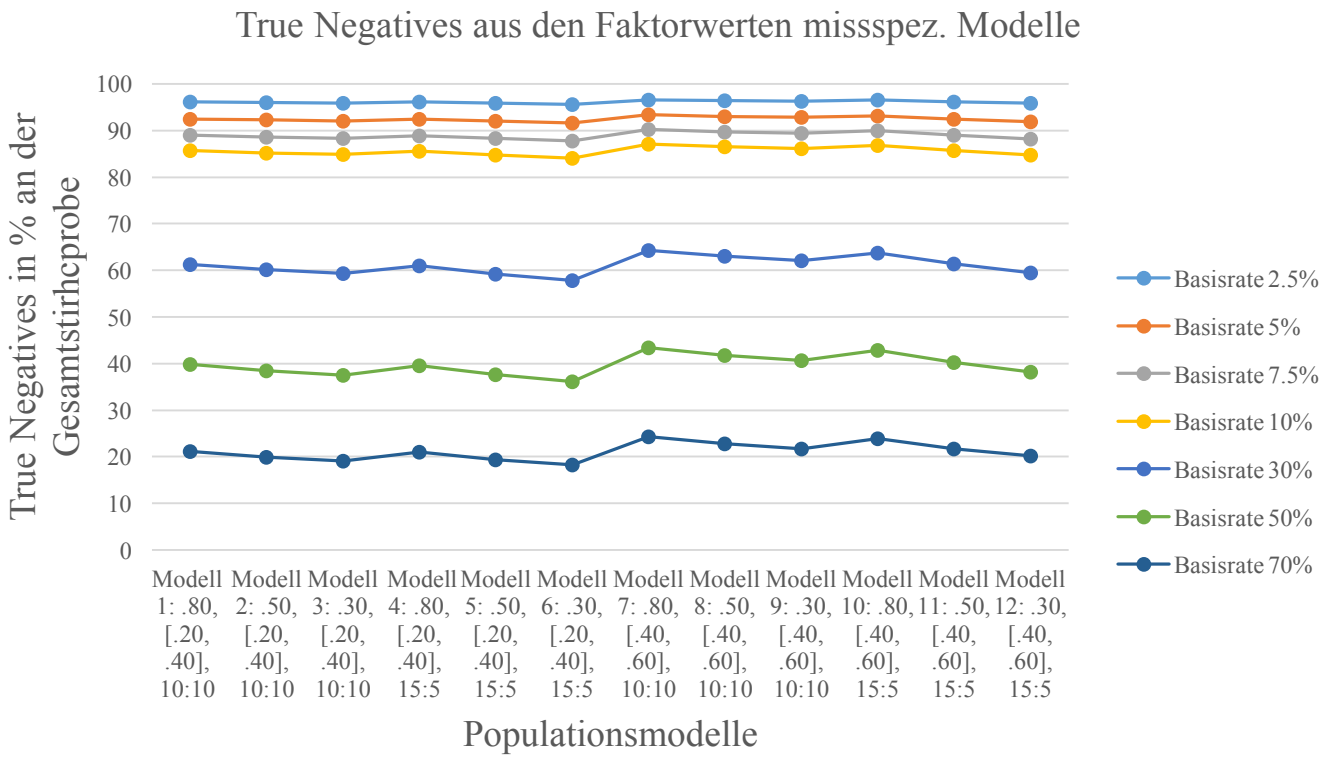

Abbildung 19. Durch die Bartlett-Faktorwerte missspezifizierter Modelle korrekt erkannte Negative

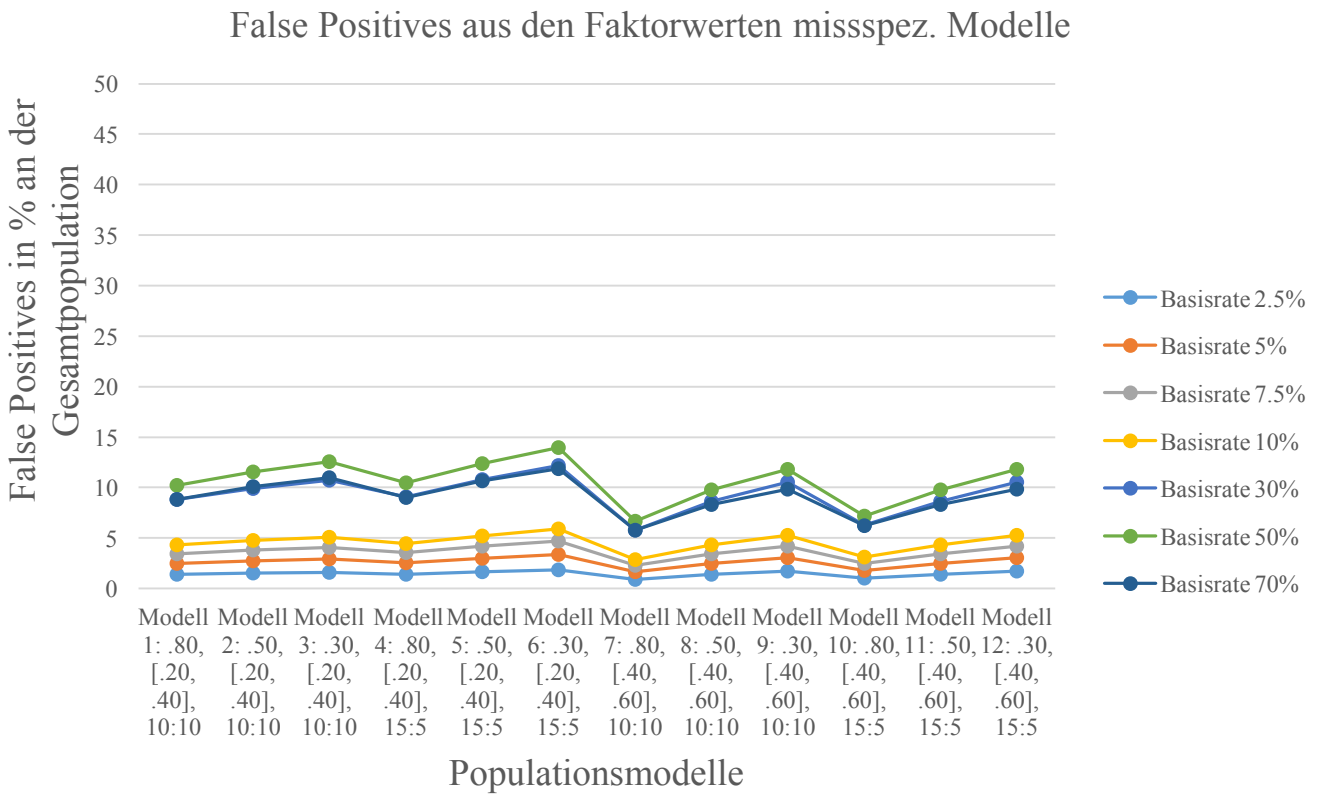

Abbildung 20. Durch die Bartlett-Faktorwerte missspezifizierter Modelle als falsch-positiv diagnostizierte Fälle

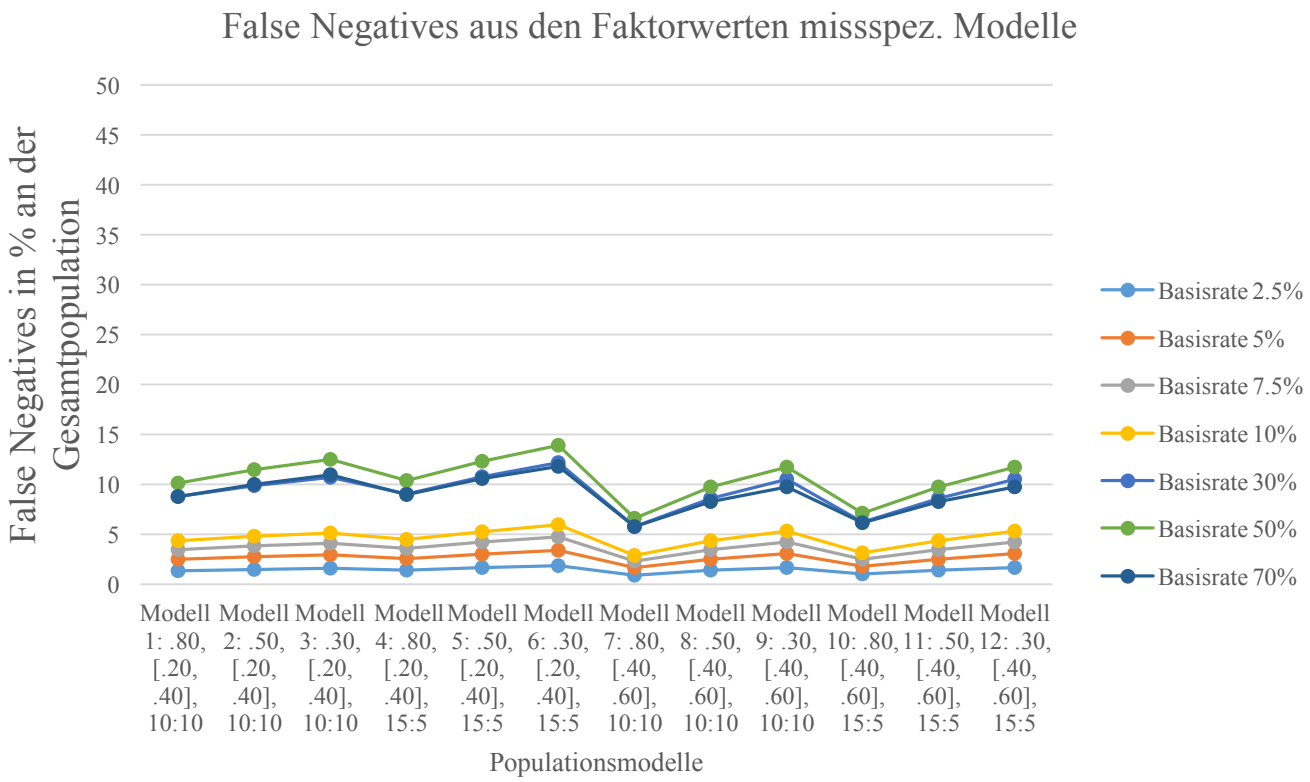

Abbildung 21. Durch die Bartlett-Faktorwerte missspezifizierter Modelle als falsch-negativ diagnostizierte Fälle

3 Diagnostische Konsistenzen Gesamtsummenwerte

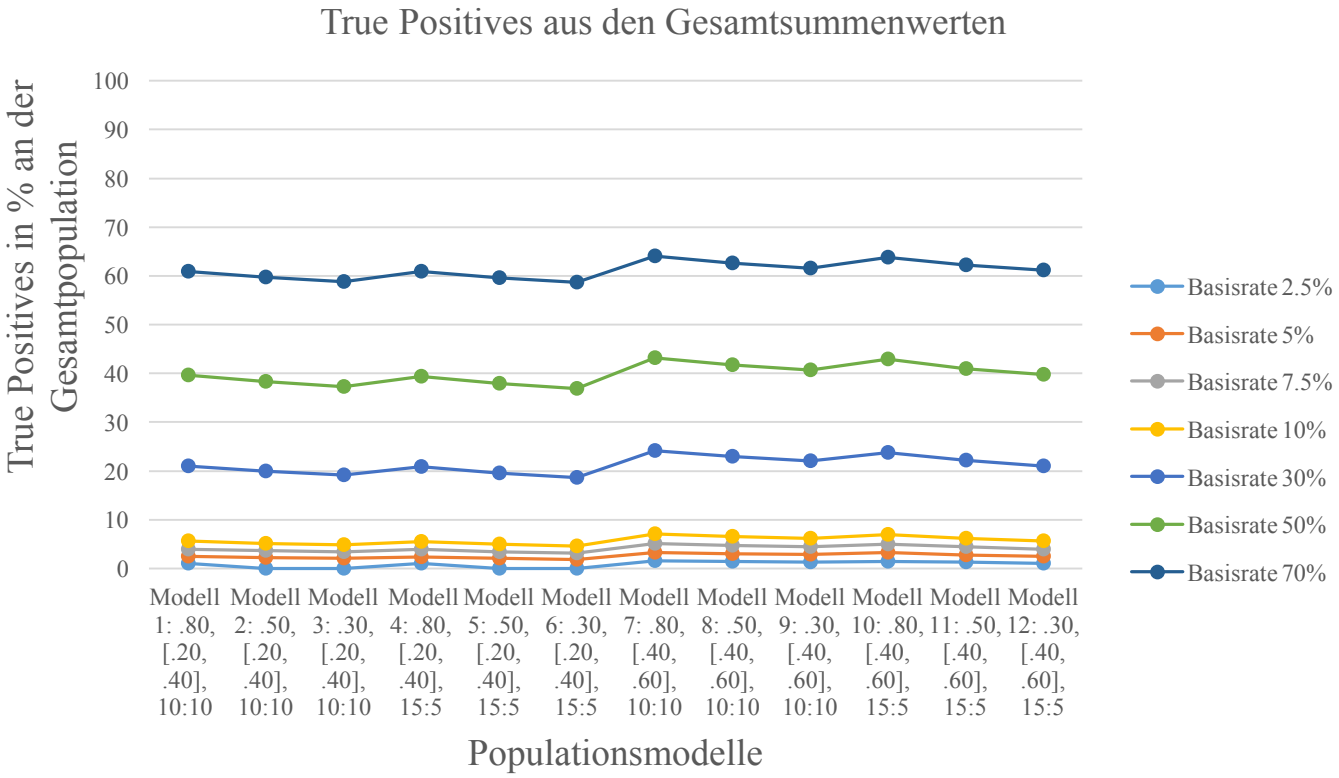

Abbildung 22. Durch die Gesamtsummenwerte korrekt erkannte Positive

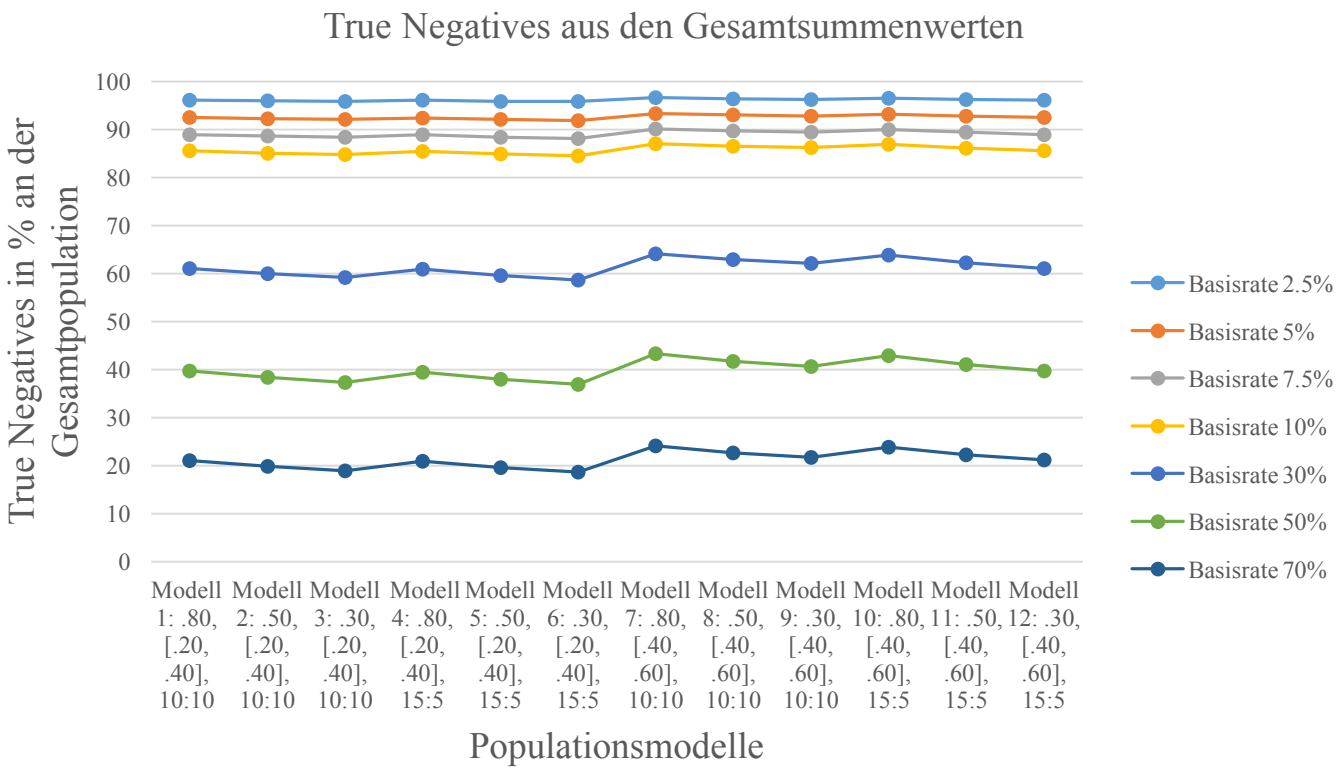

Abbildung 23. Durch die Gesamtsummenwerte korrekt erkannte Negative

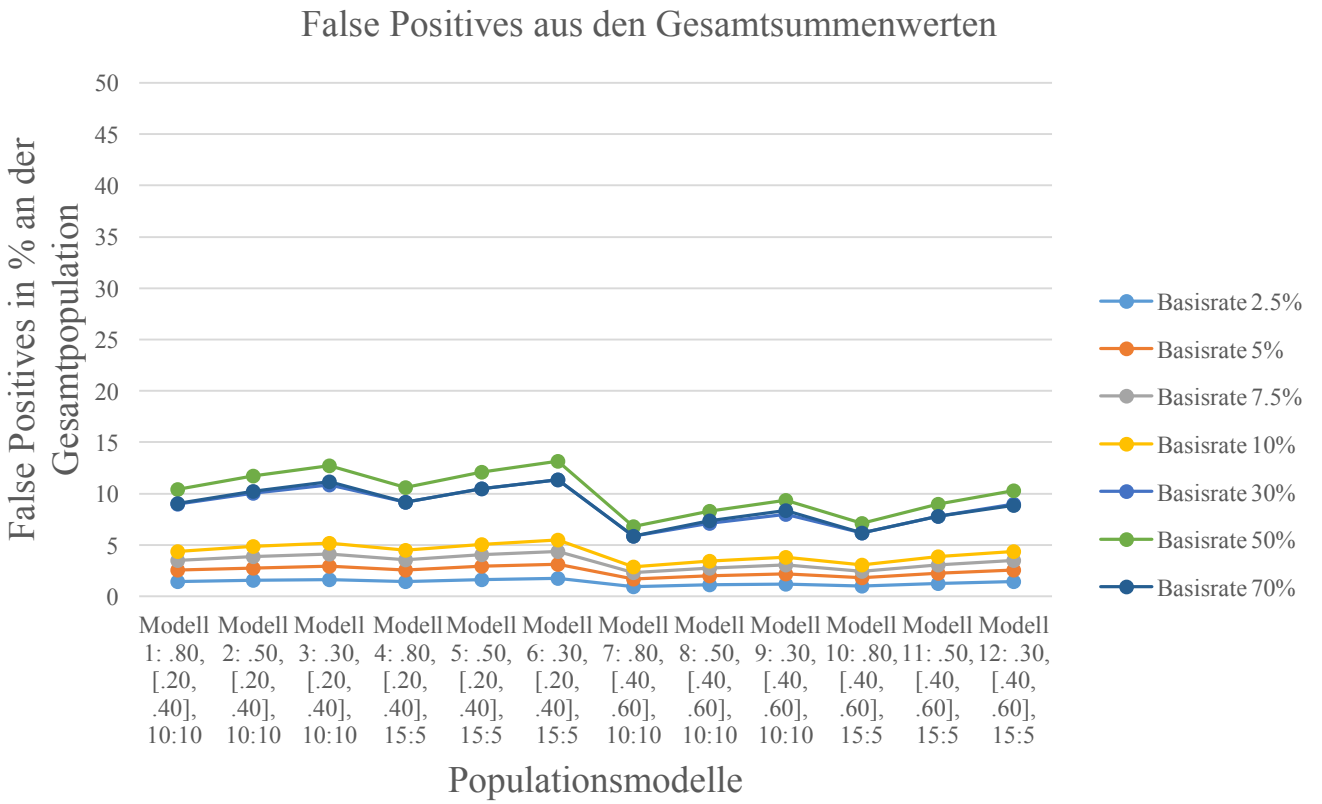

Abbildung 24. Durch die Gesamtsummenwerte als falsch-positiv Diagnostizierte

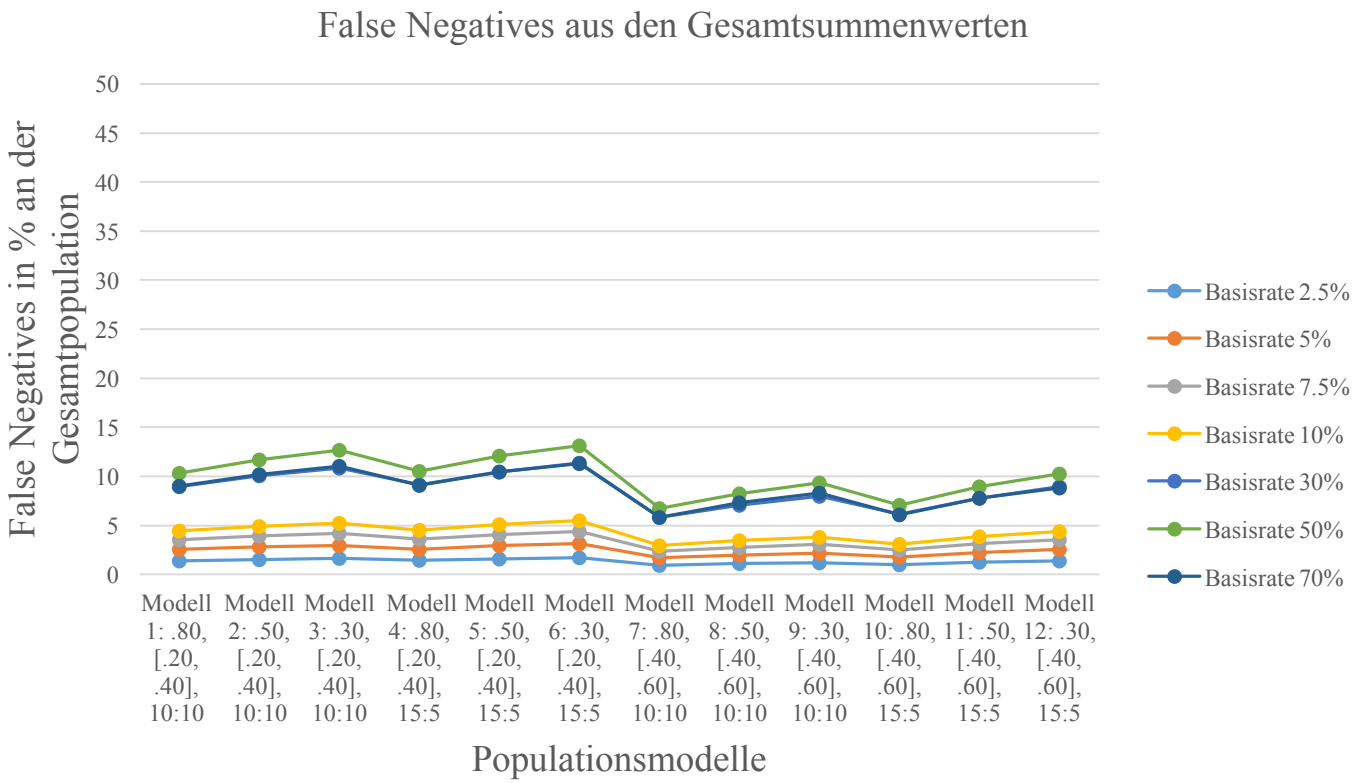

Abbildung 25. Durch die Gesamtsummenwerte als falsch-negativ Diagnostizierte