LISREL/CFA: Modelltest

im Rahmen des Interdisziplinären Seminars
Multivariate Statistik bei psychologischen Fragestellungen

Martina Feilke, Martina Unterburger, Christoph Burkhardt

Dozenten: Prof. Dr. Markus Bühner und Prof. Dr. Helmut Küchenhoff

3.Dezember 2008

Inhaltsverzeichnis

Kapitel 1

Wie gut ist die Modellanpassung?

1.1 Konfirmatorische Faktorenanalyse (CFA)

Die konfirmatorische Faktorenanalyse (CFA) ist ein Spezialfall der linearen Strukturgleichungsmodelle (LISREL). Die in diesem Vortrag vorgestellten Methoden werden hier zwar auf den Spezialfall CFA bezogen, können im Allgemeinen aber auf lineare Strukturgleichungsmodelle verallgemeinert werden.
Das Ziel der konfirmatorischen Faktorenanalyse (CFA) ist, theoretisch oder empirisch gut fundierte Modelle auf ihre Modellgüte zu testen.
Die Fragestellung dabei lautet: Wie gut passt das theoretisch aufgestellte Modell zu den beobachteten Daten / zur Stichprobe?

1.2 Exakter Modell-Fit

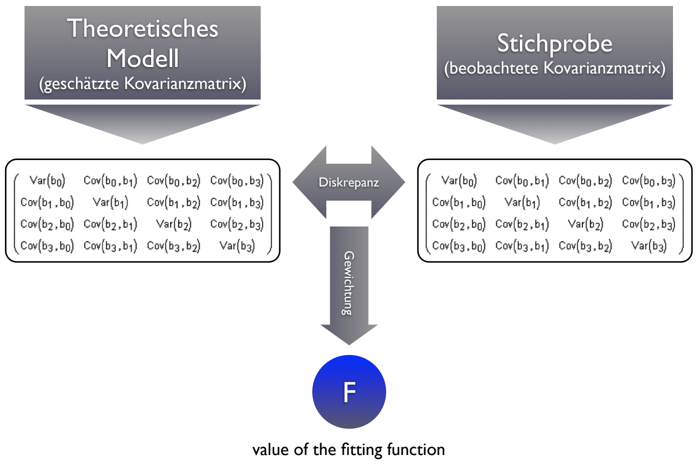
1.2.1 Exakter Modell-Fit - Testkonstruktion

Die Überprüfung der exakten Modellanpassung erfolgt mit dem χ^2-Test.

Hierbei lauten die Hypothesen:
H_0: Das Modell passt zur Datenstruktur. (Nullhypothese)
H_1: Das Modell weicht von der Datenstruktur ab. (Alternativhypothese)

Unter H_0 wird ein bestimmten Modell spezifiziert. H_1 repräsentiert im Gegensatz dazu nicht <u>ein</u> bestimmtes Modell. Dabei meint H_1, dass eine "beliebige" Kovarianzstruktur vorliegt, die nicht zum spezifizierten Modell passt.

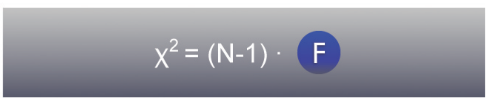
1.2.2 Berechnung des χ^2-Werts



Die Diskrepanz zwischen beobachteter (aus Stichprobe) und geschätzter (implizierter) Kovarianzmatrix wird gewichtet und zu einem Wert F zusammengefasst. Diesen Wert F bezeichnet man als "value of the fitting function".

Im letzten Vortrag wurden drei verschiedene Diskrepanzfunktionen (="fitting functions") vorgestellt, mit denen der F-Wert berechnet werden kann: OLS-, GLS- und ML-Diskrepanzfunktion. Das Prinzip der Diskrepanzfunktionen ist eine Minimierung des "Abstands" (der Diskrepanz) zwischen der beobachteten und der geschätzten Kovarianzmatrix.
Je kleiner also der F-Wert ist, umso besser stimmen beobachtete und geschätzte Kovarianzmatrix überein. Im Idealfall sollte der F-Wert gleich null sein, denn dies würde bedeuten, dass das vom Anwender aufgestellte Modell mit den beobachteten Daten perfekt übereinstimmt.

Mithilfe des F-Werts kann nun der empirische χ^2-Wert (=Testgröße) berechnet werden.



Idee des χ^2-Tests:
Unter H_0 wird ein bestimmtes Modell spezifiziert, es werden somit auch bestimmte Restriktionen getroffen. Wenn das Modell gilt, dann gelten auch die für das Modell getroffenen Restriktionen. Der χ^2-Test testet also auch, ob die Restriktionen stimmen.

Likelihood-Ratio-Rationale:
Unter H_0 liegt ein restringiertes Modell vor. Unter H_1 werden keinerlei Restriktionen vorgenommen. Mithilfe der Teststatistik wird getestet, ob H_0 eine größere Likelihood hat als H_1, also "besser" zu den Daten "passt".
Wenn die Likelihood von H_0 größer ist als die Likelihood von H_1, dann wird die Teststatistik klein bzw. negativ und es ist somit wahrscheinlich, dass sie kleiner ist als der kritische Wert der χ^2-Verteilung, also dass H_0 beibehalten wird. Wenn aber die Likelihood von H_1 größer ist als die Likelihood von H_0, dann wird der Wert der Teststatistik groß und positiv, also evtl. auch größer als der entsprechende kritische Wert der χ^2-Verteilung und somit ist es wahrscheinlicher, dass H_0 abgelehnt wird.

Wie aus der Formel zur Berechnung ersichtlich wird, ist der χ^2-Wert abhängig vom F-Wert und vom Stichprobenumfang N. Er fungiert als Testgröße und wird mit dem (1-α)-Quantil der χ^2-Verteilung (mit entsprechender Anzahl an Freiheitsgraden) verglichen.

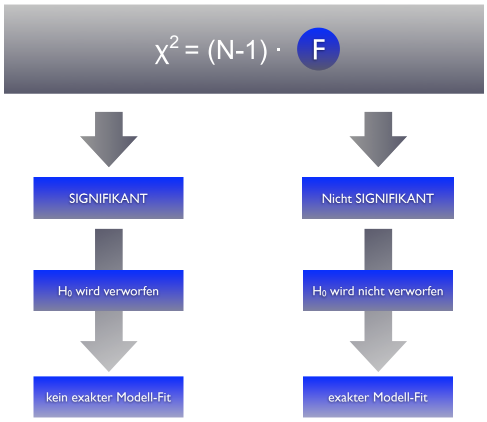
1.2.3 Testentscheidung

Berechnung des (1-α)-Quantils der χ^2-Verteilung:

Die Berechnung der Freiheitsgrade (df) erfolgt mit der Formel:

$$\mathrm{df} = \{[b \cdot \tfrac{b+1}{2}] - f\}$$

wobei b die Anzahl der Items darstellt und f die Anzahl der frei zu schätzenden Parameter im Modell (= Ladungen + Fehlervarianzen). Nach dem Einsetzen des zuvor festgelegte α's kann $\chi^2_{1-\alpha}(df)$ aus der Tabelle der χ^2-Verteilung entnommen werden.
Ist der berechnete χ^2-Wert (Testgröße) größer als das (1-α)-Quantil der χ^2-Verteilung (mit entsprechender Anzahl an Freiheitsgraden), dann wird die Nullhypothese H_0 abgelehnt, es trifft also die Alternativhypothese H_1 zu. Das Modell weicht signifikant von der beobachteten Datenstruktur ab (signifikantes Testergebnis).



Andernfalls, wenn die Testgröße kleiner ist als das entsprechende Quantil, wird die Nullhypothese (H_0: Das Modell passt zur beobachteten Datenstruktur) beibehalten. Man kann davon ausgehen, dass nichts Wesentliches gegen das Modell spricht.
Wurde der χ^2-Test durchgeführt, so sollte immer zusammen mit dem χ^2-Wert auch der p-Wert angegeben werden.

"If the model fits the data, it does not mean that it is the correct model or even the "best" model. In fact, there can be many equivalent models all of which will fit the data equally well as judged by any goodness of fit measure. To conclude that the fitted model is the "best", one must be able to exclude all models equivalent to it on logical or substantive grounds"
(Jöreskog, 1993, S.298; vgl. Cole, 1987)

Die Annahme der Nullhypothese bedeutet nicht zwingend, dass das Modell das richtige oder gar "wahre" Modell ist. Dies kann aufgrund von Hypothesentests nicht entschieden werden.

1.2.4 Probleme bei Berechnung des χ^2-Werts

Minimierung des β-Fehlers

H_0: Das Modell passt zur Datenstruktur. (Nullhypothese)
H_1: Das Modell weicht von der Datenstruktur ab. (Alternativhypothese)

$\alpha - Fehler$ (Fehler 1.Art): Entscheidung: Kein exakter Modell-Fit
Realität: Exakter Modell-Fit
$\beta - Fehler$ (Fehler 2.Art): Entscheidung: Exakter Modell-Fit
Realität: Kein exakter Modell-Fit

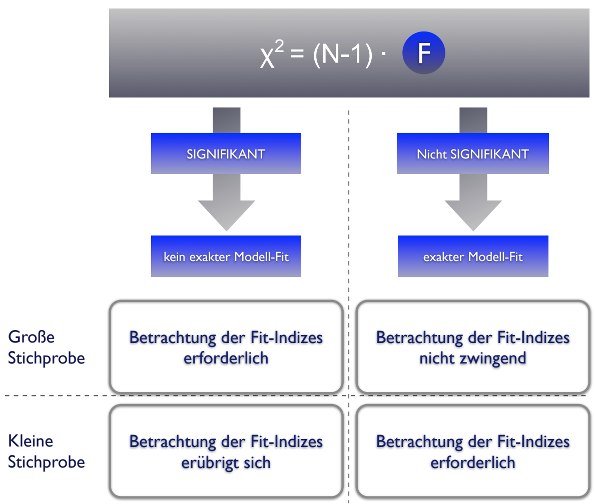
Die Wahrscheinlichkeit dafür, dass eine Entscheidung für exakten Modell-Fit eine Fehlentscheidung ist, soll möglichst gering gehalten werden. Dies entspricht einer Minimierung des β-Fehlers. Durch einen etwas höher gesetzten α-Fehler kann eine Minimierung des β-Fehler erreicht werden.

Abhängigkeit von der Stichprobengröße

Der χ^2-Wert hängt ab von der Größe der Stichprobe (N). Je größer die Stichprobe ist, desto sensitiver ist der χ^2-Test. Das bedeutet:

- Bei großen Stichproben führen bereits kleine Abweichungen von einem perfektem Modell zur Ablehnung von H_0 (Nullhypothese).

- Bei kleinen Stichproben können große Abweichungen von einem perfektem Modell trotzdem zur Annahme von H_0 (Nullhypothese) führen.



Je nach Stichprobenumfang und Testentscheidung ist somit eventuell eine zusätzliche Betrachtung der Fit-Indizes notwendig.

1.3 Approximativer Modell-Fit

1.3.1 Motivation für einen idealen Fit-Index

Aufgrund der zuvor angesprochenen Nachteile des χ^2-Werts ergibt sich die Notwendigkeit weitere Maßzahlen für die Beurteilung des Modell-Fits heranzuziehen. Dabei werden an einen **idealen Fit-Index** folgende Anforderungen gestellt:

1. Zunächst sollte ein Fit-Index auf einen bestimmten Wertebereich, idealerweise auf [0,1], normiert sein. Durch die Festlegung eines maximalen und eines minimalen Wertes wird der Index erst vergleichbar.

2. Weiterhin sollte ein solcher Fit-Index eine geringe Sensitivität gegenüber Stichprobengrößen aufweisen, da die Beurteilung der Modellgüte unabhängig vom Stichprobenumfang erfolgen soll.

1.3.2 Arten von Fit-Indizes

Es existieren zwei Arten von Fit-Indizes.

1. Zum einen gibt es die **komparativen Fit-Indizes**. Dabei wird das getestete Modell mit einem "Null-Modell" verglichen, in dem davon ausgegangen wird, dass alle Variablen unkorreliert sind.

2. Bei den **absoluten Fit-Indizes** wird das getestete Modell mit einem "saturiertem Modell" verglichen. Ein "saturiertes Modell" ist ein Modell mit perfektem Fit.

Weiterhin können Fit-Indizes in **Goodness-of-Fit-Indizes** und **Badness-of-Fit-Indizes** unterteilt werden.

1. Ein **Goodness-of-Fit-Index** gibt an, wie gut ein Modell die Daten beschreibt. Große Werte zeigen dabei einen guten approximativen Modell-Fit an.

2. **Badness-of-Fit-Indizes** spiegeln wider, wie schlecht ein Modell die Daten beschreibt. Hierbei steht ein kleiner Wert des Index für einen guten approximativen Modell-Fit.

Im Rahmen dieses Vortrags wird exemplarisch für einen komparativen Fit-Index der CFI vorgestellt. Als Vertreter der absoluten Fit-Indizes werden RMSEA und SRMR behandelt. Diese Indizes sind alle auf den Wertebereich [0,1] normiert. Neben diesen Fit-Indizes gibt es noch zahlreiche andere, die teilweise auch Werte größer als 1 annehmen können. Die für diesen Vortrag ausgewählten Indizes haben sich als in der Psychologie gängige Indizes erwiesen.

1.3.3 Komparativer Fit-Index: CFI

Bei dem **CFI** (**C**omparative-**F**it-**I**ndex) handelt es sich um einen Godness-of-Fit-Index, d.h. ein Wert nahe 1 signalisiert einen guten approximativen Modell-Fit.
Folgende Formel wird zur Berechnung des CFI verwendet:

$$\mathrm{CFI} = 1 - \frac{\chi_M^2 - df_M}{\chi_N^2 - df_N}$$

In diese Formel geht zum einen im Zähler die Diskrepanz zwischen beobachteter und geschätzter Kovarianzmatrix ein (χ_M^2). Zum anderen findet im Nenner die Diskrepanz zwischen beobachteter Kovarianzmatrix und Kovarianzmatrix des Nullmodels Eingang (χ_N^2). Dieser Bruch wird dann von 1 abgezogen.
Ist das geschätzte Modell besser als das Nullmodell, so ist die Diskrepanz im Zähler kleiner als im Nenner. Der Bruch wird somit klein und folglich liegt der CFI nahe 1 (CFI = 1 - kleiner Wert $\approx$ 1), was für einen (sehr) guten approximativen Modell-Fit spricht. Sind geschätztes Modell und Nullmodel (in etwa) gleich, d.h. das geschätzte Modell bringt gegenüber dem Nullmodel (nahezu) keine Verbesserung, so entspricht die Diskrepanz im Zähler (in etwa) der Diskrepanz im Nenner. Folglich wird der Bruch (ungefähr) 1 und der CFI nimmt (nahezu) den Wert 0 an (CFI = 1 - 1 = 0), was für einen (sehr) schlechten approximativen Modell-Fit spricht.
Vorteile des CFI sind vor allem die geringe Sensitivität gegenüber Stichprobenumfängen und die Robustheit gegenüber der Verletzung der Verteilungsannahme.

1.3.4 Absoluter Fit-Index: RMSEA

Der **RMSEA** (**R**oot-**M**ean-**S**quare-**E**rror of **A**pproximation) ist ein Badness-of-Fit-Index. Somit sprechen kleine RMSEA-Werte, nahe 0, für einen guten approximativen Modell-Fit.
Zur Berechnung des RMSEA dient folgende Formel:

$$\text{RMSEA} = \sqrt{\frac{\chi_M^2 - df_M}{N \cdot df_M}}$$

Sollte $df > \chi_M^2$ gelten, also die Anzahl der Freiheitsgrade (df_M) größer als der χ_M^2-Wert sein, so wird konventionell der Zähler gleich 0 gesetzt (um einen negativen Wert unter der Wurzel zu vermeiden).
Kernstück des RMSEA ist der χ_M^2-Wert. Je größer der χ_M^2-Wert ist, also je größer die Diskrepanz zwischen beobachteter und geschätzter Kovarianzmatrix, desto größer ist auch der RMSEA-Index und umso schlechter der approximative Modell-Fit.
Neben dem χ_M^2-Wert berücksichtigt der RMSEA-Index zusätzlich den Stichprobenumfang N sowie die Modellkomplexität, die sich in den Freiheitsgraden (df_M) niederschlägt. Je komplexer ein Modell ist, umso geringer ist die Anzahl der Freiheitsgrade. Betrachtet man die Formel des RMSEA hinsichtlich dieses Aspekts, so wird bei komplexen Modellen im Nenner mit einer kleineren Zahl (df_M) multipliziert als bei einfacheren Modellen. Dadurch wird bei komplexen Modellen der RMSEA-Index insgesamt größer. Somit folgt: je komplexer ein Modell (bei festem χ_M^2-Wert und Stichprobenumfang) ist, umso größer der RMSEA.
Vorteilhaft ist die explizite Berücksichtigung der Modellkomplexität in der Formel zur Berechnung des RMSEA.
Eine Besonderheit des RMSEA ist, dass zusätzlich zum Wert dieses Fit-Index auch noch ein Konfidenzintervall berechnet werden kann.

1.3.5 Absoluter Fit-Index: SRMR

Der **SRMR** (**S**tandardized-**R**oot-**M**ean-**R**esidual) wird den Badness-of-Fit-Indizes zugeordnet. Somit signalisiert ein kleiner SRMR-Wert, nahe 0, einen guten approximativen Modell-Fit.
Der SRMR wird durch folgende Formel berechnet:

$$\text{SRMR} = \sqrt{\Sigma_j \, \Sigma_{k<j} \, \frac{r_{jk}^2}{\frac{b \cdot (b+1)}{2}}}$$

Dabei steht r_{jk} für die Differenz zwischen dem beobachteten und dem geschätzten Korrelationskoeffizient von Item j und Item k. Wie bereits zuvor steht b für die Anzahl der Items.
Im Gegensatz zum RMSEA, berücksichtigt der SRMR nicht die Modellkomplexität. Vorteilhaft ist beim SRMR allerdings seine geringe Sensitivität gegenüber des Stichprobenumfangs.
Der SRMR ist der einzige hier vorgestellte Fit-Index, in dessen Berechnung der χ^2-Wert nicht eingeht.

1.3.6 Cut-off-Werte (aus Simulationsstudien)

In der Literatur werden für die zuvor beschriebenen Indizes folgende Cut-off-Werte angegeben (nur geeignet bei Verwendung der ML-Diskrepanzfunktion):

- CFI ≈ 0.95
- RMSEA ≤ 0.06 (N $>$ 250) bzw. ≤ 0.08 (N $<$ 250)
- SRMR < 0.11

Allerdings ist zu beachten, dass über diese Cut-off-Werte eine rege Diskussion entbrannt ist und diese stark umstritten sind. Da solche Cut-off-Werte nicht beliebig generalisierbar sondern extrem situationsabhängig sind, darf nicht vergessen werden, dass sie ausschließlich als Daumenregeln zu verwenden sind, wie folgendes Zitat treffend beschreibt:

"Decision rules such as $RMSEA < 0.05$ (...) may be useful in some solutions but they often lead to inappropriate decisions in other solutions (...) and should be considered only as rules of thumb."
(Hu & Bentler, 1999, S.4)

Beispielsweise können bei der Analyse von Fragebögen mit vielen Items diese strengen Cut-off-Werte meist nicht eingehalten werden und moderatere Cut-off-Grenzen sind durchaus vertretbar (vgl. Bühner 2006).

1.4 Zusammenfassung Modell-Fit

Um die Vor- und Nachteile der einzelnen Fit-Indizes auszugleichen, ist zu empfehlen, Fit-Indizes aus unterschiedlichen Kategorien einzubeziehen. So sollten zur Beurteilung der Modellgüte sowohl absolute und komparative Fit-Indizes als auch Goodness-of-Fit- und Badness-of-Fit-Indizes herangezogen werden. Zu Vermeiden ist eine willkürliche Auswahl derjenigen Indizes, die im Rahmen der vorliegenden Untersuchung gerade die besten passenden Werte aufweisen. In der Literatur (vgl. Bühner 2006) wird vor diesem Hintergrund zur Beurteilung des Modell-Fits häufig eine Kombination aus χ^2-Wert (incl. seines zugehörigen p-Werts), CFI, RMSEA und SRMR empfohlen.
Zu beachten ist bei allen Indizes allerdings stets, dass theoretische und logische Überlegungen unverzichtbar sind.

"It is important to remember that even a model with excellent overall fit indices can be unacceptable because of the components of the model"
(Bollen & Long, 1993, S.7)

Das vorangehende Zitat macht deutlich, dass ein Modell mit noch so guten Anpassungsmaßen trotzdem nicht unbedingt ein "gutes" Modell sein muss. Überlegungen, ob das Modell im Rahmen der Psychologie theoretisch und logisch sinnvoll ist, sind vor der Betrachtung der Modellgüte unabdingbar.

Kapitel 2

Testkonstruktion mit der CFA

2.1 Reliabilität

2.1.1 Korrelierte Fehlervariablen

Die klassische Testtheorie geht davon aus, dass Messfehler unkorreliert vorliegen solange keine systematischen Einflüsse von außen auf das Konstrukt wirken, die im Modell nicht spezifiziert sind. In der CFA ist es möglich, korrelierte Fehler zu spezifizieren und das Modell auf seine statistische Signifikanz zu testen. Da dieses Vorgehen jedoch sehr umständlich ist, sollte man die Modifikationsindizes beachten, die Aufschluss über mögliche korrelierte Fehler geben können. Die Modifikationsindizes zeigen an, um wie viel sich der Chi-Quadrat-Wert des Modelltests verändert (oft: verringert), wenn man eine spezifische Modifikation vornimmt. Problematisch an dieser Art der Modellanpassung ist, dass die statistische Verbesserung der Signifikanz eines Modells nicht inhaltlich gerechtfertigt werden kann. Da das Verfahren eher einer explorativen als einer konfirmatorischen Analyse ähnelt, ist eine Kreuzvalidierung des modifizierten Modells auch hinsichtlich der theoretischen und inhaltlichen Interpretation notwendig.

2.1.2 Konstruktreliabilität

Allgemein meint Reliabilität die "Zuverlässigkeit" von Messungen, d.h. Messwiederholungen liefern konsistente und stabile Resultate.
Die Konstruktreliabilität kennzeichnet die Qualität, mit der ein Konstrukt - also eine latente Variable - gemessen wird. Eine Schätzung der Reliabilität liefern zum einen die Kommunalität der Items (diese stellen eine Mindestschätzung dar), zum anderen die innere Konsistenz (Cronbach-Alpha). Die Konstruktreliabilität lässt sich bestimmen über:

$$H = \frac{1}{1+\left(1/\left[\sum_{i=1}^{p}\frac{a_i^2}{1-a_i^2}\right]\right)}$$

Als Richtwert sollte sie über H = 0.70 liegen.
Wenn einzelne Konstrukte nicht reliabel erfasst werden, kann es zu Interpretationsproblemen kommen. So können die latenten Variablen (Konstrukte) aus unterschiedlichen Gründen korreliert sein. Einerseits können einzelne Items auf fremden Faktoren laden, es könnte aber andererseits auch sein, dass die einzelnen Skalen (z.B. eine Skala für Intelligenz und eine für Persönlichkeit) unterschiedlich reliabel sind. In diesem Fall muss die Korrelation zwischen den latenten Variablen minderungskorrigiert werden. In der CFA wird die Reliabilität über eine Regression eines Items auf die latente Variable ermittelt.

2.2 Aggregation von Items

Aus ökonomischen Gründen lassen sich Items unter bestimmten Bedingungen zu Aggregaten zusammenfassen. Die Bildung der Aggregate kann mit Hilfe von einer exploratorischen oder einfaktoriellen konfirmatorischen Faktorenanalyse erfolgen. Die Items eines Aggregats sollten möglichst homogen sein und natürlich sollten alle die gleiche latente Variable erfassen. Weiterhin sollten alle Aggregate, die einer latenten Variable zugeordnet werden, ebenfalls homogen sein. Das resultierende Aggregat ist reliabler als die Einzelitems, es hat eine höhere Kommunalität, es weist einen höheren Prozentsatz gemeinsamer Varianz im Vergleich zur spezifischen Varianz auf und besitzt zudem meist eine günstigere Verteilung als die Einzelitems. Weiterhin von Vorteil ist, dass die Anzahl der freien Parameter dadurch reduziert wird und somit weniger Ladungen und Fehlervarianzen zu schätzen sind.

2.3 Multi-Trait-Multi-Method Ansatz

Ziel dieses Ansatzes ist es zu erkennen, wie viel Varianz der Leistung einer Person auf ihre Kompetenzen (Trait) zurückgeht und wie viel nur durch das Messinstrument (Methode) verursacht wird. Wünschenswert wäre eine geringe Methodenvarianz. In einem Assessment Center beispielsweise werden oft die gleichen Kompetenzen über verschiedene Methoden erhoben. Die hieraus resultierende Varianz soll später herausgerechnet werden. Dies ist in der CFA abbildbar. Zwei Ansätze sind denkbar. Die Modelle dazu sind das

Correlated-Uniqueness-Model (CUM) und das Correlated-Trait-Correlated-Method-Model (CTCM). Im CU-Modell sind korrelierte Fehler so spezifiziert, dass jeweils die Fehler, die durch ein- und dieselbe Methode verursacht werden, als korreliert angenommen werden. Die Methode wird hier also aus den Fehlervarianzen herausgerechnet. Im CTCM-Modell werden anstelle von korrelierten Fehlern Methodenfaktoren spezifiziert, die einen Teil der Varianz an der Leistung einer Person erklären können. Die Gesamtvarianz der Items setzt sich hierbei also aus der systematischen Varianz (Kompetenz- oder Traitvarianz), aus der Methodenvarianz, die also durch die Methode zu erklären ist und auf die latenten Methodenfaktoren zurückzuführen ist, sowie aus der Fehlervarianz zusammen.

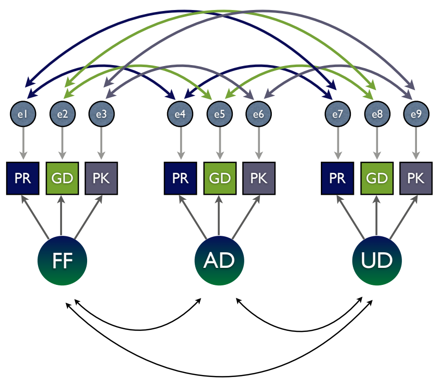


Abbildung 2.1: Correlated-Uniqueness-Model (CUM)

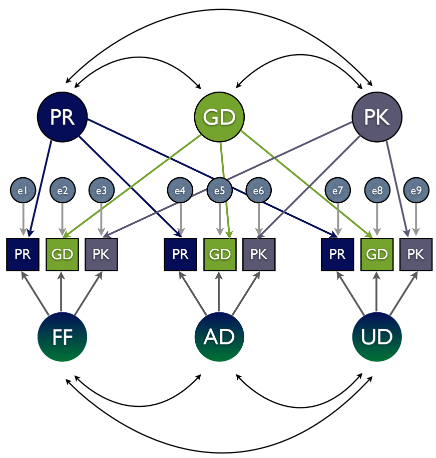


Abbildung 2.2: Correlated-Trait-Correlated-Method-Model (CTCM)

Kapitel 3

Zusammenfassung der CFA (Spezialfall von LISREL)

- Strukturgleichungsmodell aufstellen
- Datenerhebung und Berechnung der beobachteten Kovarianzmatrix (aus den Daten)
- Prüfung der Identifizierbarkeit und evtl. Treffen von Restriktionen
- Parameterschätzung (Ladungen und Fehlervarianzen)
- Berechnung der geschätzten Kovarianzmatrix aus Modellgleichung
- Modellbeurteilung (durch χ^2-Test und Fit-Indizes)
- evtl. Modellmodifikation

Literaturverzeichnis

[1] Buehner Markus, 2006: *Einführung in die Test- und Fragebogenkonstruktion*; Pearson Studium.

[2] 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.

[3] Fan X. & Thompson B. & Wang L., 1999: *Effects of Sample Size, Estimation Methods, and Model Specification on Structural Equation Modeling Fit Indexes*; Structural Equation Modeling, 6(1), 56-83.

[4] Bollen K.A. & Long J.S., 1993: *Testing Structural Equation Models*; S.7.

[5] Jöreskog K., 1993: *Testing Structural Equation Models*; in K.A. Bollen & J.S. Long (Eds.) Testing Structural Equation Models, 294-316.