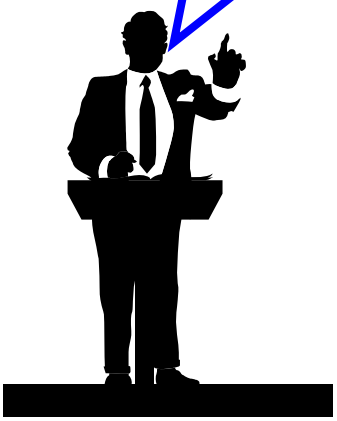
Die Messung mentaler Konstrukte:

Eine moderne Einführung in die klassischen Testtheorie und ihrer Weiterentwicklung

Datum der letzten Änderung: **Donnerstag, 12. November 2009**
Gültig bis Seite: **Ende**



Vorlesung
gehalten von
Dr. Siegfried Macho

Universität Fribourg
2009-2010

Inhaltsverzeichnis

1. Einführung: Wie misst man mentale Konstrukte?

In diesem Einführungskapitel versuche ich anhand einfacher Beispiele eine intuitive Einsicht zu vermitteln, wie die Messung mentaler Konstrukte vor sich geht. Im Zentrum dieser Darstellung steht das Konzept des *Messmodells*. Bevor ich dieses Konzept näher erläutere, werde ich die Bedeutung theoretischer Annahmen für die Messung mentaler Konstrukte erläutern.

1.1 Die Bedeutung theoretischer Vorannahmen

Die hier gegebene Darstellung basiert auf folgendem Credo:

Prinzip 1-1: *Die Bedeutung theoretischer Annahmen für die Messung:*

Jede Messung basiert auf theoretischen Annahmen über den Messprozess. Jede Messung ist daher theorieabhängig.

Die theoretischen Annahmen, welche einer Messung zugrunde liegen, lassen sich in zwei Gruppen unterteilen:

1. *Ontologisch-metaphysische Annahmen*: Diese betreffen die Existenz bestimmter Objekte, Eigenschaften und Relationen.
2. *Mess-theoretisch-statistische Annahmen:* Diese betreffen das Skalenniveau und zur Wahrscheinlichkeitsverteilung der involvierten Grössen.

Die ontologischen Annahmen betreffen im Wesentlichen:

1. Die zu messenden Konstrukte und deren Beziehungen untereinander.
2. Die Beziehungen zwischen Konstrukten und den konkreten Massen.
3. Annahmen über verschiedene Einflussgrössen, welche die Messung beeinflussen können.

Man beachte, dass *jede* Messung als theorieabhängig betrachtet wird und nicht nur die Messung theoretischer Konstrukte. Der Grund hierfür liegt in der Überzeugung, dass es keine reinen Beobachtungsgrössen gibt. Vielmehr wird jede Eigenschaft oder jedes Ding als mehr oder weniger beobachtungsnah betrachtet. Die folgende Aussage von Karl Popper bringt diese Tatsache auf den Punkt:

> *»**Es gibt keine reinen** Beobachtungen: sie sind von Theorien durchsetzt und werden von Problemen und Theorien geleitet.« (Popper, 1989; S. 76 [Hervorhebungen im Original]).*

Im Folgenden befassen wir uns die *ontologisch-metaphysischen* Annahmen (Die messtheoretisch-statistischen Annahmen werden später besprochen). Hierbei handelt es sich um Annahmen, welche das Vorhandensein bestimmter Dinge in der Welt betreffen.

Unter Prinzip 1-1 wurden drei wesentliche Aspekte, auf die sich ontologische Annahmen beziehen, erwähnt. Wie sehen diese nun im Falle der Messung mentaler Konstrukte im Rahmen des hier vertretenen Ansatzes konkret aus?

1.1.1 Ontologische Annahmen des Latenten-Variablen-Ansatzes

Der hier vertretene Ansatz wird als *Latenter-Variablen-Ansatz* bezeichnet (Borsboom, 2005; 2006). Mentale Konstrukte werden als latente Variablen betrachtet, welche kausalen Beziehungen sowohl untereinander als auch zu den beobachteten Messungen aufweisen. Dieser Ansatz ist in natürlicher Weise mit der metaphysischen Position des *Realismus* verbunden: Mentale Konstrukte, sowie deren kausale Beziehungen existieren und das Ziel einer Messung besteht darin, die Konstrukte sowie deren Beziehungen zu quantifizieren.

Die folgenden drei Annahmen sind zentral für den Latenten-Variablen-Ansatz:

1. *Annahmen über die zu messenden Konstrukte und deren Beziehung:*

   Mentale Konstrukte werden als real existierend angenommen. Konstrukte, wie Intelligenz, Persönlichkeit, soziale Kompetenz usw. beziehen sich auf reell existierende Fertigkeiten bzw. Eigenschaften von Personen, die jedoch nicht direkt beobachtbar sind. Diese Position wird als *Realismus* bezeichnet. Die Relationen zwischen mentalen Konstrukten wollen wir in zwei grobe Kategorien unterteilen:

   (i) *Kausale Beziehung:* Ein bestimmtes mentales Konstrukt übt einen kausalen Einfluss auf ein anderes Konstrukt aus. So beeinflusst z.B. die Leistung des Arbeitsgedächtnisses die Fähigkeit zur Einspeicherung neuer Informationen.

   (ii) *Unbestimmte Beziehung:* Es bestehen keine spezifischen Annahmen bezüglich der Beziehung zwischen zwei Konstrukten. Es wird jedoch angenommen, dass sich diese Beziehungen bei näherer Kenntnis durch kausale Relationen erklären lassen.

Bemerkung zum Verhältnis von Realismus und Kausalität:
Die Annahme, dass die wesentlichen Beziehungen zwischen Konstrukten kausaler Natur sind, setzt eine realistische Position voraus, da kausale Effekte nur zwischen real existierenden Dingen denkbar sind.

2. *Annahmen zur Relation zwischen Konstrukte und Messungen:*

   Die Beziehung zwischen abstrakten Konstrukten und konkreten Messungen wird ebenfalls als eine kausale betrachtet: Abstrakte Konstrukte üben eine kausalen Einfluss auf die Messung aus. Nur aufgrund dieser kausalen Beziehung wird die Messung des Konstrukts überhaupt möglich. Liegt keine derartige Beziehung vor, so ist keine valide Messung möglich.

3. *Annahmen über verschiedene Einflussgrössen auf die Messung:*

   Neben den zu messenden Konstrukten existieren im Normalfall alternativen Einflussgrössen, welche auf die Messung Einfluss nehmen. Diese lassen sich in zwei Kategorien einteilen:

   (i) *Kovariaten:* Hierbei handelt es sich um Grössen, deren Einfluss festgestellt und – idealer Weise – quantifiziert werden kann. Ein typisches Beispiel für eine derartige Kovariate ist die verwendete Methode. Diese kann einen grossen Einfluss auf die Messung ausüben.

   (ii) *Fehler:* Unter dieser Kategorie werden alle ungemessenen Einflussgrössen subsumiert. Ein wichtiger Aspekt von Messung besteht darin, die Grösse des Fehlers zu quantifizieren, denn eine Messung mit unbekanntem Messfehler ist praktisch wertlos.

Hinsichtlich des Auftretens von Messfehlern gilt das folgende Prinzip:

Prinzip 1-2: *Ubiquität von Messfehler:*

Messungen sind im Allgemeinen fehlerbehaftet.

1.1.2 Zum Unterschied zwischen dem Latenter-Variablen-Ansatz und dem True-Score-Ansatz der klassischen Testtheorie

Die klassische Testtheorie, wie sie in dem Klassiker von Lord und Novick (1968) entwickelt wird, beruht auf dem Konzept des *True-Score*. Das Problem dieser Konzeption besteht darin, dass die ontologischen Grundlagen dieses Konzepts nicht ausreichend klar formuliert wurden (siehe Borsboom, 2005). Es lassen sich (mindestens) drei mögliche Interpretationen von True-Scores unterscheiden, die hinsichtlich ihrer ontologischen Annahmen variieren:

1. Die operationalistische Konzeption von True-Scores:

Gemäss dieser Konzeption ist der True-Score der erwartete Testwert einer Person (Mr. Brown), die einen Test viele Male durchführt, wobei nach jedem Test eine »Gehirnwäsche« erfolgt, sodass die Erinnerung an den Test verloren geht.

Diese Konzeption ist insofern operationalistisch, als die Definition des True-Score auf ein Verfahren zur Ermittlung des True-Score reduziert wird. Man beachte jedoch, dass das Verfahren ein fiktives ist, welches in der Realität nicht durchgeführt werden kann, da der Test idealer Weise unendlich oft und nach jedem Test eine Gehirnwäsche durchgeführt werden müsste.

Man beachte, dass die operationalistische Konzeption keinerlei Annahmen über die Existenz latenter Konstrukte macht. Er entspricht daher einer zentralen Zielsetzung der Autoren, die klassische Testtheorie mit einem Minimum an theoretischen Annahmen zu entwickeln.

2. *True-Score als Erwartungswert einer personeninternen Propensitätsverteilung:*

Gemäss dieser Konzeption hat eine Person eine variierende innere Tendenz (Propensität) die unterschiedlichen möglichen Testergebnisse auf einen Test zu produzieren. Diese innere Tendenz kann durch eine Wahrscheinlichkeitsverteilung über die verschiedenen Testwerte repräsentiert werden. Der True-Score ist der Mittelwert dieser Verteilung (Und die Fehlervarianz entspricht der Varianz dieser Verteilung).
Diese Interpretation erinnert stark an Poppers (1959) *Propensitätstheorie der Wahrscheinlichkeit.* Gemäss dieser Konzeption besitzt ein physikalisches System eine bestimmte Propensität, die verschiedenen möglichen Ergebnisse zu produzieren. So besitzt z.B. ein faires Münzwurf-Szenario die Tendenz, in 50% der Fälle »Kopf« und in 50% »Zahl« zu produzieren. Diese Propensität kann durch eine Wahrscheinlichkeitsverteilung repräsentiert werden.
Ob Lord und Novick (1968) tatsächlich Poppers Propensitätstheorie im Hinterkopf hatten, ist allerdings fraglich, da sie permanent die Propensitätskonzeption mit der frequentistischen Konzeption der Wahrscheinlichkeit vermengen. Poppers Ziel bestand hingegen darin, eine neue Konzeption von Wahrscheinlichkeit zu formulieren, welche nicht auf Frequenzen beruht und welche er als für die Quantenphysik adäquater betrachtete.
Die Interpretation von True-Score als Erwartungswert einer personeninternen Propensitätsverteilung ist – im Gegensatz zur operationalistischen Sichtweise – eine Latente-Variablen-Konzeption, da es sich bei der Propensitätsverteilung um ein theoretisches Konstrukt handelt.

3. *True-Score als Amalgam von latenten Konstrukten und Relationen zwischen latenten Konstrukten und Messwerten:*

Im letzten Kapitel ihres Buches behandeln Lord und Novick (1968) latente Variablenmodelle. In diesem Kontext betrachten sie True-Scores als eine Zusammensetzung von latenten Variablen und deren Relationen zu beobachteten Grössen.
Auch diese Konzeption von True-Score kann als eine Latente-Variablen-Interpretation aufgefasst werden.

Die Darstellung von Lord und Novick (1968) lässt daher keine eindeutige inhaltliche Interpretation des Konzepts des True-Score zu. Es gibt jedoch eine Gemeinsamkeit: Unter allen drei Interpretationen entspricht der Wert des True-Scores dem erwarteten Wert des Tests innerhalb der Population der getesteten Personen.
Aufgrund dieser Schwierigkeiten bei der Interpretation von True-Score werden wir im Folgenden vollständig auf dieses Konzept zugunsten der Latenten-Variablen-Konzeption verzichten. Die klassischen Testmodelle und deren Erweitungen lassen sich im Rahmen dieser Konzeption rekonstruieren und verstehen (siehe Kapitel 2).

Nach diesen Ausführungen zur Bedeutung theoretischen Annahmen für die Messung und die Konkretisierung dieser Annahmen im Rahmen des Latenten-Variablen-Ansatzes, werden im folgenden Abschnitt mögliche problematische Aspekte dieser Konzeption diskutiert.

1.2 Mögliche Kritikpunkte des Latenten-Variablen-Ansatzes

Der vorliegende Latente-Variablen-Ansatz weist trotz seiner unbestreitbaren Vorteile auch problematische Aspekte auf. Im Folgenden werden drei mögliche Kritikpunkte des Ansatzes kurz diskutiert:

1. Kritik der Verwendung theoretischer Konstrukte.
2. Kritik des wissenschaftlichen Realismus.
3. Kritik der kausalen Interpretation von Relation zwischen Variablen.

1.2.1 Das Problem der Verwendung theoretischer Konstrukte

Theoretische Konstrukte beziehen sich auf Dinge, welche nicht direkt beobachtbar sind. Daher stellt sich die Frage, ob die Verwendung derartiger Konstrukte überhaupt sinnvoll ist. Ein gewichtiger Einwand gegen die Verwendung theoretischer Konstrukte ergibt sich aus dem folgenden fundamentalen Prinzip der Rationalität:

Prinzip 1-3: *Ockhams Rasiermesser (Occam's Razor):*

»Entia non sunt multiplicanda praeter (sine) necessitatem. (Entitäten dürfen nicht über das Notwendige hinaus vermehrt werden)«. [William von Ockham (vermutlich 1285-1347)]

Das Prinzip besagt, dass die Ontologie von Theorien nicht mit Entitäten belastet werden sollte, welche keinen Beitrag zur Erklärung liefern.

Wenn zwei gleich gute Erklärungen vorliegen, von denen aber eine mit weniger Vorannahmen auskommt, so ist diese zu bevorzugen.

Dieses Prinzip ist von eminenter Bedeutung für die Wissenschaften im Allgemeinen und für die Statistik im Speziellen. Es verhindert die Entstehung von unnötig komplexen Modellen.
Das Occamsche Rasiermesser verbietet jedoch nicht die Verwendung theoretischer Konstrukte an sich. Es richtet sich nur gegen den *ungerechtfertigten* Gebrauch derartiger Annahmen ein. Die Frage ist daher, ob die Annahme latenter theoretischer Konstrukte innerhalb wissenschaftlicher Theorien im Allgemeinen bzw. die Annahme mentaler Konstrukte im Rahmen psychologischer Theorien gerechtfertigt ist.
Es gilt heute als unbestreitbar, dass fortgeschrittene Wissenschaften nicht ohne die Verwendung theoretischer Konstrukte auskommen. Auch innerhalb der Psychologie hat sich nach dem Ende des Behaviorismus die Meinung durchgesetzt, dass eine Erklärung menschlichen

Verhaltens ohne den Rückgriff auf mentale Konstrukte nicht möglich ist.

1.2.2 Das Problem des wissenschaftlichen Realismus

Der Latente-Variablen-Ansatz basiert auf der Position des wissenschaftlichen Realismus. Gemäss dieser Auffassung beziehen sich theoretische Terme auf real existierende Dinge. Im aktuellen Fall bedeutet dies, dass mentale Fertigkeiten als existierend angenommen werden. Die Messung besteht darin, Aspekte dieser Fertigkeiten zu messen.
Die Gegenposition zum Realismus – der Antirealismus – umfasst eine Menge unterschiedlicher Auffassungen, wie *Instrumentalismus, konstruktiven Empirizismus* (*constructive empiricism*) [Van Fraassen, 1980] oder *NOA* (*natural ontological attitude*) [Fine, 1984]. Die bekannteste Gegenposition ist der *Instrumentalismus*. Dieser behauptet, dass theoretische Konstrukte lediglich als geistige Konstruktionen zu betrachten sind, welche es ermöglichen, unsere Erfahrungen zu strukturieren.
Der wissenschaftliche Realismus hält einer strengen Prüfung nicht Stand. Dies liegt darin, dass aufgrund des Erfolgs einer Theorie bezüglich der Erklärung und Vorhersage von Ereignissen sich keine zwingenden Folgerungen hinsichtlich der Existenz der in den Theorien postulierten Entitäten ergeben. Denn, einerseits können Theorien, welche nicht-existent Entitäten (zumindest gemäss unserer derzeitigen Erkenntnis) postulieren, erfolgreich sein. Klassische Beispiele hierfür sind die Theorie des Äthers oder die Phlogistontheorie der Verbrennung. Andererseits können Theorien, die korrekte Mechanismen postulieren, zumindest zeitweise wenig erfolgreich sein. Ein Beispiel hierfür ist die Urknall-Theorie der Entwicklung des Universums (vgl. Singh, 2005) oder das Kopernikanische Sonnensystem.

Bemerkung:
Die letztgenannten Beispiele zeigen, dass eine einfache Form des Falsifikationismus, wonach eine Theorie, welche durch die Daten nicht bestätigt wurde, als widerlegt zurückzuweisen ist, kontraproduktiv sein kann.
Es muss jedoch angemerkt werden, dass diese Art von Falsifikationismus niemals von Popper vertreten wurde (Watkins, 1984).

Trotz der Tatsache, dass der Realismus nicht durch den Erfolg einer Theorie gerechtfertigt werden kann, werden wir im Folgenden die realistische Position beibehalten. Der Grund liegt darin, dass diese Position heute in der Psychologie und vermutlich in den meisten Wissenschaften als die dominante Sichtweise betrachtet werden kann. Weiters entspricht eine realistische Sichtweise auch der Auffassung von Laien, was die Studien zum psychologischen Essentialismus belegen (Medin & Ortony, 1989).

Bemerkung zum psychologischen Essentialismus:
Der psychologische Essentialismus lässt sich wie folgt zusammenfassen:

> *Wissenschaftliche Laien und auch Angehörige von Naturvölkern glauben, dass Dinge einen inneren unabänderlichen Kern besitzen (eine zentrale innere Eigenschaft), welcher das Wesen des Dinges ausmacht.*
> *Dieser Glaube ist auch dann unerschüttert, wenn der Person vorgeführt wird, dass sie nicht imstande ist, eine definierende Eigenschaft anzugeben.*

1.2.3 Das Problem der kausalen Interpretation

Eng verbunden mit dem wissenschaftlichen Realismus ist die Annahme der Existenz kausaler Wirkungen. Die Latente-Variablen-Konzeption der Messung nimmt an, dass zwischen den latenten Variablen kausale Wirkungen bestehen und dass die latenten Konstrukte eine Wirkung auf die zu messenden Variablen ausüben. Dieser kausale Einfluss von mentalen Entitäten auf die Indikatoren wird als eine notwendige Bedingung dafür betrachtet, dass eine Messung des Konstrukts überhaupt möglich ist.
Die letztgenannte Annahme ist umstritten. So kritisiert Zumbo (2007) die Idee von Borsboom, Mellenbergh und Van Heerden (2004), Validität auf die kausale Beziehung zwischen dem gemessenen Konstrukt und der Messung zurückzuführen. Er führt hierbei an, dass das Konzept der Kausalität selbst nicht vollständig klar und auch zu eingeschränkt sei:

> *»I am not found of the exclusive reliance on "causal" model of explanation of the sort that Borsboom and his colleagues suggest. Their causal notions give us a restricted view of measurement because of the well-known objections to the causal model of explanation – briefly, that we do not have a fully adequate analysis of causation, there are non-causal explanations, and that it is too weak of permissive, that it undermines our explanatory practices.« (Zumbo, 2007; S. 53).*

Das Konzept der Validität wird im Abschnitt 2.4 ausführlich behandelt. In diesem Zusammenhang wird auch die Konzeption von Borsboom et al. (2004) dargestellt. Zur Kritik von Zumbo (2007) seien hier nur zwei Anmerkungen gemacht:

1. Das Konzept der Kausalität ist zentral für alle Naturwissenschaften. (Kitcher & Salmon, 1989). Die Behauptung, wonach auf dieses Konzept, da nicht völlig geklärt, zu verzichten sei, ist kontraproduktiv: Mit dem gleichen Argument kann man andere Konzepte, wie jenes der *Gesetzesaussage* oder das der *Wahrscheinlichkeit* aus den Wissenschaften ausschliessen.
2. Der Behauptung von Zumbo, wonach es neben kausalen Erklärungen noch andere relevante Arten von wissenschaftlichen Erklärungen gibt, welche nicht auf Kausalerklärungen reduzierbar sind, –

wie z.B. funktionale, historische oder hermeneutische, – entbehrt jeder Grundlage.

Bemerkung zu nichtkausale Erklärungen:
Mir ist nur eine Erklärung bekannt, welche – zumindest zum aktuellen Zeitpunkt – nicht auf eine Kausalerklärung reduzierbar ist.
Diese betrifft die Vorhersage von seltenen Metallen mit Hilfe des Pauli-Verbots (im Rahmen des Bohrschen Atommodells).
Das Pauli-Verbot besagt – vereinfacht ausgedrückt – dass es in einem Quantensystem keine zwei Elemente mit völlig übereinstimmenden Quantenzuständen geben kann. Hier handelt es sich um kein kausales Prinzip.
Das Pauli-Verbot sagt die Existenz von chemischen Elementen vorher, in denen das Auffüllen der Schalen des Atoms mit Elektronen nicht dem Prinzip der geringsten Energie entspricht (eine klare und gut lesbare Diskussion hierzu findet man in Pais, 1993).

Die Kritik von Zumbo (2007) ist aufgrund der angeführten Argumente kaum gerechtfertigt. Die Verwendung einer kausalen Interpretation vereinfacht darüber hinaus das Verständnis für zentrale Konzepte der Testtheorie, wie z.B. *Validität* und *Reliabilität*.

Bevor wir die Diskussion zur kausalen Interpretation beenden, möchte ich noch einer Behauptung von Borsboom, Mellenberg und Van Heerden (2003) entgegentreten, der zufolge eine konstante Grösse keine Ursache sein kann. Im Speziellen behaupten die Autoren:

> *»On cannot say, however, that Subject A's latent variable value caused his item response, because there is no covariation between his position on the latent variable and his item response. An individual's position on the latent variable is, in a standard measurement model, conceptualized as a constant, and a constant cannot be a cause.« (Borsboom et. al., 2003; S. 211).*

Diese Position führt zu absurden Schlussfolgerungen. So dürfte man demzufolge – um ein Beispiel von Glymour (1986, Seite 966) zu zitieren – nicht behaupten, dass der Urknall die Ursache für die Hintergrundstrahlung sei, da der Urknall ein einmaliges Ereignis ist.

Meines Erachtens verwechseln Borsboom et al. (2003) das ontologische Problem der Bedingungen für das Vorliegens einer Ursache mit dem epistemologischen Problem der Voraussetzungen für das Erkennen einer Ursache. Für letzteres benötigt man oft Beobachtungen über das gemeinsame Auftreten von Ursache und Wirkung (oder aber – wie im Falle des Urknalls – eine gut bestätigte Theorie, welche die kausale Relation impliziert).

Nach dieser Diskussion der Bedeutung theoretischer Annahmen für die Fundierung der Latenten-Variablen-Konzeption des Messens mentaler

Konstrukte, wenden wir uns nun dem zentralen Konzept des *Messmodells* zu.

1.3 Messmodelle

Das Ziel dieses Abschnittes besteht darin, verschiedene Aspekte des Konzepts des Messmodells anhand eines einfachen Beispiels zu demonstrieren.
Wir betrachten hier ein Messmodell *als eine Repräsentation der Messsituation, welche die kausalen Einflussgrössen, die die Messung beeinflussen, beinhaltet.* Dies bedeutet, dass das Messmodell alle Annahmen über die in der Messung involvierten Entitäten und deren Beziehungen beinhaltet (Eine präzisere Definition wird weiter unten präsentiert, vgl. Konzept 2-1).
Das folgende Beispiel soll dazu dienen, das Konzept des Messmodells zu illustrieren.

Bsp.1-1: Ein einfaches Prozessbaummodell zur Messung von Wissen und Raten auf das Lösen binärer Items:

Gegeben:
Ein Menge von Testitems eines Leistungstests mit jeweils zwei Antwortmöglichkeiten, eine davon wahr, die andere falsch.
Zielsetzung:
Messung des Wissens der Person unter Berücksichtigung der Möglichkeit des Ratens einer korrekten Antwort.
Die grundlegenden Annahmen des Messmodells:
Das Messmodell in Abb. 1-1 macht folgende Annahmen:

1. Bei Präsentation eines Testitems ist der Proband mit Wahrscheinlichkeit π in der Lage, das zur Lösung relevante Wissen abzurufen (oberer Zweig in Abb. 1-1). In diesem Fall erfolgt die korrekte Antwort mit Wahrscheinlichkeit 1.
2. Mit Wahrscheinlichkeit $1-\pi$ ist das Wissen nicht abrufbar (entweder weil es nicht vorhanden ist oder weil nicht darauf zugegriffen werden kann). In diesem Fall rät die Person mit Wahrscheinlichkeit von 0.5 die korrekte Antwort.

Der Parameter π:
Das Modell enthält eine unbekannte Grösse π. Hierbei handelt es sich um einen so genannten (Populations-) Parameter (siehe hierzu Konzept 1-1 und Konzept 1-2, sowie Bsp. 1-2).
Statistische Parameter kennzeichnen so genannte Grundgesamtheiten oder Populationen. *Genauer:* Parameter kennzeichnen *die Verteilung* von Merkmalen einer Population (Konzept 1-2). Sie werden daher auch *Kennwerte* genannt.

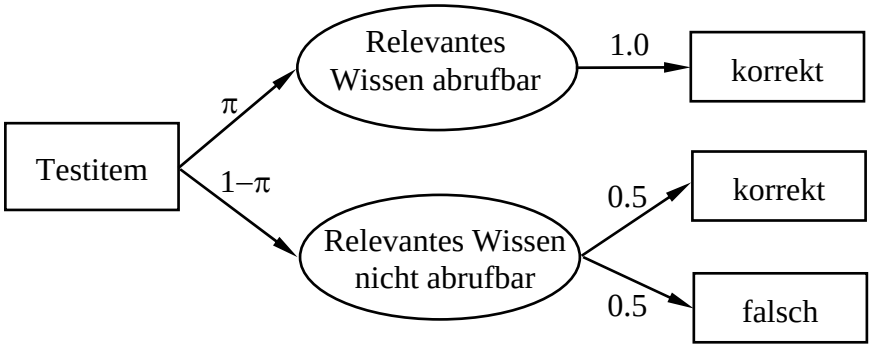


Abb. 1-1: *Prozessbaum-Modell zur Messung von relevantem Wissen und Raten.*

Alle im Folgenden zu besprechenden Messmodelle enthalten Parameter. *Diese Parameter sind die zentralen Grössen eines Messmodells, denn sie repräsentieren im Wesentlichen die Messung.*

Im aktuellen Beispiel repräsentiert π die Wahrscheinlichkeit, dass eine Zielperson bei Vorliegen eines Testitems das relevante Wissen zur Lösung abrufen kann.

Die betrachtete Population besteht daher aus einem Pool von Testitems, die alle (annähernd) den gleichen Schwierigkeitsgrad aufweisen.

Beispielsweise könnte der Itempool aus eine Menge von multiple-choice Fragen mit je zwei Wahlmöglichkeiten zu einem Fachgebiet bestehen. Aus diesem Pool von Fragen, wird zufällig eine Stichprobe gewählt und dem Probanden vorgelegt.

Wie oben erwähnt, repräsentieren die Parameter den zentralen Aspekt einer Messung. Für das aktuelle Beispiel ist dies leicht nachzuvollziehen, denn der Prüfer ist daran interessiert, wie viel der Proband weiss. Der Anteil korrekter Antworten ist offensichtlich kein ideales Maß für das Wissen, da in dieser Gösse die Anteile von Wissen und Raten konfundiert sind.

In diesem Zusammenhang spricht man auch von *prozessreinen Maßen* (*process pure measures*). Hierbei handelt es sich um Maße welche einen Prozess oder ein Konstrukt »gereinigt» von anderen Einflüssen und Prozessen repräsentieren.

So kann man π als prozessreines Maß des Wissens betrachten, falls das Modell korrekt ist. Der Prozentsatz korrekter Antworten repräsentiert Wissen und Raten und ist hingegen kein reines Maß des Wissens der Person.

1.3.1 Syntax und Semantik von Messmodellen

Bei der Betrachtung von Messmodellen lassen sich zwei Aspekte unterscheiden:

1. *Syntaktischer Aspekt:* Dieser betrifft die formale mathematische Struktur des Modells. In der Regel handelt es sich hierbei um ein System von Gleichungen.
   Das Ziel von Messmodellen besteht darin, die Verteilung der Messungen zu modellieren. Hierbei wird davon ausgegangen, dass die Messungen einem bestimmten Typ von Verteilungen folgen. Die bei weitem wichtigsten dieser Verteilungen sind die *multivariate Normalverteilung* und die *Multinomialverteilung*. Diese Verteilungen besitzen *Kennwerte*, auch *Parameter* genannt.

Konzept 1-1: *Variable, Parameter und Konstante:*

Ein (*Funktions*-) *Parameter* ist eine Grösse, die hinsichtlich ihrer Variabilität zwischen einer *Variablen* und einer *Konstanten* angesiedelt ist, d.h. ein Parameter ändert sich weniger oft als eine Variable. Er ist jedoch auch keine Konstante, deren Wert sich niemals ändert.

Innerhalb der Statistik spielen Parameter eine zentrale Rolle:

Konzept 1-2: *Populationsparameter und Stichprobenkennwert (Statistik):*

Ein *Populationsparameter* ist ein Kennwert, der zur Charakterisierung von Verteilungseigenschaften innerhalb einer Population verwendet wird.
Ein *Stichprobenkennwert* (*Statistik*) ist eine Funktion der Stichprobe.

Bsp. 1-2: Populationsparameter und Stichprobennkennwerte:

Klassische Beispiele für Populationsparameter sind:

1. Der Mittelwert μ der Grösse der erstsemestrigen Studentinnen von Fribourg,
2. Die Streuung σ um den besagten Mittelwert,
3. Der Regressionskoeffizient β der Regression des Gewichts auf die Grösse, innerhalb der besagten Population der erstsemestrigen Studentinnen,
4. Der Korrelationskoeffizient ρ der Korrelation zwischen Gewicht und Grösse, innerhalb der besagten Population der erstsemestrigen Studentinnen.

Die mit den Populationsparametern korrespondierenden Statistiken sind:

1. Der Mittelwert der Stichprobe: $\bar{x} = \frac{1}{N} \cdot \sum_{i=1}^{n} x_i$, (1-1)
2. Die Streuung innerhalb der Stichprobe:

$$s = \sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \,, \quad (1\text{-}2)$$

3. Der aus der Stichprobe ermittelte Regressionskoeffizient:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \,, \quad (1\text{-}3)$$

4. Der aus der Stichprobe ermittelte Korrelationskoeffizient:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (1\text{-}4)$$

Beachte:

Populationsparameter und Statistiken sind strikt zu unterscheiden. Populationsparameter sind – bei gegebener Population – fixe Grössen (zumindest innerhalb der klassischen Statistik), während Statistiken Zufallsvariablen sind: Zieht man eine neue Stichprobe, so ergeben sich (wahrscheinlich) andere Stichprobenkennwerte.

Notationskonvention:

Die Unterscheidung zwischen Populationsparametern und Stichprobenkennwerte wird durch die Verwendung einer bestimmten Notationskonvention hervorgehoben (Bsp. 1-2):

Populationsparameter werden mit griechischen Buchstaben bezeichnet: μ, σ, β, ρ, etc.

Die Modellgleichungen dienen dazu, die Parameter der Verteilung zu modellieren. Dies bedeutet, die Originalparameter der Verteilung werden als Funktion von neuen Parametern repräsentiert. Die Anzahl dieser neuen Parameter ist gewöhnlich geringer als die Anzahl der zu modellierenden Parameter (Vgl. Abschnitt 1.3.2).

2. *Semantischer Aspekt:* Dieser betrifft die inhaltliche Interpretation der Variablen und Parameter des Modells, sowie der Relationen zwischen den Variablen (Da die Relationen zwischen den Variablen des Modells oft durch die Parameter repräsentiert werden, fällt die Interpretation von Parametern und Relationen zwischen Variablen zusammen).

Bsp.1-3: Syntax und Semantik des Prozessbaummodells zur Messung des Einflusses von Wissen und Raten (Fortsetzung von Bsp.1-1):

Gegeben:
Das Modell von Abb. 1-1.
Formales syntaktisches Modell:
Das Modell besteht aus einem Gleichungssystem mit den beiden Gleichungen:

$$\begin{aligned} P(\text{korrekt}) &= \pi + 0.5 \cdot (1-\pi) \\ P(\text{falsch}) &= 0.5 \cdot (1-\pi) \end{aligned} \qquad (1\text{-}5)$$

Hierbei ist $P(\text{korrekt})$ die Wahrscheinlichkeit einer korrekter und $P(\text{falsch})$ die Wahrscheinlichkeit einer falscher Antwort.
Das Modell repräsentiert (oder modelliert) also die Wahrscheinlichkeit einer korrekten bzw. falschen Antworten als Funktion des Parameters π.
Die Grösse $\theta = P(\text{korrekt})$ ist aber selbst ein Parameter einer Verteilung, nämlich der Wahrscheinlichkeitsparameter der Binomialverteilung:

$$P(\text{Anzahl korrekter Antworten} = x) = \binom{n}{x} \cdot \theta^x \cdot (1-\theta)^{n-x} \qquad (1\text{-}6)$$

Hierbei ist $\binom{n}{x} = \frac{n!}{x! \cdot (n-x)!}$ der Binomialkoeffizient,

n ist die Anzahl präsentierter Testitems,
x ist die Anzahl korrekter Antworten $(x = 0,1,\ldots,n)$.
Die Binomialverteilung spezifiziert also für jede mögliche Anzahl korrekter Antworten *x* deren Wahrscheinlichkeit, wobei der Parameter θ (= Wahrscheinlichkeit, dass ein Item korrekt beantwortet wird) vorgegeben ist.
Gewöhnlich ist die Anzahl *x* der korrekten Antworten gegeben und man möchte den Parameter θ ermitteln. Im aktuellen Fall ist dies sehr einfach, denn es existiert ein einfacher und guter Schätzer von θ, der *Maximum-Likelihood Schätzer* des Parameters θ:

$$\hat{\theta} = \frac{x}{n}. \qquad (1\text{-}7)$$

d.h. $\theta = P(\text{korrekt})$ wird durch den Anteil korrekter Antworten geschätzt.

Konzept 1-3: *Schätzer vs. Schätzung eines Parameters:*

Der Wert eines Parameters wird mit Hilfe von Daten geschätzt. Hierbei gilt:

Der *Schätzer des Parameters* ist die Formel bzw. Gleichung, aufgrund dessen sich der Parameter aus den Daten ermitteln lässt. Diese Gleichung muss nicht eine explizite Formel sein, wie in Gleichung (1-7). In vielen Fällen ist der Schätzer nur implizit durch ein System von Gleichungen gegeben, welches nur mittels numerischer Methoden gelöst werden kann.

Die *Schätzung eines Parameters* entspricht dem konkreten Wert, der sich aufgrund der aktuell vorliegenden Daten ergibt.

Notationskonvention:
Der Schätzer eines Parameters θ wird durch ein »Dach« symbolisiert: $\hat{\theta}$, $\hat{\mu}$, $\hat{\beta}$, $\hat{\rho}$, etc..

Der griechische Buchstabe θ (theta) wird meist verwendet, um einen nicht näher bestimmten Parameter zu symbolisieren (In Bsp.1-3 wurde er allerdings nur deshalb verwendet, weil das natürlichere Symbol π bereits »besetzt« ist).

Aufgrund der Schätzung von $\theta = P(\text{korrekt})$ und den Modellgleichungen von (1-5) kann man auch den Wert des Modellparameters π ermitteln (schätzen). Es gilt:

$$\hat{\pi} = 2 \cdot \frac{x}{n} - 1 \qquad (1\text{-}8)$$

Gleichung (1-8) ist die klassische Formel zur Ratekorrektur, wie man sie in vielem Lehrbüchern findet (siehe z.B. Lord & Novick, 1968; Macmillan & Creelman, 2005).
Inhaltliche Interpretation des Modells:
Die inhaltliche Interpretation des Modells wurde bereits oben präsentiert: So wird π als die Wahrscheinlichkeit des Abrufs des relevanten Wissens interpretiert und der Wert 0.5 beschreibt die Wahrscheinlichkeit des korrekten Ratens, falls kein relevantes Wissen abgerufen werden kann.

Bemerkung / Entwarnung:
Wir werden uns im Folgenden nicht mit dem Problem der Schätzung von Parameter befassen. Hierzu verwenden wir Programme, welche diese Schätzungen für uns durchführen.
Im Zentrum stehen für uns verschiedene Testmodelle, ihre Bedeutung, Vorannahmen und Anwendungen.

1.3.2 Prüfung von Messmodellen

Messmodelle implementieren die theoretischen Annahmen über den Messprozess. Klarerweise möchte man nun prüfen, ob das Modell korrekt ist, d.h. ob die Annahmen, die in die Konstruktion des Messmodells eingingen, korrekt sind. Eine notwendige (aber nicht hin-

reichende) Voraussetzung für die Prüfung eines Messmodells besteht im Vorliegen von Relationen zwischen den theoretischen Grössen im Modell und beobachtbaren oder anderen bekannten Grössen. Nur wenn im Modell derartige Beziehungen realisiert sind, kann es Vorhersagen liefern und ist somit potentiell testbar.

Bemerkung:
Bei den *anderen bekannten Grössen*, die nicht direkt messbar bzw. beobachtbar sind, könnte es sich z.B. um theoretische Vorhersagen von gut bestätigten Modellen handeln. In diesem Fall wird zur Prüfung des Modells ein anderes theoretisches Modell benötigt.

Im Modell von Abb. 1-1 sind die Beziehungen zwischen den theoretischen und den beobachtbaren Grössen durch die Pfeile symbolisiert. Diese besagen, dass die postulierten inneren Zustände zu bestimmten beobachtbaren Verhalten führen. In den Modellgleichungen (1-5) ist die Beziehung realisiert, indem die beobachteten Grössen als Funktion der Modellparameter dargestellt sind.
Die Existenz einer Relation zwischen beobachteten und theoretischen Grössen ist jedoch nicht hinreichend für die Prüfbarkeit eines Modells. So ist z.B. das Modell in Abb. 1-1 nicht prüfbar. Der Grund liegt darin, dass das Modell einen freien Parameter (π) besitzt, mit dem es einen freien Datenpunkt (die Wahrscheinlichkeit einer korrekten Antwort) erklärt. Man beachte, dass die Wahrscheinlichkeit einer falschen Antwort keinen zusätzlicher freien Datenpunkt darstellt, da sich diese direkt aus der Wahrscheinlichkeit einer korrekten Antwort ergibt.
Erweitert man das Modell von Abb. 1-1 derart, dass es als Messmodell für Items mit $n>2$ möglichen Antwortoptionen fungieren kann (Übung 1-1), so ist das resultierende Modell prüfbar. Denn in diesem Falle gibt es $n-1$ freie Datenpunkte und nur einen freien Parameter zur Erklärung derselben. Das Modell stellt also eine »Ersparnis« gegenüber den Daten dar, d.h. es erklärt die Daten unter Verwendung einer geringeren Anzahl frei schätzbarer Grössen.

1.3.3 Komplexität und Fehlerhaftigkeit von Messmodellen

Für alle Modelle (insbesondere für jene in der Psychologie) gilt das folgende Prinzip:

Prinzip 1-4: *Fehlerhaftigkeit von (Mess-) Modellen:*

Jedes (Mess-) Modell ist (mehr oder weniger) falsch!

Was das Modell von Abb. 1-1 betrifft, so kann man davon ausgehen, dass das Modell mit Sicherheit falsch ist. Der Grund hierfür liegt in der Tatsache begründet, dass die Dichotomie zwischen Wissen und (reinem) Raten in dieser einfachen Form nur extrem selten auftritt. Vielmehr muss das Wissen in den meisten Fällen als graduell betrach-

tet werden. Es handelt sich in diesem Falle um kein reines Raten mehr, sondern eher um ein wissensbasiertes Raten.
Trotz der Fehlerhaftigkeit von Modellen können sie für bestimmte Zwecke mehr oder weniger nützlich sein. So ist das Modell der Ratekorrektur sicher nützlich, falls die Fragen schwierig und keine »Fangoptionen« vorhanden sind. Denn in diesem Falle spiegelt das Modell die Messsituation relativ gut wider. Dies hat folgenden Grund: Wenn die Fragen sehr schwierig sind, so muss das Wissen eine bestimmte Schwelle überschreiten, damit die Frage korrekt beantwortet werden kann, andernfalls ist die Antwort tatsächlich ungefähr auf Zufallsniveau, solange keine »Fangoptionen« vorliegen, welche die Antworten der ratenden Personen in eine falsche Richtung lenken.

Bemerkung:
Das Ratemodell wird auch als *Hochschwellen-Modell* bezeichnet, weil es annimmt, dass eine hohe Schwelle überschritten werden muss, damit die Antwort mit Sicherheit korrekt wiedergegeben werden kann. Wird die hohe Schwelle nicht überschritten, kann die Person nur raten.
Die eben geschilderte Situation entspricht in etwa diesen Annahmen des Modells.

Bsp.1-4: Fehlspezifikation des linearen multiplen Regressionsmodells mit normal verteilten Fehlern

Gegeben:
Multiples lineares Regressionsmodell:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k + \varepsilon \qquad (1\text{-}9)$$

wobei gilt:

Y	ist die abhängige Variable,
$X_1, X_2, \cdots, X_k$	sind die unabhängigen Variablen, von denen angenommen wird, dass sie ohne Fehler gemessen werden,
$\beta_1, \beta_2, \cdots, \beta_k$	sind die Regressionskoeffizienten (auch β–Gewichte genannt),
ε	ist der Residualterm, von dem angenommen wird, dass er unabhängig normal verteilt ist, mit Mittelwert 0 und Varianz σ^2.

Bemerkungen:

1. Das Modell von Gleichung (1-9) repräsentiert ein System von *n*) *n* = Anzahl der Einheiten) linearen Gleichungen, wobei für jede Einheit *i* gilt:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \cdots + \beta_k \cdot x_{ik} + \varepsilon_i \qquad (i = 1, \ldots, n)$$

wobei gilt:

y_i	ist der Wert der abhängige Variable *Y* für die Einheit *i*.
$x_{i1}, x_{i2}, \cdots, x_{ik}$	sind die Werte von Einheit *i* auf den unabhängigen Variablen $X_1, X_2, \cdots, X_k$
ε_i	repräsentiert den Wert des Residuums für Einheit *i*.

2. Das Wort »unabhängig« in der Wendung »unabhängig normal verteilt« bedeutet, dass die Fehlerterme der einzelnen Einheiten unabhängig voneinander und auch unabhängig von den unabhängigen Variablen $X_1, X_2, \cdots, X_k$ verteilt sind.

Das lineare Regressionsmodell ist in den meisten sozialwissenschaftlichen Anwendungen fehlerhaft, d.h. es spiegelt die Situation nicht korrekt wider. Dies hat zwei Gründe:

1. Die Annahme der fehlerfreien Messung der unabhängigen Variablen ist meist nicht korrekt (ausgenommen es handelt sich um experimentelle Variablen, die vom Versuchsleiter kontrolliert werden).
2. Die Annahme, dass die Residuen unabhängig von den unabhängigen Variabeln sind, ist nicht realistisch (ausgenommen es handelt sich experimentell kontrollierte Variablen). Dies würde nämlich bedeuten, dass alle Variablen, die mit den unabhängigen Variablen korreliert sind und einen Einfluss auf die abhängige Variable ausüben, als in das Modell einbezogen wurden (als unabhängige Variablen).

Trotz der Tatsache, dass das lineare Regressionsmodell die Situation nicht völlig korrekt abbildet, ist es in vielen Anwendungen von Nutzen, wenn folgende Bedingungen erfüllt sind:

1. Die Messfehler der unabhängigen Variablen sind im Verhältnis zu deren systematischen Variation gering.

2. Die in das Modell einbezogenen unabhängigen Variablen erklären den Grossteil der Varianz von *Y*, sodass man davon ausgehen kann, dass alle relevanten (unabhängigen) Variablen in das Modell einbezogen wurden. Im Speziellen, sollte es nicht der Fall sein, dass die Einbeziehung weiterer Variablen die Werte der Regressionskoeffizienten drastisch verändert.

Die bisherige Diskussion lässt sich wie folgt zusammenfassen: Obwohl alle (Mess-) Modelle fehlerbehaftet sind – d.h. sie spiegeln die Situation nicht exakt wider – können sie dennoch nützlich sein, falls die Abweichungen von der Realität nicht so drastisch sind, dass sich fehlerhafte inhaltliche Schlussfolgerungen ergeben (Letzteres wäre zum Beispiel der Fall, wenn die Einbeziehung weiterer unabhängiger Variablen in das lineare Regressionsmodell die Vorzeichen eines oder mehrerer Regressionskoeffizienten ändern würde).

Um zu erreichen dass ein Modell die Realität besser wiedergibt, muss man oft dessen Komplexität erhöhen. Zum Beispiel kann man im linearen Regressionsmodell (siehe Bsp.1-4) weitere unabhängige Variablen einbeziehen, um eine bessere Erklärung der Varianz von *Y* zu erreichen. Mit der Erhöhung der Komplexität bzw. mit der Bildung komplexer Modelle sind zwei mögliche Gefahren verbunden:

1. *Die Spezifikation unnötig komplexer Modelle:*
   Ein typischer Anfängerfehler besteht in der Konstruktion zu komplexer Modelle. Einfache Modelle haben eine Reihe von Vorzügen gegenüber komplexeren. Sie sind einfacher zu verstehen, einfacher zu schätzen und meist robuster (d.h. kleine Änderungen in den Daten führen nicht zu grossen Veränderungen in den Modellparametern). Zusätzlich entsprechen sie eher dem Occamschen Rasiermesser (vgl. Prinzip 1-3).
   Es daher sinnvoll, mit einfachen Modellen zu beginnen. »Komplizierter kann man immer werden«. Weiter ist es wichtig, Maße der Modellgüte zu verwenden, welche die Modellkomplexität stärker berücksichtigen (vgl. Kapitel **xxxx**). Simluationsstudien belegen, dass Maße der Modellgüte, welche die Modellkomplexität nicht ausreichend berücksichtigen (und dies ist für die meisten verwendeten Maße der Fall), zu komplexe Modell in ungerechtfertigter Weise bevorzugen (siehe z.B. Camstra, & Boomsma, 1992).
2. *Das Problem der Modellierung des Zufalls:*
   In vielen Fällen wird ein vorgegebenes Messmodell durch die Daten der Stichprobe nicht bestätigt. Das übliche Vorgehen besteht nun darin, das Modell zu verändern (meist in Richtung höherer Komplexität).
   Gegen dieses Verfahren ist grundsätzlich nichts einzuwenden. Es sollte dabei jedoch immer im Auge behalten werden, dass die

Daten durch den Zufall mitbestimmt wird. Dies hat zur Folge, dass – vor allem bei komplexen Daten – manche signifikanten Effekte zufallsbedingt sind. Modifiziert man nun das Modell derart, dass es gut zu den vorliegenden Daten passt, so besteht die Gefahr, dass man Zufallseinflüsse mit modelliert (im Englischen spricht man hier von »capitalization on chance«). Da man den Zufall nicht modellieren kann (da er keine systematischen Komponenten enthält), kann dies dazu führen, dass das Modell zwar die Daten der aktuellen Stichprobe gut erklärt aber bei neuen Stichproben kläglich versagt. Es ist daher notwendig, das modifizierte Modell an neuen Daten zu testen, bzw. eine so genannte Kreuzvalidierung durchzuführen (vgl. Kapitel **xxxx**).

Bemerkung:

Wenn man ein Modell aufgrund einer vorgegebenen Stichprobe modifiziert ohne das Modell an neuen Daten zu testen, so muss dies in einer wissenschaftlichen Publikation explizit vermerkt werden.
Andernfalls liegt eine unerlaubte Täuschung vor, welche grossen Schaden anrichten kann, da dann andere Forscher in ungerechtfertigter Weise auf die Korrektheit (und Robustheit) des Modells vertrauen.

Nach diesen allgemeinen Ausführungen zur Messung latenter Konstrukte wenden wir uns nun der Diskussion der klassischen Messmodelle zu.

1.4 Übungen zu Kapitel 1

***Übung 1-1**: Modell zur Ratekorrektur für multiple-choice Items mit n Optionen:*

1. Erstelle – analog zu Modell in Abb. 1-1 – ein Modell zur Messung von Wissens- und Rateprozessen für Items mit n Antwortmöglichkeiten (anstelle von $n = 2$).
   Gehe von der Annahme aus, dass im Falle des Ratens jede Antwortalternative mit gleicher Wahrscheinlichkeit gewählt wird.
2. Erstelle die Modellgleichungen
3. Ermittle die Formel zur Ratekorrektur für dieses Modell.

2. Konzepte, Prinzipien und Methoden der klassischen Testtheorie

In diesem Abschnitt werden zentrale Konzepte der klassischen Testtheorie im Rahmen des latenten-Variablen-Ansatzes (siehe Kapitel 1) behandelt.

Die klassische Testtheorie beschäftigt sich mit drei Arten von Grössen, die sich in zwei Gruppen einteilen lassen:

1. Erwartungswerte,
2. Varianzen und Kovarianzen.

Die Erwartungswerte betreffen den *Leistungsaspekt*. Der Begriff »Leistungsaspekt« soll hier in einem weiten Sinne verstanden werden. Es ist damit nicht nur die Leistung bei leistungsbezogenen Tests gemeint, sondern – allgemeiner – als der *Grad*, die *Höhe* oder das *Ausmass* einer mentalen Eigenschaft.
Die Varianzen und Kovarianzen repräsentieren *strukturelle Aspekte*, d.h. die Variabilität und das Muster des gemeinsamen Auftretens verschiedener mentaler Eigenschaften.

Aufgrund unserer Annahme, dass die mentalen Eigenschaften einen kausalen Einfluss auf die Testwerte ausüben, gehen wir davon aus, dass sowohl der Grad einer mentalen Eigenschaft als auch die Variabilität und die Beziehungen zwischen verschiedenen mentalen Kapazitäten in den Erwartungswerten und Kovarianzen der Test widergespiegelt werden. Daher geben uns letztere Grössen Aufschluss über Ausmass und Struktur der gemessenen mentalen Eigenschaften und eventuell von anderen Grössen, welche einen kausalen Einfluss auf die Testergebnisse ausüben.

Bemerkung:

Im Folgenden beschäftigen wir uns nur mit den strukturellen Aspekten, d.h. mit den Varianzen und Kovarianzen. Die Behandlung von Mittelwertsstrukturen erfolgt später (siehe Abschnitt **xxxx**)

Der vorliegende Abschnitt ist wie folgt aufgebaut: Zuerst wird das Konzept des Messmodells konkretisiert. Dem folgt die Besprechung der klassischen Testmodelle. Sodann werden die beiden zentralen Konzepte der klassischen Testtheorie – *Reliabilität* und *Validität* – eingehend behandelt.

2.1 Messmodelle als Kausalmodelle

Im dem hier verfolgten Ansatz werden Messmodelle als Kausalmodelle betrachtet.

Konzept 2-1: *Messmodell:*

Ein Messmodell ist ein kausales Modell, welches die Messsituation repräsentiert. Es umfasst folgende Komponenten:

(i) Die Menge der verwendeten Maße (Indikatoren).
(ii) Die Menge aller relevanten Einflussgrössen, welche einen direkten Einfluss auf die Messung aufweisen, sowie deren Beziehung untereinander.
(iii) Eine detaillierte Spezifikation der (kausale) Relation zwischen den Maßen und den Einflussgrössen

Bsp.2-1: Veranschaulichung der Komponenten eines Messmodells:

Abb. 2-1 veranschaulicht die einzelnen Komponenten eines Messmodells:

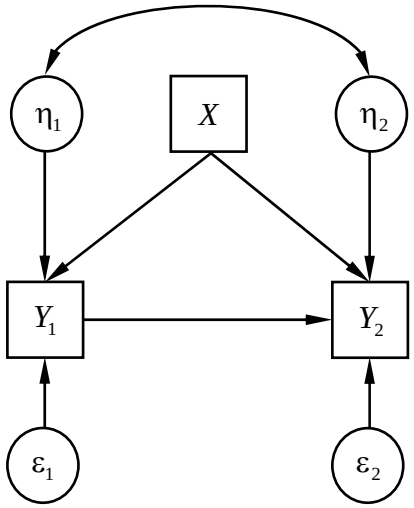


Abb. 2-1: *Die Komponenten eines Messmodells.*

1. Die Maße werden durch die Variablen Y_1 und Y_2. Die Maße sind immer manifeste – d.h. beobachtbare – Variablen. Sie werden – wie alle manifesten Grössen – durch Rechtecke symbolisiert.
2. Die relevanten Einflussgrössen sind die Variablen:

   η_1, η_2, ε_1, ε_2, X und Y_1

   Hierbei werden η_1, η_2, ε_1 und ε_2 als latente (nicht beobachtbare) Grössen betrachtet. Latente Grössen werden immer durch Kreise symbolisiert. Im Speziellen gilt:

   (i) η_1 und η_2 repräsentieren die zu messenden latenten Konstrukte.
   (ii) Die Symbole ε_1 und ε_2 repräsentieren Messfehler.

Die beiden anderen Einflussgrössen, X und Y_1, sind hingegen manifest.

3. Die Pfeile stellen die kausalen Einfüsse dar, wobei die genaue Art und Weise (z.B. linear) sowie die Stärke des Einflusses nicht genau spezifiziert ist.

Das Messmodell in Abb. 2-1 weist eine Reihe spezifischer Merkmale auf, die es zu beachten gilt:

1. Variable X im Modell repräsentiert weder eine Messung, welche modelliert werden soll, noch handelt es sich um eine latenten Grösse, welche durch eine Messung repräsentiert wird. Da X jedoch einen Einfluss auf die Messungen hat, muss sie einbezogen werden. Andernfalls kann es zu verzerrten Schätzungen der eigentlichen Messrelationen (d.h. der Beziehungen zwischen Indikatoren und zugehörigen latenten Variablen) kommen (z.B. kann die Reliabilität der Maße (siehe Konzept 2-6) unterschätzt werden).
2. Der Bogen mit den Doppelpfeilen zwischen den Kreisen, welche die beiden zu messenden Konstrukte symbolisieren, repräsentiert eine kausal nicht näher spezifizierte Relation zwischen den beiden Konstrukten.
3. Das Modell stellt auch die *Abwesenheit* von Relationen zwischen den »reinen« Ursachenvariablen dar. So wird eine Beziehung zwischen den Fehlertermen oder zwischen Fehler und latenten Konstrukten explizit ausgeschlossen. Dies ist in Abb. 2-1 durch die fehlenden Bögen zwischen den Fehlertermen und den latenten Konstrukten indiziert.

   Bemerkung:
   Mit reinen Ursachenvariablen sind jene Variablen gemeint, welche selbst im Modell keine Ursachen aufweisen. Diese Variablen werden auch exogene Variablen genannt (siehe Prinzip 2-1).
4. Im Messmodell wird ein direkter Einfluss des Maßes Y_1 auf Y_2 angenommen. Dieser Fall kann z.B. auftreten, wenn es sich um wiederholte Messungen handelt (siehe hierzu auch Bsp.2-6). Im Allgemeinen sind Einflüsse zwischen Messungen (entweder direkt, wie im dargestellten Fall oder indirekt über Drittvariablen) in das Messmodell einzubeziehen.
5. Die einzelnen Messungen werden als *Effekte* betrachtet, die durch die latenten Konstrukte, durch die zugehörigen Fehlerterme, sowie – eventuell – durch andere beobachtete Grössen kausal beeinflusst werden.

   Wir betrachten im Folgenden nur Messmodelle, in denen die Messungen als Effekte der latenten Konstrukte konzipiert sind und nicht als Ursachen. Derartige Modelle werden als *Modelle mit Effektindikatoren* bezeichnet.

Modelle, in denen die Messungen als Ursachen der latente Konstrukte aufgefasst werden (so genannte *Formative Modelle* oder Modelle mit *Ursachenindikatoren*), werden in Abschnitt **xxxx** kurz behandelt. Diese letztere Art von Modellen werde hier nicht als Messmodelle betrachtet.

Die Definition eines Messmodells in Konzept 2-1 umfasst eine grosse und heterogene Gruppe von Messmodellen. Wir beschränken im Folgenden unsere Überlegungen auf eine spezielle Untergruppe von Messmodellen.

Konzept 2-2: *Lineares Messmodell zur Erklärung der Kovarianzstruktur der Messungen:*

Ein *lineares Messmodell zur Erklärung der Kovarianzstruktur der Messungen* ist ein Messmodell mit folgenden speziellen Eigenschaften:

1. Die kausalen Beziehungen werden als lineare Beziehungen betrachtet.
2. Nicht näher spezifizierte Relationen werden als Kovarianzen interpretiert.
3. Der Zweck des Modells besteht in der Erklärung der Kovarianzstruktur der Messungen.

Das Modell eignet sich in idealer Weise zur Modellierung von multivariat normal verteilten Messungen.

Das folgende Beispiel verdeutlicht das Konzept.

Bsp.2-2: Lineares Messmodell zur Erklärung der Kovarianzstruktur der Messungen:

Abb. 2-2 zeigt nochmals das Messmodel von Abb. 2-1. Im Unterschied zur obigen Darstellung wird das Modell mit Modellparameter gezeigt. Diese Parameter stehen in engem Zusammenhang mit der Interpretation des Modells als lineares Messmodell zur Erklärung der Kovarianzstruktur der Messungen.

Wie sind nun die einzelnen Komponenten des Modells von Abb. 2-2 im Detail zu interpretieren?

Prinzip 2-1: *Interpretation linearer Messmodelle zur Erklärung der Kovarianzstruktur:*

Ein lineares Messmodell zur Erklärung der Kovarianzstruktur umfasst die folgenden Aspekte:

1. *Typen von Variablen:*

   Ein Messmodell umfasst zwei Typen von Variablen:

   (a) *Exogene Variablen:* Hierbei handelt es sich um Variablen, deren Kovarianzstruktur im Modell nicht erklärt wird.

Die exogenen Variablen lassen sich wiederum in drei Gruppen unterteilen:

(i) Die zu messenden latenten Konstrukte;
(ii) Kovariate, welche neben den zu messenden Konstrukten einen Einfluss auf die Messungen ausüben;
(iii) Die Fehlerterme.

Im kausalen Graphen sind exogene Variablen dadurch gekennzeichnet, dass kein Pfeil auf diese Variablen hinführt.
Die exogenen Variablen in Abb. 2-2 sind:

(i) η_1 und η_2: Die latenten Konstrukte;
(ii) *X*: Eine Kovariate;
(iii) ε_1 und ε_2: Die Fehlerterme.

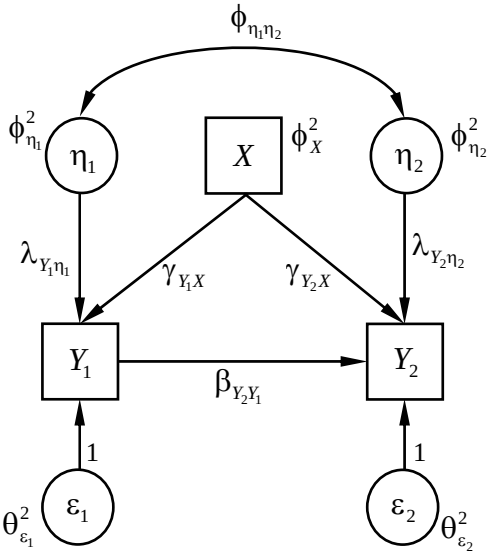


Abb. 2-2: *Das Messmodell von Abb. 2-1 als lineares Messmodell mit Parametern.*

(b) Endogene Variablen: Hierbei handelt es sich um Variablen, deren Kovarianzstruktur mit Hilfe des Modells erklärt werden soll.
Im kausalen Graphen erkennt man endogene Variablen daran, dass ein Pfeil auf diese Variablen hinführt.
Die endogenen Variablen in Abb. 2-2 sind die beiden Messungen Y_1 und Y_2.

2. *Typen von Relationen:*
Ein Messmodell umfasst zwei Typen von Relationen:
(a) *Direkte Kausalrelationen:* Diese sind im Diagramm durch Pfeile symbolisiert. Die Stärke dieser Relationen wird mit Hilfe von Parametern repräsentiert.

Die direkten Kausalrelationen des Modells lassen sich in vier verschiedene Gruppen unterteilen:

(i) Direkte kausale Einflüsse von latenten Konstrukten auf die Messung. Die Parameter zur Repräsentation der Stärke dieser Relationen werden durch den griechischen Buchstaben λ (lambda) symbolisiert.

(ii) Direkte kausale Einflüsse von Kovariaten auf die Messungen. Deren Stärke wird durch den griechischen Buchstaben γ (gamma) symbolisiert.

(iii) Direkte kausale Einflüsse zwischen endogenen Variablen. Ihre Stärke ist durch den griechischen Buchstaben β (beta) repräsentiert.

(iv) Direkte kausale Einflüsse der Fehler, d.h. aller ungemessenen Grössen, welche einen Einfluss auf die Messung ausüben. Deren Stärke wird auf den Wert 1.0 fixiert.

Bemerkung: Direkte, indirekte und totale Effekte:

Die Pfadkoeffizienten repräsentieren die Stärke der *direkten Effekte* einer Variablen auf eine andere.

Neben den direkten Effekten existieren noch zwei weitere wichtige Typen von Effekten:

1. *Indirekte Effekte:* Hierbei handelt es sich um Effekte, die durch eine Zwischenvariable (*Mediatorvariable* oder kurz: *Mediator*) vermittelt werden.

   Im Modell von Abb. 2-2 übt z.B. das latente Konstrukt η_1 einen indirekte Effekt auf Y_2 aus, wobei Y_1 als Mediator fungiert: $\eta_1 \rightarrow Y_1 \rightarrow Y_2$.

2. *Totale Effekte:* Hierbei handelt es sich um die Summe aus direkten und indirekten Effekten einer Variable auf eine andere.

   Im Modell von Abb. 2-2 übt z.B. die Kovariate *X* sowohl einen direkten Effekt als auch einen indirekten Effekt (via Y_1) auf die Messung Y_2 aus.

   Der totale Effekt von *X* auf Y_2 entspricht der Summe dieser beiden Effekte.

(b) Kovarianzbeziehungen: Hierbei handelt es sich um Relationen zwischen exogenen Variablen, deren kausaler Status nicht näher analysiert wird. Diese Relationen werden im Diagramm durch Kovarianzbögen (Bögen mit Doppelpfeilen) symbolisiert.
Hierbei sind drei Dinge zu beachten:

(i) Endogene Variablen besitzen keine unanalysierten kausalen Relationen weder untereinander noch zu den exogenen Variablen. Daher ist eine exogene Variable *niemals* durch einen Kovarianzbogen mit einer anderen Variablen verbunden.

(ii) Fehlerterme weisen keine Kovarianzbeziehungen mit anderen exogenen Variablen auf.

(iii) Gewöhnlich werden alle exogenen Variablen eines Modells mit Ausnahme der Fehlerterme mit Kovarianzbögen verbunden. Die Abwesenheit von Kovarianzbögen zwischen exogenen Variablen (wie in Abb. 2-2) symbolisiert explizit die Abwesenheit einer Beziehung zwischen den beiden Variablen.

3. *Modellparameter:*

Die Modellparameter repräsentieren alle quantitativen Aspekte des Modells. Aufgrund der Werte der Modellparameter wird die Kovarianzstruktur der Messungen ermittelt.

Die Parameter lassen sich in folgende Klassen unterteilen:

(a) Varianzen und Kovarianzen der latenten Konstrukte und Kovariaten: Diese werden mit ϕ (phi) bezeichnet.
Im Modell von Abb. 2-2 gehören zu dieser Kategorie die Parameter:

$$\phi^2_{\eta_1},\ \phi^2_{\eta_2},\ \phi^2_X \text{ und } \phi_{\eta_1\eta_2}.$$

Man beachte, dass die Parameter $\phi_{\eta_1 X}$ (= Kovarianz zwischen η_1 und *X*), sowie $\phi_{\eta_2 X}$ (= Kovarianz zwischen η_2 und *X*) beide als 0 betrachtet werden, was durch die fehlenden Kovarianzbögen angezeigt wird.

(b) Varianzen und Kovarianzen der Fehler: Diese werden mit θ (theta) bezeichnet.
Im Modell von Abb. 2-2 gehören in diese Kategorie die Parameter:

$$\theta^2_{\varepsilon_1}, \text{ und } \theta^2_{\varepsilon_2}.$$

Die Abwesenheit eines Kovarianzbogens zwischen den Fehlertermen symbolisiert, dass der Kovarianz zwischen den Fehlern, repräsentiert durch den Kovarianzparameter $\theta_{\varepsilon_1\varepsilon_2}$ den Wert 0 besitzt.

Beachte:
Endogene Variablen besitzen keinen zugehörigen Varianzparameter. Dies erklärt sich dadurch, dass die Varianzen (und die Kovarianzen) der endogenen Variablen mit Hilfe der Kovarianzstruktur der exogenen Variablen und den im Modell spezifizierten kausalen Relationen erklärt wird.

(c) Die Pfadkoeffizienten: Hierbei handelt es sich um die Parameter, welche die Stärke der direkten kausalen Einflüsse repräsentieren. Die Koeffizienten lassen sich in drei Gruppen unterteilen:

(i) Ladungskoeffizienten (oder einfach Ladungen): Diese betreffen den Einfluss der latenten Konstrukte auf die Messungen. Sie werden mit dem Symbol λ (lambda) bezeichnet.
Im Modell von Abb. 2-2 gibt es zwei Ladungskoeffizienten: $\lambda_{Y_1\eta_1}$ und $\lambda_{Y_2\eta_2}$.

(ii) Koeffizienten der Pfade von Kovariaten zu endogenen Variablen: Diese repräsentieren die Stärke des direkten Einflusses der Kovariaten auf die Messungen. Sie werden durch das Symbol γ (gamma) repräsentiert.
Im Modell von Abb. 2-2 gehören die Koeffizienten γ_{Y_1X} und γ_{Y_2X} in diese Kategorie.

(iii) Koeffizienten der Pfade zwischen endogenen Variablen. Sie repräsentieren die Stärke des direkten Einflusses einer Messung auf eine andere. Sie werden durch das Symbol β (beta) repräsentiert
Im Modell von Abb. 2-2 gehört der Koeffizient $\beta_{Y_2Y_1}$ in diese Kategorie.

Die Pfadkoeffizienten der Pfade von den Fehlertermen zu den Messungen wurden auf den Wert 1.0 fixiert.
Es ist möglich, auch diese Pfadkoeffizienten frei zu schätzen. Von dieser Möglichkeit wird jedoch in der Praxis (mit ganz wenigen Ausnahmen) nicht Gebrauch gemacht.

Notationskonvention:
Die Reihenfolge der Indizes der Pfadkoeffizienten ist immer derart, dass zuerst die *Zielvariable* (d.h. jene Variable, auf die der Pfeil hindeutet) aufscheint, gefolgt von der *Quellvariable* (= jene Variable, welche den Ausgangspunkt des Pfeils bildet).

Bsp.: Die Reihenfolge des Indizes des Koeffizienten $\lambda_{Y_1\eta_1}$ zeigt an, dass das latente Konstrukt η_1 als die Ursache (Ursprung des Pfeils) und die Messung Y_1 als die Wirkung (Endpunkt des Pfeils) betrachtet wird.

4. *Syntax des Models: Die linearen Modellgleichungen*

   Die vollständige formale Struktur des Modells beschreibt, wie die Varianzen der Messungen und die Kovarianzen zwischen diesen mit Hilfe der Modellparameter repräsentiert werden.
   Einen Teil dieser formalen Struktur bilden die linearen Modellgleichungen. Hierbei handelt es sich um ein lineares Gleichungssystem mit einer Gleichung pro endogener Variable.
   Die einzelnen Gleichungen lassen sich sehr einfach mit Hilfe des kausalen Diagramms erstellen:

   Die linke Seite der Gleichung bildet die endogene Variable.
   Die rechte Seite besteht aus der gewichteten Summe jener Variablen, von denen ein Pfeil auf die endogene Variable führt. Als Gewichte dienen hierbei die Pfadkoeffizienten, welche die Stärke des kausalen Einflusses repräsentieren.

   Das lineare Gleichungssystem für das Modell in Abb. 2-2 lautet:

$$\begin{aligned} Y_1 &= \lambda_{Y_1\eta_1} \cdot \eta_1 + \gamma_{Y_1X} \cdot X + \varepsilon_1 \\ Y_2 &= \lambda_{Y_2\eta_2} \cdot \eta_2 + \gamma_{Y_2X} \cdot X + \beta_{Y_2Y_1} \cdot Y_1 + \varepsilon_2 \end{aligned} \qquad (2\text{-}1)$$

Wir haben nun alle Komponenten des allgemeinen linearen Messmodells zur Modellierung der Kovarianzstruktur der Messungen behandelt. Es sei hier noch erwähnt, dass es sich bei dieser Familie von Messmodellen um so genannte *lineare Strukturgleichungsmodelle* handelt. Es existiert eine Reihe von Programmen – wie AMOS, EQS, Lisrel, Mplus, Ramona und Mx (auch als OpenMx innerhalb des Programms R) – welche zur Schätzung dieser Modelle verwendet werden können.
Wir wenden uns nun der Besprechung der klassischen Testmodelle zu.

2.2 Die Messmodelle der klassischen Testtheorie

Das grundlegende Modell der klassischen Testtheorie, sowie die hierauf basierenden spezifischeren Testmodelle können alle – wenn sie als latente Variablenmodelle interpretiert werden – als Spezialfälle des linearen Messmodells mit Effektindikatoren betrachtet werden können. Diese Betrachtungsweise der klassischen Testmodelle erleichtert wesentlich das Verständnis dieser Modelle und der damit in Zusammenhang stehenden testtheoretischen Konstrukte der *Reliabilität* und

Validität. Andererseits eröffnet sie die Möglichkeit der Konstruktion erweiterter Testmodelle.
Die Interpretation der klassischen Testmodelle als latente Variablenmodelle und die Anwendung effizienter Methoden zur Schätzung dieser Modelle als lineare Strukturgleichungsmodelle (oder genauer: als Modelle der Konfirmativen Faktorenanalyse) wurde erstmals von Jöreskog (1971) durchgeführt.

2.2.1 Das grundlegende Modell der klassischen Testtheorie

Das Kausaldiagramm in Abb. 2-3 zeigt das Kausalmodell (für 4 Testitems), welches die fundamentalen Annahmen der klassischen Testtheorie bezüglich der Kovarianzstruktur impliziert.

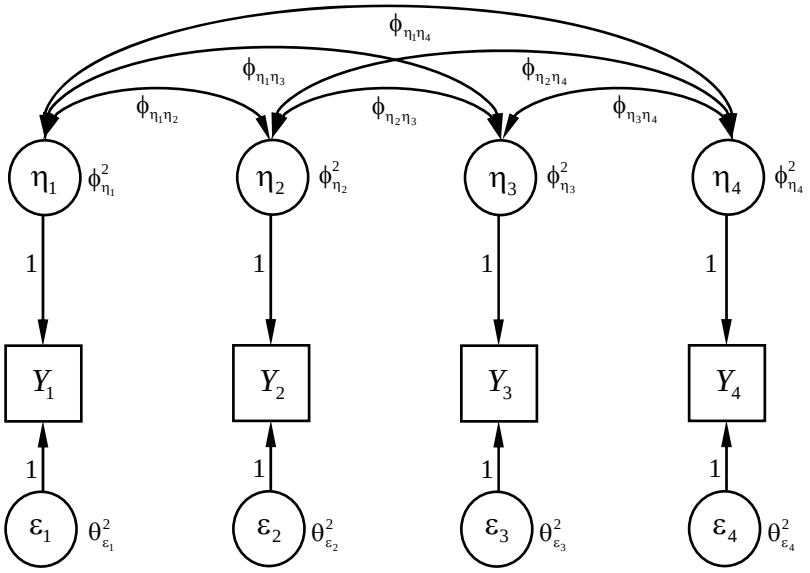


Abb. 2-3: *Das grundlegende klassische Testmodell für vier Tests.*

Die vier beobachteten Testwerte sind mit Y_1, Y_2, Y_3 und Y_4 bezeichnet, die Werte der zugehörigen latenten Konstrukte mit η_1, η_2, η_3 und η_4 und die Fehlerterme mit $\varepsilon_1, \varepsilon_2, \varepsilon_3$ und ε_4. Es ist unmittelbar evident, dass es sich hier um einen Spezialfall eines linearen Messmodells handelt.
Die aus diesem Messmodell folgenden zentralen Annahmen, welche jenen der klassischen Testtheorie entsprechen, sind:

1. Der beobachtete Testwert Y_i innerhalb einer Population von Untersuchungseinheiten ergibt sich aufgrund des Einflusses zweier Grössen: Dem Konstruktwertes η_i für die einzelnen Einheiten und dem Fehlerterm ε_i. Das lineare Gleichungssystem, welches dem Modell zugrunde liegt, besitzt die allgemeine Form:

$$Y_i = \eta_i + \varepsilon_i \qquad (i = 1,2,3,4). \qquad (2\text{-}2)$$

Man sagt auch, dass das Modell den beobachteten Testwert Y_i in den (wahren) Wert des Konstrukts η_i und den Messfehler ε_i zerlegt.

2. Konstruktwerte und Messfehler sind unkorreliert, sowohl innerhalb eines Tests als auch zwischen verschiedenen Tests:

   $\mathrm{Kov}(\eta_i, \varepsilon_j) = 0$, für alle *i, j*.

   Im Modell von Abb. 2-3 ist dies durch das Fehlen von Kovarianzbögen zwischen den Kreisen, welche Fehlerterme repräsentieren und den Kreisen, welche die latenten Konstrukte symbolisieren, ausgedrückt.

3. Die Fehlerterme ε_i und ε_j zu zwei verschiedenen Tests sind ebenfalls unkorreliert:

   $\mathrm{Kov}(\varepsilon_i, \varepsilon_j) = 0$.

   Ähnlich wie im Falle der fehlenden Kovarianz zwischen Fehler und Konstrukten ist die fehlende Kovarianz zwischen den Fehlerntermen im Modell von Abb. 2-3 durch die Abwesenheit von Kovarianzbögen repräsentiert.

Bemerkung zur fehlenden Kovarianz zwischen den Fehlertermen:
In manchen Abhandlungen wird die letzte Bedingung (d.h. die fehlende Kovarianz zwischen den Fehlertermen) nicht als genuine Annahme der klassischen Testtheorie betrachtet, sondern als zusätzliche Annahme.
Das Modell von Abb. 2-3 kann durch Einführung von Kovarianzbögen zwischen den Fehlertermen erweitert werden. Dieses erweiterte Modell würde die einschränkende Bedingung der fehlenden Kovarianzen zwischen den Fehlern nicht mehr repräsentieren.

Im Modell von Abb. 2-3 fungieren die Varianzen von η_i als freie Parameter, während die *Pfadkoeffizienten* der Pfade $\eta_i \rightarrow Y_i$ alle auf den Wert 1 fixiert wurden.
Ein äquivalentes Modell, welches die Kovarianzstruktur gleich gut erklärt wie jenes von Abb. 2-3 ergibt sich, indem man die Varianzen der latenten Konstrukte alle auf 1 setzt und stattdessen die Ladungskoeffizienten als freie Parameter schätzt. Dieses Modell ist in Abb. 2-4 dargestellt. Man sagt auch, dass es sich um das gleiche Modell in *alternativer Parametrisierung* handelt.
In dieser Parametrisierung entsprechen die Kovarianzen zwischen den latenten Konstrukten den Korrelationen.

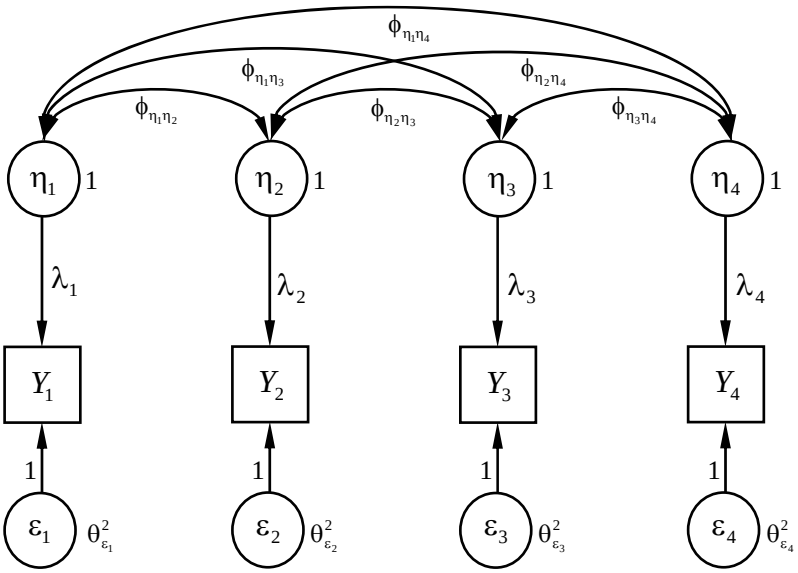


Abb. 2-4: *Das grundlegende klassische Testmodell von Abb. 2-3 in alternativer Parametrisierung.*

2.2.2 Spezialfälle klassischer Testmodelle: Das kongenerische, τ (tau) – äquivalentes und parallele Testmodell

Im Folgenden betrachten wir drei Testmodelle, welche als Spezialfälle des allgemeinen Messmodells von Abb. 2-3 zu betrachten sind. Diese Spezialfälle ergeben sich, indem Beschränkungen auf bestimmte Modellparameter spezifiziert werden. Aufgrund der Beschränkungen machen diese Modelle spezifischere Annahmen bezüglich der Kovarianzstruktur der beobachteten Indikatoren. Diese Annahmen können – falls genügend Indikatoren vorliegen, an Hand von Daten getestet werden.

Die drei Modelle können als *ineinander geschachtelt* (*eingebettet*) betrachtet werden. Dies bedeutet, dass jedes speziellere Modell auch alle Annahmen der übergeordneten Modelle beinhaltet, bzw. dass jedes speziellere Modell aus den übergeordneten Modellen durch Beschränkung weiterer Parameter entsteht (die Beschränkungen der übergeordneten Modelle bleiben in den spezifischeren Modellen erhalten).

Das *kongenerische Modell* stellt das generellste Modell dar. Diesem folgt vom *Modell τ -äquivalenter Messungen*. Das *parallele Modell* ist das spezifischste Modell mit den meisten Restriktionen (und mit den stärksten Annahmen).

Wir beginnen mit der Besprechung des generellsten der drei Modell.

2.2.2.1 Das Modell kongenerischer Tests

Das kongenerische Testmodell nimmt an, dass die Kovarianzen zwischen den beobachteten Testwerten mit Hilfe eines einzigen Faktors erklärt werden können.

Konzept 2-3: *Kongenerische Maße (Tests):*

Zwei (oder mehrere) Maße (Indikatoren, Tests) heissen *kongenerisch*, wenn sie von ein und demselben gemeinsamen latenten Faktor (und nur von diesem) kausal beeinflusst werden und die Fehlerterme unkorreliert sind.
Bemerkung:
Anstelle *kongenerischen* Maßen, spricht man auch von *homogenen oder eindimensionalen* Maßen.

Das kongenerische Modell mit vier Indikatoren ist in Abb. 2-5 gezeigt.

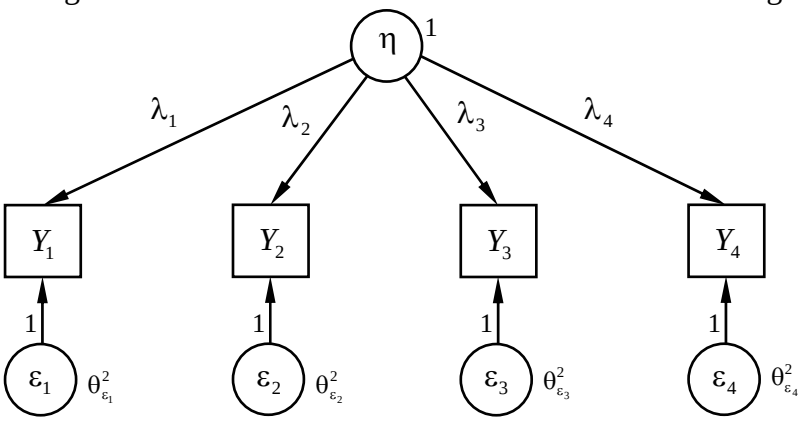


Abb. 2-5: *Das Testmodell kongenerische Maße.*

Dieses Modell ergibt sich aus dem allgemeinen Messmodell von Abb. 2-3 bzw. Abb. 2-4 durch Einführung der zusätzlichen Annahme, dass die Konstruktwerte für die einzelnen Tests perfekt korreliert sind:

$$\text{Korr}(\eta_i, \eta_j) = 1, \text{ für alle } i \neq j, \qquad (2\text{-}3)$$

In diesem Falle können vier Variablen statistisch durch eine einzige ersetzt werden, wobei jedoch die Ladungen unterschiedlich sind.

Wir wollen nun etwas genauer untersuchen, wie sich das Modell von Abb. 2-5 aus den Modellen von Abb. 2-3 und Abb. 2-4 durch Fixieren von Parametern ergibt.

Bezüglich des Modells von Abb. 2-3 ergibt sich eine Schwierigkeit, welche darin besteht, dass die Korrelationen keine Parameter des Modells sind. Daher können die obigen Restriktionen nicht direkt gesetzt werden.
Ein einfacher Trick ermöglicht es jedoch, die Parameter des Modells derart zu fixieren, dass die Korrelationen zwischen latenten Variablen den Wert 1 aufweisen.

Methode 2-1: *Fixieren der Korrelation zwischen zwei latenten Variablen auf den Wert 1.0:*

Gegeben:
Das Teilmodell von Abb. 2-6 mit den beiden latenten Variablen η_i und η_j. Diese weisen die Varianzen ϕ_i^2 und ϕ_j^2 auf, sowie die Kovarianz ϕ_{ij}.

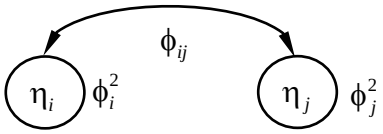


Abb. 2-6: *Teilmodell bestehend aus zwei latenten Variablen*

Gesucht:
Beschränkung der Korrelation auf den Wert 1.0.
Abb. 2-7 zeigt das Modell, welches die gewünscht Restriktion implementiert.

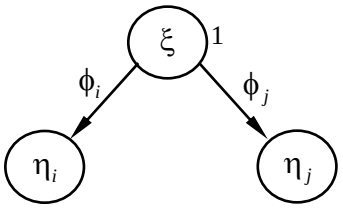


Abb. 2-7: *Modell mit der Beschränkung* $\text{Korr}(\eta_i, \eta_j) = 1.0$.

Es wurde eine Pseudovariable ξ mit Varianz 1 eingeführt.
Die Koeffizienten der Pfade $\xi \rightarrow \eta_i$ und $\xi \rightarrow \eta_j$ werden frei geschätzt. Sie wurden hier mit den Symbolen ϕ_i und ϕ_j bezeichnet und nicht – wie üblich – durch den Buchstaben λ.
Dies soll anzuzeigen, dass die Koeffizienten den Standardabweichungen der beiden Variablen η_i und η_j entsprechen.
Man beachte, dass keine Residuenterme für die Variablen η_i und η_j gesetzt wurden. Dies kann zu Warnungen des Programms führen. Um diese zu vermeiden, kann man Residuen mit Varianz 0 verwenden.
Der Beweis, dass bei Anwendung dieser Methode die Korrelation zwischen η_i und η_j tatsächlich den Wert 1 aufweist, ist auf einfache Weise durch Anwendung der Kovarianzrechnung zu erbringen und bleibt der Leserin überlassen (Übung 2-1).

Mit Hilfe dieser Methode können nun im klassischen Testmodell von Abb. 2-3 die Beschränkungen zur Implementation eines kongeneri-

schen Modells gesetzt werden. Das resultierende Modell mit den Beschränkungen ist in Abb. 2-8 dargestellt.

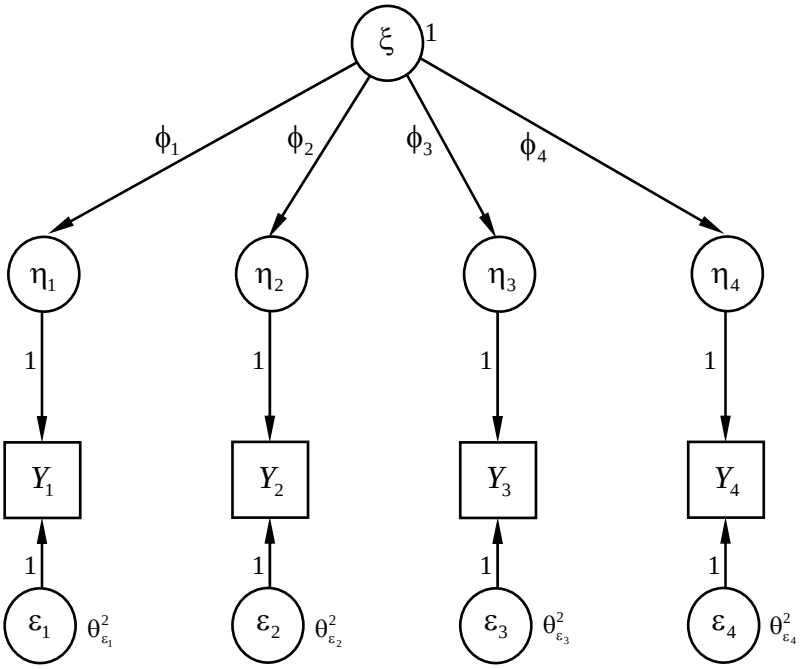


Abb. 2-8: *Das kongenerisches Modell für vier Tests.*

Das gezeigte Modell ist identisch zu jenem in Abb. 2-5. Dies ergibt sich aufgrund der folgenden Überlegung:
Der totale Effekt der Variablen ξ auf die Messung Y_i ergibt sich aus dem Produkt der Pfadkoeffizienten der Pfade auf dem Weg von ξ zu Y_i. Somit ist der totale Effekt gleich $(1 \cdot \phi_1)$ bzw. ϕ_i. Setzt man $\phi_i = \lambda_i$, wobei λ_i dem Ladungskoeffizienten für das Maß Y_i in Abb. 2-5 darstellt, so ergibt sich unmittelbar die Äquivalenz der beiden Modelle, indem man im Modell von Abb. 2-8 die (überflüssigen) Variablen η_i eliminiert und die Pfade direkt von ξ nach Y_i verlaufen lässt.

Für das Modell in Abb. 2-4 ist die Implementation der Beschränkungen (2-3) hingegen einfach, da in diesem Modell die Kovarianzen den Korrelationen zwischen den latenten Konstrukten entsprechen. Setzt man daher diese Kovarianzen auf den Wert 1, so erhält man das kongenerische Modell (Abb. 2-9).
In alle drei Modellen (Abb. 2-5, Abb. 2-8 und Abb. 2-9) werden die beobachteten Varianzen und Kovarianzen wie folgt durch die Parameter repräsentiert:

$$\mathrm{Kov}(Y_i, Y_j) = \lambda_i \cdot \lambda_j \left(= \phi_i \cdot \phi_j\right) \qquad (2\text{-}4)$$

und

$$\mathrm{Var}(Y_i) = \lambda_i^2 + \theta_{\varepsilon_i}^2 \,. \qquad (2\text{-}5)$$

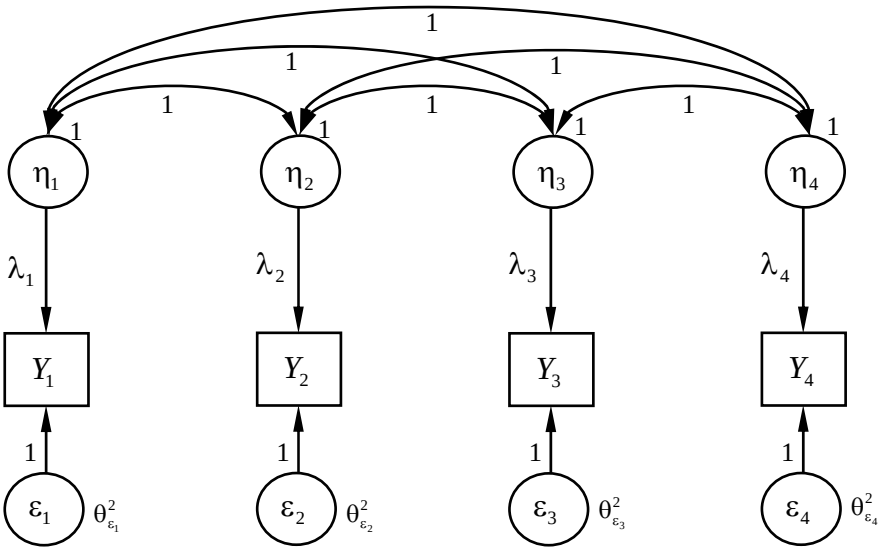


Abb. 2-9: *Das kongenerische Testmodell für vier Tests.*

Hierbei repräsentiert λ_i^2 die durch den kausalen Einfluss von η_i auf Y_i induzierte Varianz, während die Varianz des Fehlers durch den Term $\theta_{\varepsilon_i}^2$ repräsentiert wird. Die durch das latente Konstrukt bedingte Varianz wird auch als True-Score Varianz bezeichnet (siehe Konzept 2-7).

Das kongenerische Testmodell (ohne Zusatzbeschränkungen) ist für drei Messungen exakt identifiziert, d.h. es hat gleich viele eindeutige Parameter wie freie Datenpunkte. Zur Prüfung des Modells benötigt man daher (ohne Einführung zusätzlicher Beschränkungen) mindestens vier Tests.

Bemerkung zur Identifikation von Parametern und Modellen:

Das Konzept der *Identifikation von Modellen und Parametern* wird in Abschnitt xxxx näher ausgeführt.

Für die gegenwärtige Diskussion reicht die folgende Information:

1. Nur Modelle und Parameter, die identifiziert sind, sind inhaltlich interpretierbar.
2. *Exakt identifizierte Modelle* sind nicht prüfbar, da sie gleich viele freie Parameter wie Datenpunkte enthalten
3. *Überidentifizierte Modelle* sind prüfbar, da sie weniger Parameter enthalten als Datenpunkte vorhanden sind.

2.2.2.2 Das Modell τ (tau) – äquivalenter Tests

Das Modell τ-äquivalenter Tests ist ein Spezialfall des kongenerischen Modells:

Konzept 2-4: *τ-äquivalente Maße (Tests):*

Zwei (oder mehrere) Maße (Indikatoren, Tests) sind *tau-äquivalent*, falls sie kongenerisch sind und (zusätzlich) der Einfluss des latenten Faktors auf alle Maße identisch ist.

Das Modell τ-äquivalenter Tests ergibt sich daher aus dem kongenerischen Modell durch Gleichsetzung der Ladungen:

$$\lambda_1 = \lambda_2 = \cdots = \lambda_p . \tag{2-6}$$

Dieses Modell impliziert daher die folgende Struktur der Varianzen und Kovarianzen:

$$\mathrm{Kov}\left(Y_i, Y_j\right) = \lambda^2 \tag{2-7}$$

und

$$\mathrm{Var}\left(Y_i\right) = \lambda^2 + \theta^2_{\varepsilon_i} . \tag{2-8}$$

Dies ergibt die folgende Struktur der vom Modell implizierten Kovarianzmatrix:

$$\hat{\boldsymbol{\Sigma}} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & \cdots & Y_p \end{array} \\ \begin{bmatrix} \lambda^2 + \theta_1^2 & \lambda^2 & \cdots & \lambda^2 \\ \lambda^2 & \lambda^2 + \theta_2^2 & \cdots & \lambda^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda^2 & \lambda^2 & \cdots & \lambda^2 + \theta_p^2 \end{bmatrix} \end{array} \tag{2-9}$$

Hierbei gilt:

- λ symbolisiert die (für alle Indikatoren identischen) Ladungskoeffizienten.
- θ_i^2 bezeichnet die Varianz des Fehlerterms ε_i.

Das Modell τ-äquivalenter Tests ist exakt identifiziert, falls mindestens 2 Indikatoren vorhanden sind (und die Metrik der latenten Variablen fixiert wurde, entweder durch Fixieren der Varianz des latenten Konstrukts oder durch Fixieren der Ladungen [in beiden Fällen werden die Grössen meist auf den Wert 1 fixiert]). Das Modell ist bei Verwendung von mindestens 3 Indikatoren prüfbar. In diesem Falle liegen 3 unabhängige Kovarianzen vor, die – gemäss Modell – alle drei gleich sein müssen.

Bemerkung:

In der Testliteratur gibt es die Unterscheidung zwischen *τ-äquivalenten* und *essentiell τ-äquivalenten* Tests. Im ersten Fall sind auch die beobachteten Mittelwerte der Tests identisch im zweiten Fall nicht.

Da wir keine Mittelwerte schätzen, entfällt die Unterscheidung.

2.2.2.3 DAS MODELL PARALLELER TESTS

Ein weiteres, spezifischeres Modell als jenes τ-äquivalenter Tests ist das Modell paralleler Tests.

Konzept 2-5: *Parallele Maße (Tests):*

Zwei (oder mehrere) Maße (Indikatoren, Tests) sind *parallel*, falls sie τ-äquivalent sind und zusätzlich die Fehlervarianzen für beide Maße identisch sind.

Das parallele Modell ergibt sich daher aus dem τ-äquivalenten, durch Gleichsetzung der Fehlervarianzen:

$$\theta^2_{\varepsilon_1} = \theta^2_{\varepsilon_2} = \cdots = \theta^2_{\varepsilon_p} \left(= \theta^2\right). \tag{2-10}$$

Dieses Modell impliziert daher die folgende Struktur der Varianzen und Kovarianzen:

$$\mathrm{Kov}\left(Y_i, Y_j\right) = \lambda^2 \tag{2-11}$$

und

$$\mathrm{Var}\left(Y_i\right) = \lambda^2 + \theta^2 . \tag{2-12}$$

Dies ergibt die folgende Struktur der vom Modell implizierten Kovarianzmatrix:

$$\hat{\boldsymbol{\Sigma}} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & \cdots & Y_p \end{array} \\ \begin{bmatrix} \lambda^2+\theta^2 & \lambda^2 & \cdots & \lambda^2 \\ \lambda^2 & \lambda^2+\theta^2 & \cdots & \lambda^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda^2 & \lambda^2 & \cdots & \lambda^2+\theta^2 \end{bmatrix} \end{array} \tag{2-13}$$

Hierbei gilt:

- λ symbolisiert die (für alle Indikatoren identischen) Ladungskoeffizienten.
- θ^2 bezeichnet die (für alle Indikatoren) identische) Varianz der Fehlerterme ε_i.

Das Modell paralleler Tests ist exakt identifiziert, falls mindestens 1 Indikator vorhanden ist (und die Metrik der latenten Variablen fixiert wurde). Das Modell ist bei Verwendung von mindestens 2 Indikatoren prüfbar. In diesem Falle macht das Modell die prüfbare Aussage, dass die Varianzen der beiden Indikatoren identisch sind.

Bemerkung:

Analog zur Unterscheidung zwischen τ-äquivalenten und essentiell τ-äquivalenten Tests gibt es im parallen Fall jene zwischen *strikt parallelen* Tests (identische Mittelwerte) und *parallelen* Tests (unterschiedliche Mittelwerte).
Wie zuvor ist die Unterscheidung für die aktuelle Diskussion irrelevant.

2.2.2.4 Illustration der drei Testmodelle

Zur Illustration der Konzepte betrachten wir ein Beispiel von Jöreskog (1971), in welchem Daten von Lord (1957) neu analysiert wurden.

Bsp.2-3: *Prüfung von Testmodellen (Jöreskog, 1971):*

Gegeben: 4 Arten von Vokabeltests:

X_1, X_2 sind zwei Tests, bestehend aus je 15 Items, die ohne Zeitdruck präsentiert wurden.

Y_1, Y_2 sind zwei Tests, bestehend aus je 75 Items, die unter Zeitdruck präsentiert wurden.

Die Anzahl der Testpersonen betrug $N = 649$. Tab. 2-1 zeigt die Kovarianzmatrix der 4 Tests.

	X_1	X_2	Y_1	Y_2
X_1	86.3979			
X_2	57.7751	86.2632		
Y_1	56.8651	59.3177	97.2850	
Y_2	58.8986	59.6683	73.8201	97.8192

Tab. 2-1: *Kovarianzmatrix von vier Tests (Nach Jöreskog, 1971).*

Jöreskog untersuchte die folgenden 4 Hypothesen:

H_1: X_1 und X_2, sowie Y_1 und Y_2 sind jeweils parallel. Die beiden Paare sind jedoch nicht kongenerisch.

H_2: X_1 und X_2, sowie Y_1 und Y_2 sind jeweils parallel und alle 4 Tests sind kongenerisch.

H_3: X_1 und X_2, sowie Y_1 und Y_2 sind jeweils kongenerisch, aber die beide Paare (zusammengenommen) sind nicht kongenerisch.

H_4: Die 4 Tests sind kongenerisch, jedoch nicht notwendigerweise parallel.

Die einfachste Methode zur Untersuchung dieser 4 Hypothesen besteht darin, jenes Modell zu erzeugen, welches der allgemeinsten Hypothese entspricht, und die den anderen Hypothesen entsprechenden Modelle durch Beschränkung der Parameter zu spezifizieren.

Das allgemeinste Modell ist durch Hypothese H_3 gegeben. Das zugehörige lineare Messmodell ist in Abb. 2-10 dargestellt.

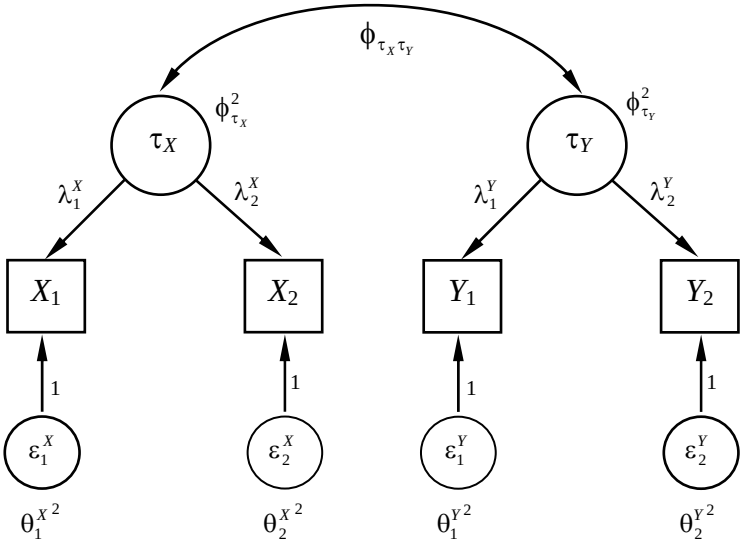


Abb. 2-10: *Lineares Messmodell zur Prüfung von Hypothese H_3.*

Hieraus ergibt sich das zu Hypothese H_1 gehörige Modell durch Einführung der Beschränkungen:

$$\lambda^X_1=\lambda^X_2=\lambda^X \qquad \lambda^Y_1=\lambda^Y_2=\lambda^Y \qquad \text{und} \qquad \theta^{X\,2}_1=\theta^{X\,2}_2=\theta^{X\,2} \qquad \theta^{Y\,2}_1=\theta^{Y\,2}_2=\theta^{Y\,2}.$$

Das resultierende Modell ist identifiziert, sobald die Skala der beiden Faktoren fixiert wird. Dies kann durch Setzen der beiden Varianzparameter $\phi^2_{\tau_X}=1$ und $\phi^2_{\tau_Y}=1$ geschehen.

Aus dem Modell, welches Hypothese H_1 repräsentiert, ergibt sich das Modell für H_2, indem zusätzlich der Parameter $\phi_{\tau_X\tau_Y}=1$ gesetzt wird. Dies entspricht der Spezifikation einer perfekten Korrelation $\rho_{\tau_X\tau_Y}=1$ zwischen den beiden latenten Variablen τ_X und τ_Y.

Das resultierende Modell ist identisch zu einem Modell mit nur einer latenten Variablen, d.h. einem Modell, welches annimmt, dass alle 4 Indikatoren (mindestens) kongenerisch sind.

Das Modell zur Prüfung von Hypothese H_4 ergibt sich schliesslich aus dem Modell zu H_3, indem $\phi_{\tau_X\tau_Y}=1$ gesetzt wird.

Beachte:

Die verwendet Äquivalenz $\phi_{\tau_X\tau_Y}=1 \Leftrightarrow \rho_{\tau_X\tau_Y}=1$ gilt nur, falls die Varianzparameter auf $\phi^2_{\tau_X}=1$ und $\phi^2_{\tau_Y}=1$ gesetzt wurden. Andernfalls gilt: $\phi_{\tau_X\tau_Y}=\phi_{\tau_X}\cdot\phi_{\tau_Y}$.

Nach dieser Behandlung der 3 Arten von Tests wenden wir uns dem Konzept der *Reliabilität* zu, welches innerhalb der Testtheorie einen zentralen Stellenwert hat.

2.3 Reliabilität: Konzept und Schätzung

In diesem Abschnitt wollen wir das Konzept der *Reliabilität*, sowie deren Messung näher betrachten. Es wird sich zeigen, dass mit den uns zur Verfügung stehenden Methoden (graphische Darstellung des Messmodells und Kovarianzrechnung) sowohl das Konzept der Reliabilität als auch die der Messung (bzw. der Schätzung) zugrunde liegenden Annahmen einfach zu illustrieren sind.

2.3.1 Darstellung des Konzepts der Reliabilität

Wir beginnen mit der Definition des Konzepts.

Konzept 2-6: *Reliabilität:*

Die *Reliabilität einer abhängigen Variablen* ist der durch die unabhängigen Variablen erklärte Varianzanteil.

Bemerkungen zur Definition der Reliabilität:

1. Um die Definition möglichst allgemein zu halten, wurde nicht von der *Reliabilität einer Messung* oder der *Reliabilität eines Tests* gesprochen.
2. Falls die Beziehung zwischen den unabhängigen Variablen $X_1, X_2, \ldots, X_n$ und der abhängigen Variable Y linear ist, so ist der multiple quadrierte Korrelationskoeffizient $R^2_{Y.X_1,X_2,\ldots,X_n}$ ein unverzerrter Schätzer der Reliabilität. Dieser entspricht dem Quadrat der Korrelation zwischen Y und der Schätzung $\hat{Y} = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n$ von Y.

Konzept 2-7: *True-Score Varianz:*

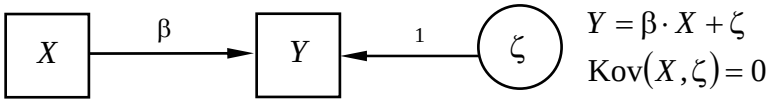
Die *True-Score Varianz* ist der durch die latenten Konstrukte erklärte (bedingte, verursachte) Varianzanteil in der Messung.

Bemerkungen:

1. Von True-Score Varianz spricht man im Allgemeinen nur, wenn die unabhängigen Variablen latente Konstrukte sind.
2. Es ist auch möglich von der True-Score Varianz einer Teilmenge der im Modell vorhandenen Konstrukte zu sprechen. In diesem Falle handelt es sich um die durch diese Konstrukte bedingte Varianz in der Messung.
3. Die True-Score Varianz ist strikt von der Varianz der latenten Konstrukte zu unterscheiden. Letztere geht allerdings in die Berechnung der True-Score Varianz ein.

Bsp.2-4: Reliabilität und lineares Regressionsmodell:

Gegeben: Das linearen Regressionsmodell interpretiert als lineares Kausalmodell:



$$Y = \beta \cdot X + \zeta$$
$$\mathrm{Kov}(X, \zeta) = 0$$

Wir zeigen zuerst, dass sich die Gesamtvarianz von *Y* in zwei additive Komponenten zerlegen lässt:

1. *Die durch X erklärte Varianz in Y*: Varianz von *Y* aufgrund von Änderungen der unabhängigen Variable *X*.
2. *Die residuale Varianz:* Der Varianzanteil von *Y*, der nicht durch *X* erklärbar ist. Dies ist die Varianz des Residuums ζ.

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Kov}(Y, Y) \\ &= \mathrm{Kov}(\beta \cdot X + \zeta, \beta \cdot X + \zeta) \\ &= \beta^2 \cdot \mathrm{Kov}(X, X) + \mathrm{Kov}(\zeta, \zeta) \\ &= \beta^2 \cdot \mathrm{Var}(X) + \mathrm{Var}(\zeta)\end{aligned}$$

Dividiert man beide Seiten durch $\mathrm{Var}(Y)$, so erhält man die Varianzanteile:

$$1 = \frac{\beta^2 \cdot \mathrm{Var}(X)}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}(\zeta)}{\mathrm{Var}(Y)}$$

Der erste Term auf der rechten Seite repräsentiert den durch die unabhängige Variable *X* erklärten Anteil an Varianz in *Y* und der zweite den Varianzanteil aufgrund anderer Einflüssen, die durch das Residuum ζ repräsentiert sind. Es gilt somit:

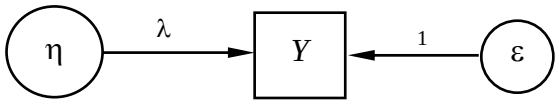
$$\mathrm{Rel}(Y) = \frac{\beta^2 \cdot \mathrm{Var}(X)}{\mathrm{Var}(Y)}. \qquad (2\text{-}14)$$

Auf einfache Weise lässt sich auch zeigen, dass die Reliabilität der quadrierten Korrelation $R^2_{Y.X}$ zwischen *Y* und $\hat{Y} = \beta \cdot X$ dem durch die unabhängige Variable *X* vorhergesagten Werten von *Y* entspricht (Übung 2-6). $R^2_{Y.X}$ ist im Falle einer unabhängigen Variable identisch zur quadrierten Korrelation R^2_{XY} zwischen *X* und *Y*.

Das nächste Beispiel demonstriert die Berechnung der Reliabilität eines Indikators.

Bsp.2-5: Reliabilität eines Indikators:

Gegeben: Ein einfaches Messmodell:



$$Y = \lambda \cdot \eta + \varepsilon$$
$$\mathrm{Kov}(\eta, \varepsilon) = 0$$

Abb. 2-11: *Ein einfaches lineares Messmodell.*

Dieses Modell ist strukturgleich zum linearen Regressionsmodell von Bsp.2-4. Der einzige Unterschied besteht darin, dass die unabhängige Variable nun latent ist (und die Bezeichnung der Grössen verändert wurde). Die Reliabilität von *Y* ermittelt sich daher völlig analog zu Bsp.2-4:

$$\mathrm{Rel}(Y) = \frac{\lambda^2 \cdot \mathrm{Var}(\eta)}{\mathrm{Var}(Y)} \qquad (2\text{-}15)$$

Der Zähler $\lambda^2 \cdot \mathrm{Var}(\eta)$ repräsentiert die True-Score Varianz.

Analog zu Bsp.2-4 lässt sich zeigen, dass die Reliabilität von *Y* der quadrierten Korrelation zwischen *Y* und dem durch η vorhergesagten Wert $\hat{Y} = \lambda \cdot \eta$ entspricht, bzw. der Korrelation zwischen *Y* und η:

$$\mathrm{Rel}(Y) = \frac{[\mathrm{Kov}(\eta, Y)]^2}{\mathrm{Var}(\eta) \cdot \mathrm{Var}(Y)} = R^2_{Y.\eta} = R^2_{Y\eta} \qquad (2\text{-}16)$$

Am Ende dieser Darstellung des Konzepts der *Reliabilität* noch zwei Bemerkungen:

1. Der Begriff *Reliabilität* bezieht sich auf die Messung und nicht auf das Konstrukt, d.h. eine Messung ist reliabel aber nicht das gemessene Konstrukt (das Gleiche gilt übrigens auch für die Validität eines Tests).
2. Reliabilität ist eine populationsbezogene Grösse. Dies bedeute, dass von Reliabilität nur im Zusammenhang mit einer Population, *in welcher der Wert des zu messenden Konstrukts variiert*, gesprochen werden kann. Falls das Konstrukt in der Population konstant ist, so ist die Reliabilität der Messung gleich Null.

2.3.2 Traditionelle Ansätze zur Messung der Reliabilität von Tests

Zur Messung der Reliabilität von Tests verwendet man traditionellerweise eine der folgenden drei Methoden:

Methode 2-2: *Traditionelle Methoden zur Messung der Reliabilität eines Test:*

1. *Test-Retest – Methode:*
   Der gleiche Test wird zu zwei verschiedenen Zeitpunkten angewendet.

2. *Alternativformen:*
   Der Test liegt in zwei Formen vor. Diese werden zu zwei verschiedenen Zeitpunkten appliziert.
3. *Testhälften:*
   Die zur Messung eines latenten Faktors verwendeten Testitems werden in zwei Klassen aufgeteilt (z.B. ungerade und gerade Testitems). Die Messung beider Teile erfolgt zum gleichen Zeitpunkt.

Je nach Erhebungsmethode haben die Indikatoren Y_1 und Y_2 unterschiedliche Bedeutung:

1. Im Falle von *Test-Retest* handelt es sich um die Messergebnisse für ein und dasselbe Testitem zu zwei verschiedenen Zeitpunkten.
2. Bei *Verwendung* von *Alternativformen* handelt es sich um eine Messungen von verschiedenen Testitems zu verschiedenen Zeitpunkten.
3. Bei *Verwendung* von *Testhälften* handelt es sich um die Messungen verschiedener Testitems zum gleichen Zeitpunkt.

Die Reliabilität wird dann meist mit Hilfe des Korrelationskoeffizienten zwischen den beiden Messungen Y_1 und Y_2 quantifiziert und zwar unabhängig davon, welche empirische Methode zur Messung verwendet wurde.

2.3.3 Probleme und Grenzen des traditionellen Ansatzes

Der im letzten Absatz beschriebene Ansatz zur Messung der Reliabilität hat einen markanten »Schönheitsfehler«: Er spezifiziert nicht, unter welchen Bedingungen der Korrelationskoeffizient ein korrekter Schätzer der Reliabilität darstellt. Die Gültigkeit hängt aber wesentlich vom Messmodell zur Beschreibung der Messsituation ab. Es gilt nämlich das folgendes Prinzip, welches als Spezialfall von Prinzip 1-1 betrachtet werden kann:

Prinzip 2-2: *Prinzip der modellabhängigen Messung der Reliabilität eines Test:*

Die Messung der Reliabilität eines Tests ist modellabhängig, d.h. das Ergebnis der Messung ist abhängig vom spezifizierten Messmodell.

Aus Prinzip 2-2 folgt, dass zur Ermittlung der Reliabilität eines Test ein Messmodell spezifiziert und dessen Gültigkeit geprüft werden muss.

Bezüglich der traditionellen Vorgehensweise der Verwendung der Korrelation zwischen zwei Tests als Maß für die Reliabilität ergibt sich aufgrund des Ergebnisses von Übung 2-9 die folgende Schlussfolgerung:

Bei Vorliegen paralleler Messungen (siehe Konzept 2-5) – d.h. das parallele Messmodel beschreibt die Struktur der Indikatoren korrekt – ist der Korrelationskoeffizient ein unverzerrter Schätzer der Reliabilität.

Das folgende Beispiel zeigt mögliche Ursachen für Abweichungen vom parallelen Modell auf und präsentiert die zugehörigen erweiterten Modelle.

Bsp.2-6: Ursachen für mögliche Abweichungen vom parallelen Modell im Falle wiederholter Messung

Gegeben: Drei verschieden Messmodelle zur Modellierung wiederholter Messung (Abb. 2-12):

Modell (a):
Dieses Modell repräsentiert eine Situation, in der die erste Messung einen direkten Einfluss auf die zweite Messung ausübt. Diese Situation liegt zum Beispiel vor, wenn die Messungen kurz aufeinander folgen und daher die Wahrscheinlichkeit besteht, dass ein Teil der Personen sich an die Antwort im ersten Test erinnert.

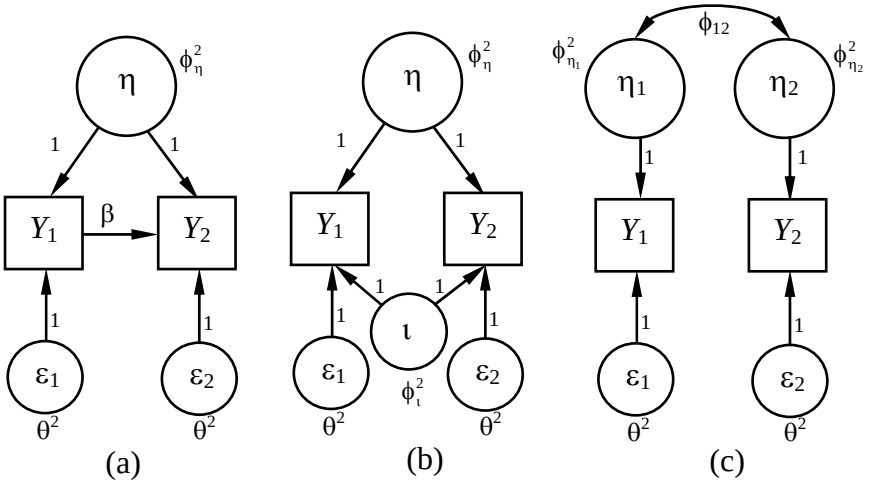


Abb. 2-12: *Mögliche Messmodelle zur Beschreibung der Messsituation im Falle wiederholter Messung: (a) Erste Messung hat Effekt auf die zweite Messung, (b) Vorhandensein eines itemspezifischen Faktors ι und (c) Es liegt kein stabiles Merkmal vor.*

Modell (b):
Dieses Modell repräsentiert eine Situation, in der ein itemspezifischer Faktor ι vorliegt, Dieser induziert eine zusätzliche Abhängigkeit zwischen den beiden Messungen. So können zum Beispiel Fragen eines Fragebogens einen derartigen itemspezifischen Faktor aufweisen, indem sie in irgendeiner Hinsicht extrem sind (z.B. peinlich, schlecht formuliert, etc.).

Modell (c):
Dieses Modell beschreibt eine Situation, in der kein unveränderliches Merkmal vorliegt. Die Korrelation zwischen den wahren Werten ist daher nicht gleich 1. Diese Situation tritt z.B. bei Messungen von Emotionen und Stimmungen auf, die keine festen Merkmale sondern transiente Zustände darstellen (Vergleiche Übung 2-5).

Eine weit verbreitete Praxis besteht in der Bildung der Summen *Y* aus den einzelnen Messungen $Y_1, Y_2, \ldots, Y_m$: $Y = Y_1 + Y_2 + \cdots + Y_m$ (siehe Anhang). In diesem Zusammenhang stellt sich die Frage nach der Reliabilität $\mathrm{Rel}(Y)$ des Summenwertes.

2.3.4 Reliabilität der Summe von Messungen

Die bekanntesten Reliabilitätsmaße ergeben sich durch die *Spearman-Brown – Formel* und durch *Cronbachs α (Koeffizient α)*. Im Folgenden beschäftigen wir uns mit der Berechnung dieser Grössen mit Hilfe von Strukturgleichungsmodellen, sowie mit den Problemen, die sich im Zusammenhang mit deren Anwendung ergeben.

Konzept 2-8: *Die Spearman-Brown – Formel der Testverlängerung:*

Gegeben:
m parallele Tests: $Y_1, Y_2, \ldots, Y_m$. Die Reliabilität der Items betrage ρ (identisch für alle Items).

Es gilt:
Die Reliabilität $\mathrm{Rel}(Y)$ der Summe $Y = Y_1 + Y_2 + \cdots + Y_m$ ist gegeben durch:

$$\mathrm{Rel}(Y) = \frac{m \cdot \rho}{1 + (m-1) \cdot \rho} \qquad (2\text{-}17)$$

Gleichung (3-36) ist die *Spearman-Brown* – Formel.

Beachte:
Zentrale Voraussetzung für die Gültigkeit der Spearman-Brown – Formel ist die Parallelität der Testitems, denn nur parallele Testitems messen ein Konstrukt mit identischer Reliabilität.

Konzept 2-9: *Cronbachs α:*

Gegeben:
m τ-äquivalente Tests: $Y_1, Y_2, \ldots, Y_m$.

Es gilt:
Die Reliabilität der Summe $Y = Y_1 + Y_2 + \cdots + Y_m$ ergibt sich durch:

$$\alpha = \frac{m}{m-1} \cdot \frac{\sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \mathrm{Kov}(Y_i, Y_j)}{\mathrm{Var}(Y)} \qquad (2\text{-}18)$$

bzw.

$$\alpha = \frac{m}{m-1} \cdot \left(1 - \frac{\sum_{i=1}^{m} \mathrm{Var}(Y_i)}{\mathrm{Var}(Y)} \right) \qquad (2\text{-}19)$$

Bemerkungen:

1. Falls die m Tests parallel sind, so ist Cronbachs α identisch zur Spearman-Brown – Formel (Übung 2-12).
2. Falls die *m* Tests kongenerisch sind, so ergibt Cronbachs α eine untere Grenze der Reliabilität.

Zum besseren Verständnis von Gleichung (2-18) bzw. (2-19) betrachten wird die schematische Kovarianzmatrix von Abb. 2-13. Die Matrix ist in drei Bereiche unterteilt: Die Hauptdiagonale enthält die Varianzen. Dieser Bereich der Matrix wurde daher mit *V* bezeichnet. Die mit *C* bezeichneten Teile der Matrix enthalten die Kovarianzen.

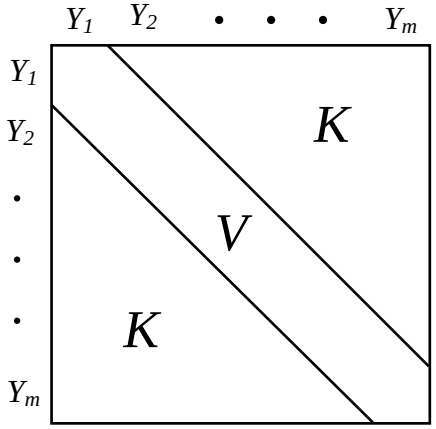


Abb. 2-13: *Schematische Darstellung der Kovarianzmatrix der Variablen $Y_1, Y_2, \ldots, Y_m$: K = Kovarianzen, V = Varianzen.*

Der Zähler von Gleichung (2-18) enthält die Summe aus allen Einträgen in den beiden mit *K* bezeichneten Bereichen. Der Nenner enthält die Summe aus allen Einträgen der Kovarianzmatrix (siehe Anhang, B-2). Dieser Bruch wird dann noch mit $m/(m-1)$ multipliziert.

Gleichung (2-18) lässt sich daher schematisch wie folgt schreiben:

$$\alpha = \frac{m}{m-1} \cdot \frac{K+K}{K+K+V}, \tag{2-20}$$

wobei K und V in Gleichung (2-18) die Summen der Einträge in den zugehörigen Regionen der Kovarianzmatrix bezeichnen.
Die Identität von (2-18) und (2-19) ergibt sich durch folgende Überlegung: Die Varianz $\mathrm{Var}(Y)$ ist die Summe der Einträge aus allen drei Bereichen der Kovarianzmatrix von Abb. 2-13. Zieht man daher von $\mathrm{Var}(Y)$ die Summe der Einträge in der Hauptdiagonale (d.h. innerhalb des mit V bezeichneten Bereichs von Abb. 2-13) ab, so verbleibt die Summe aller Einträge in den mit K bezeichneten Bereichen. Damit ergibt sich (2-19) wie folgt aus (2-18):

$$\begin{aligned}
\alpha &= \frac{m}{m-1} \cdot \frac{K+K}{K+K+V} \\
&= \frac{m}{m-1} \cdot \frac{K+K+V-V}{K+K+V} \\
&= \frac{m}{m-1} \cdot \left(\frac{K+K+V}{K+K+V} - \frac{V}{K+K+V} \right) \\
&= \frac{m}{m-1} \cdot \left(1 - \frac{V}{K+K+V} \right)
\end{aligned} \tag{2-21}$$

2.3.5 Die Berechnung der Reliabilität einer Summe von Testwerten mit Hilfe linearer Strukturgleichungsmodelle

Die Grundidee zur Schätzung der Reliabilität von Summenwerten mit Hilfe linearer Strukturgleichungsmodelle besteht in der Bildung von Phantomvariablen, welche die Summen repräsentieren.

Konzept 2-10: *Phantomvariablen:*

Eine *Phantomvariable* ist eine latente Variable mit den folgenden beiden Eigenschaften:

4. Sie hat keinen Einfluss auf die Schätzung des Modells. Dies bedeutet, dass durch die Einführung der Phantomvariable in das Modell weder die Schätzung der restlichen Modellparameter noch die Modellvorhersagen verändert werden.
5. Die Einführung von Phantomvariablen in das Modell dient ganz bestimmten Zwecken. Die beiden wichtigsten sind:
   (i) Die Spezifikation bestimmter Parameter;
   (ii) Die Schätzung spezifischer Effekte: Dies ist auch der Zweck, den die Phantomvariablen im Folgenden spielen.

Wir demonstrieren nun das Verfahren der Reliabilitätsschätzung für das Modell kongenerischer Tests. Das Verfahren funktioniert ebenso für τ-äquivalente und parallele Tests.

Methode 2-3: *Ermittlung der Reliabilität von Summen kongenerischer, τ-äquivalenter oder paralleler Tests (Miller, 1995, Rykov, 1997).*

Gegeben: Das Modell kongenerischer Tests (Abb. 2-14):

1. Bilde eine Phantomvariable *Y* zur Repräsentation der Summe $Y = Y_1 + Y_2 + \cdots + Y_m$.
2. Verbinde die *m* Testwerte $Y_1, Y_2, \ldots, Y_m$ mit der Phantomvariable *Y* durch Pfeile $Y_i \rightarrow Y$, wobei die zugehörigen Regressionskoeffizienten alle auf den Wert 1 fixiert werden (Siehe Abb. 2-15).

   Bemerkung:
   Da *Y* keinen zugehörigen Fehlerterm besitzt, wird durch die Pfeile die Gleichung $Y = Y_1 + Y_2 + \cdots + Y_m$ realisiert.
3. Die Reliabilität der Summe $Y = Y_1 + Y_2 + \cdots + Y_m$ entspricht der quadrierten Korrelation zwischen η und *Y*.

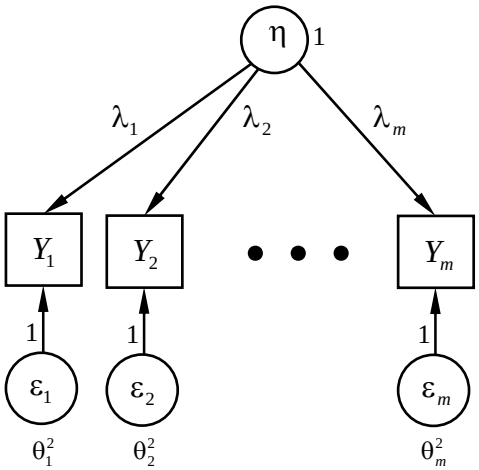


Abb. 2-14: *Das Modell kongenerischer Tests.*

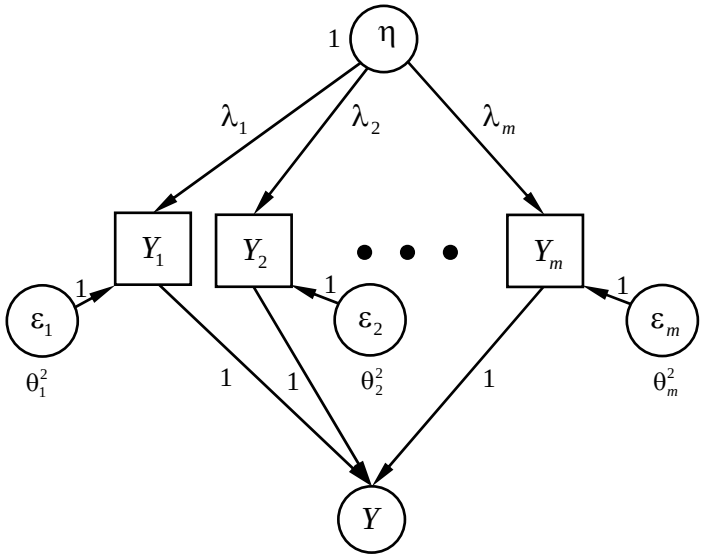


Abb. 2-15: *Das Modell zur Ermittlung der Reliabilität der Summe der Tests im kongenerischen (τ-äquivalenten, parallelen) Modell.*

Bemerkung:
Falls das Modell τ-äquivalent ist, so entspricht die ermittelte quadrierte Korrelation Cronbachs α.

Um die Korrelation zwischen den latenten Variablen η und *Y* im Output zu erhalten, müssen die Optionen:

View/Set / Analysis Properties / Output / All implied moments

sowie:

View/Set / Analysis Properties / Output / Standardized estimates

gesetzt werden.

Bsp.2-7: Schätzung der Reliabilität der Summe der Testwerte im kongenerischen Modell:

Gegeben: Die Kovarianzmatrix von 5 kongenerischen Tests (Tab. 2-2):

	Y_1	Y_2	Y_3	Y_4	Y_5
Y_1	3.62	1.80	0.36	4.32	1.08
Y_2	1.80	2.50	0.20	2.40	0.60
Y_3	0.36	0.20	2.02	0.48	0.12
Y_4	4.32	2.40	0.48	6.84	1.44
Y_5	1.08	0.60	0.12	1.44	3.24

Tab. 2-2: *Kovarianzmatrix von fünf kongenerischen Testitems.*

Gesucht: Die Reliabilität der Summe:

$$Y = Y_1 + Y_2 + Y_3 + Y_4 + Y_5 .$$

Verwendet man Cronbachs α zu Schätzung der Reliabilität, so ergibt sich: $\alpha = .730$.
Der korrekte Wert der Reliabilität der Summe beträgt jedoch: $\mathrm{Rel}(Y) = .822$. Zur Ermittlung der Reliabilität wurde das Modell in Abb. 2-16 verwendet. Die Korrelation zwischen η und Y beträgt .9064. Das *Quadrat dieses Wertes* entspricht der Reliabilität.
Bemerkung Modell von Abb. 2-16:
Im Modell wurde für die Phantomvariable Y ein Residuumsterm ζ mit Varianz $\psi_\zeta^2 = 0$ eingeführt, um eine Warnung durch das Programm zu verhindern. Die Verwendung eines derartigen Residuums mit Varianz 0 hat den gleichen Effekt, wie das Weglassen des Terms.

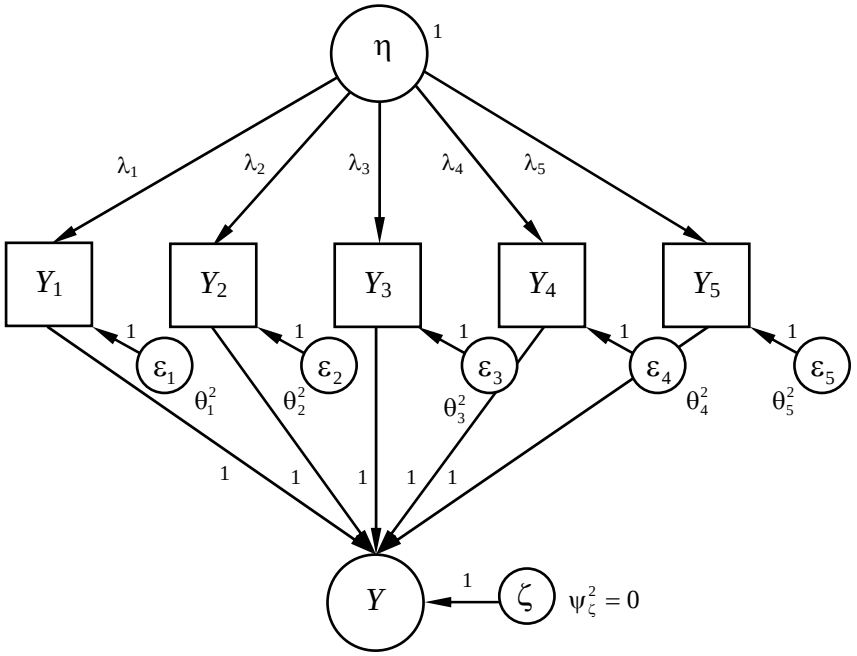


Abb. 2-16: *Modell zur Schätzung der Reliabilität von fünf kongenerischen Testitems.*

Die dargestellte Methode zur Ermittlung der Reliabilität einer Summe von Testwerten kann generalisiert werden, sodass sie auf das allgemeine Testmodell mit mehreren nicht perfekt korrelierten Faktoren anwendbar ist.

Methode 2-4: *Ermittlung der Reliabilität der Summe von Testwerten im klassischen Testmodell.*

Gegeben: Das klassische Testmodell (Abb. 2-3):

1. Bilde eine Phantomvariable *Y* zur Repräsentation der Summe $Y = Y_1 + Y_2 + \cdots + Y_m$ und verbinde die *m* Testwerte mit der Phantomvariable *Y* durch Pfeile $Y_i \rightarrow Y$, mit zugehörigen Regressionskoeffizienten 1.
2. Bilde eine Phantomvariable η zur Repräsentation der gewichteten Summe $\eta = \lambda_1 \cdot \eta_1 + \lambda_2 \cdot \eta_2 + \cdots + \lambda_m \cdot \eta_m$ und verbinde die *m* Faktoren mit η durch Pfeile $\eta_j \rightarrow \eta$, mit zugehörigen Regressionskoeffizienten λ_j (identisch zum unstandardisierten Ladungskoeffizienten, siehe Abb. 2-17).
3. Die Reliabilität der Summe $Y = Y_1 + Y_2 + \cdots + Y_m$ entspricht der quadrierten Korrelation zwischen den beiden Phantomvariablen η und *Y* oder – alternativ – der Varianz von η dividiert durch die Varianz von *Y* (Da die Varianz von η der True-Score Varianz entspricht).

Interpretation:
Die so ermittelte Reliabilität der Summe entspricht der Reliabilität des Indikators *Y* im Modell von Abb. 2-18:

Hierbei gilt:

$Y = Y_1 + Y_2 + \cdots + Y_m$ und $\mathrm{Var}(Y)$ entspricht der Summe aller Einträge in der Kovarianzmatrix der $Y_1, Y_2, \ldots, Y_m$.

$\theta^2 = \theta_1^2 + \theta_2^2 + \cdots + \theta_m^2$, d.h. die Fehlervarianz im Modell von Abb. 2-18 entspricht der Summe der einzelnen Fehlervarianzen im Modell von Abb. 2-17.

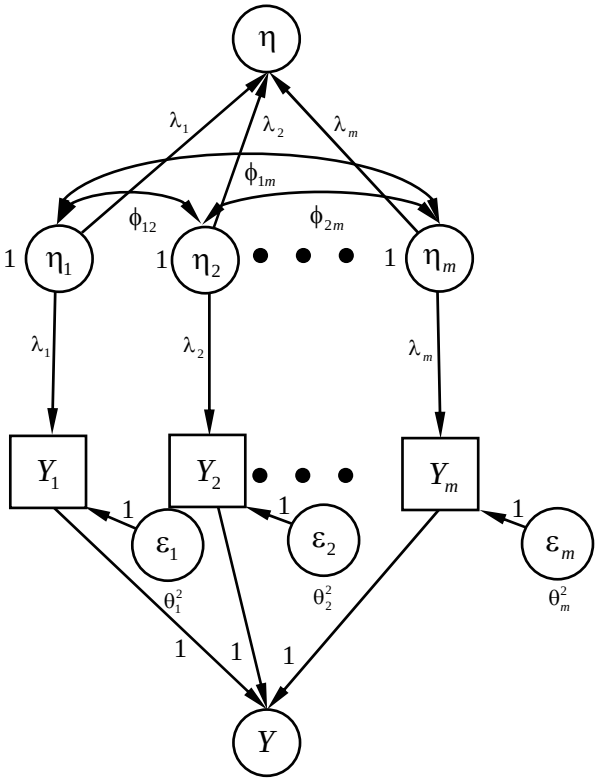


Abb. 2-17: *Das Modell zur Ermittlung der Reliabilität der Summe der Tests im allgemeinen klassischen Testmodell.*

Bsp.2-8: Schätzung der Reliabilität der Summe der Testwerte im klassischen Testmodell:

Gegeben: Die Kovarianzmatrix von Tab. 2-1, Bsp.2-3): X_1 und X_2, sowie Y_1 und Y_2 sind jeweils paarweise parallel aber die vier Tests sind nicht kongenerisch (Hypothese H_1).

Gesucht: Die Reliabilität der Summe:

$$Y = X_1 + X_2 + Y_1 + Y_2 .$$

Verwendet man Cronbachs α zu Schätzung der Reliabilität (basierend auf der vom Modell implizierten Kovarianzmatrix), so ergibt sich: $\alpha = .888$.

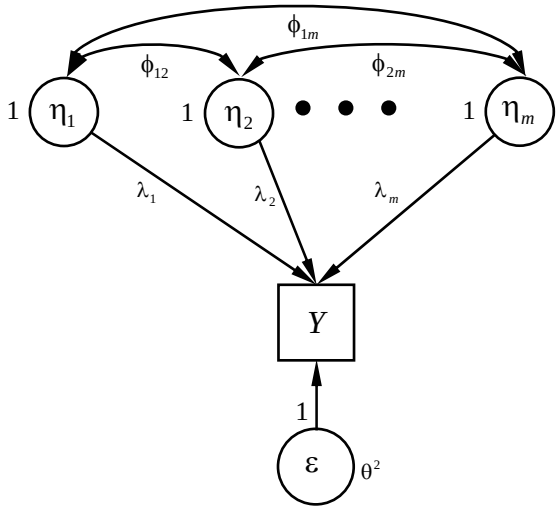


Abb. 2-18: *Das Modell zur Ermittlung der Reliabilität der Summe der Tests im allgemeinen klassischen Testmodell.*

Der korrekte Wert der Reliabilität der Summe beträgt jedoch: $\text{Rel}(Y) = .905$. Zur Ermittlung der Reliabilität wurde das Modell in Abb. 2-19 verwendet. Die Korrelation zwischen η und *Y* beträgt .9513. Das Quadrat dieses Wertes entspricht der Reliabilität.

Alternativ ergibt sich die Reliabiltät durch:

$$\text{Rel}(Y) = \frac{\text{Var}(\eta)}{\text{Var}(Y)} = \frac{994.344}{1098.760} = .905 .$$

Auch in diesem Beispiel wird die Reliabilität durch Cronbachs α unterschätzt.

Die dargestellte Methode zur Berechnung der Reliabilität einer Summe von Testwerten mit Hilfe linearer Strukturgleichungsmodelle lässt sich weiter verallgemeinern. Mit Hilfe dieser verallgemeinerten Methode lässt sich die Reliabilität jeder (gewichteten) Summe von Testwerten im allgemeinen faktorenanalytischen Modell berechnen.

Bemerkung:

Das allgemeine faktorenanalytische Modell wird in Abschnitt **xxxx** im Detail besprochen. Für die gegenwärtige Darstellung reicht es zu wissen, dass es sich um ein lineares Messmodell mit einer Anzahl latenter Faktoren handelt.

Die Kovarianzstruktur der latenten Konstrukte ist nicht festgelegt und jedes Konstrukt kann auf jede Messung einwirken.

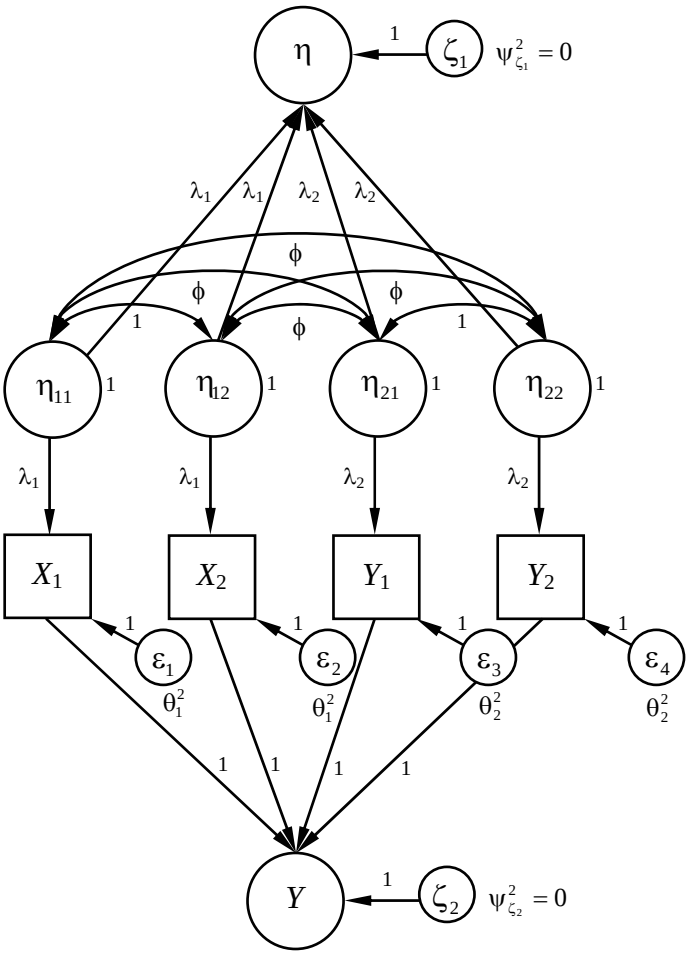


Abb. 2-19: *Modell zur Schätzung der Reliabilität von vier nicht kongenerischen Tests.*

Methode 2-5: *Ermittlung der Reliabilität der gewichteten Summe von Testwerten im allgemeinen faktorenanalytischen Modell.*

Gegeben:

Das allgemeine faktorenanalytische Modell mit *n* latenten Variablen $\eta_1, \eta_2, \ldots, \eta_n$ und *m* Tests $Y_1, Y_2, \ldots, Y_m$:

1. Bilde eine Phantomvariable Y zur Repräsentation der gewichteten Summe $Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + \cdots + w_m \cdot Y_m$ und verbinde die *m* Testwerte mit der Phantomvariable Y durch Pfeile $Y_i \rightarrow Y$, mit zugehörigen Regressionskoeffizienten w_i.

2. Bilde eine Phantomvariable η, deren Varianz die gesamte True-Score Varianz von *Y* repräsentiert und weitere Phantomvariablen ξ_{ij}. Letztere dienen dazu, im Modell mehrere Verbindungen zwischen den η_j und η herzustellen(siehe Abb. 2-20).
3. Verbinde alle ξ_{ij} mit η durch Pfeile $\xi_{ij} \rightarrow \eta$, mit Pfadkoeffizienten w_i (Abb. 2-20).
4. Verbinde die latente Variable η_j mit allen zugehörigen ξ_{ij} (d.h. mit jenen ξ, welche denselben Index *j* an zweiter Stelle aufweisen) durch einen Pfeil $\eta_j \rightarrow \xi_{ij}$, mit zugehörigem Pfadkoeffizienen λ_{ij} (d.h. mit dem Ladungskoeffizienten der Ladung von Variable Y_i auf η_j).
5. Führe Schritt 4 für alle η_j $(j = 1,\ldots,n)$ durch.
6. Die Reliabilität der Summe $Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + \cdots + w_m \cdot Y_m$ entspricht der quadrierten Korrelation zwischen den beiden Phantomvariablen η und Y, bzw. – alternativ – der Varianz von η dividiert durch die Varianz von Y (Da die Varianz von η der True-Score Varianz von *Y* entspricht).

Abb. 2-20 zeigt das resultierende lineare Strukturgleichungsmodell.

Bsp.2-9: Ermittlung der Reliabilität der Summe der Testwerte im allgemeinen faktorenanalytischen Modell:

Gegeben: Das Faktorenmodell von Abb. 2-21 zur Modellierung der Kovarianz zwischen den 5 Tests Y_1-Y_5.

Gesucht: Die Reliabilität der Summe der fünf Messungen.

Cronbachs α ergibt einen Wert von 0.5. Die korrekte Reliabilität beträgt 0.588.

Letzter ergibt sich entweder durch Berechnung der quadrierten Korrelation zwischen den Phantomvariablen *Y* und η, oder durch Division von $\mathrm{Var}(\eta)$ durch $\mathrm{Var}(Y)$.

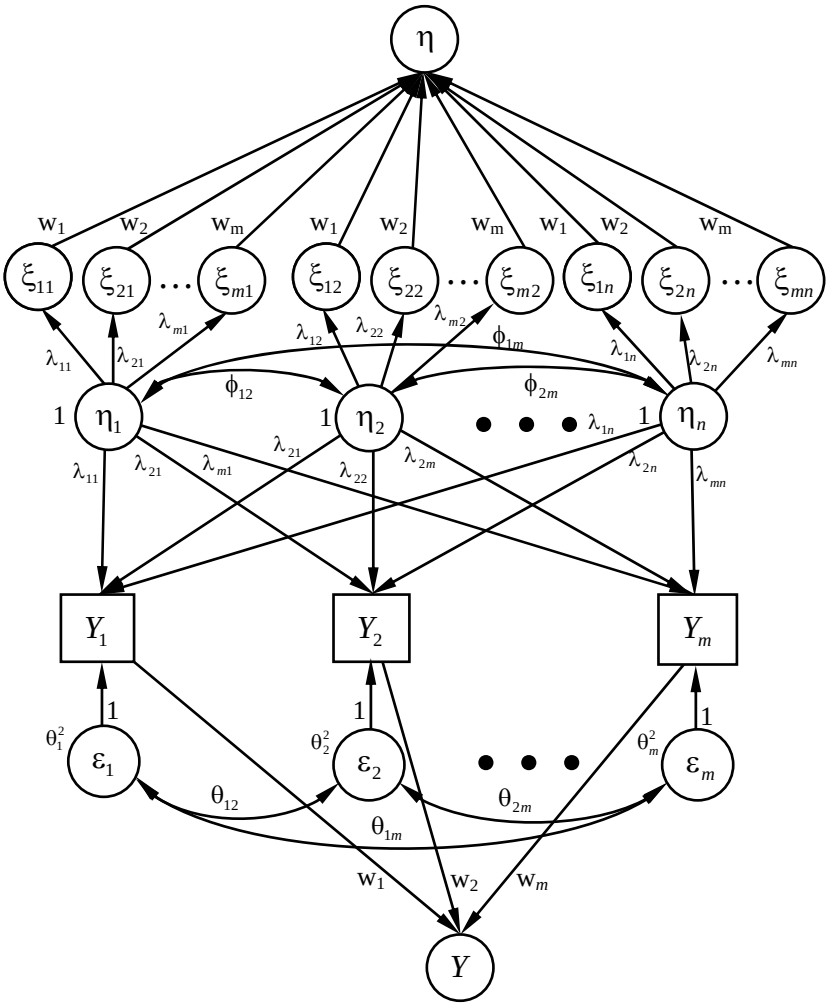


Abb. 2-20: *Das Modell zur Ermittlung der Reliabilität der Summe der Tests im allgemeinen faktorenanalytischen Modell.*

Methode 2-5 zur Ermittlung einer unverzerrten Schätzung der Reliabilität einer (gewichteten) Summe von Messungen im allgemeinen faktorenanalytischen Modell ist für komplexe Modelle mit vielen latenten Variablen und Messungen ungeeignet (da zu aufwendig). In diesem Fall ist es sinnvoll, die relevanten Varianzen mit Hilfe einfacher Matrizenmultiplikationen zu berechnen, wobei die vom Modell gegebenen Schätzungen der Kovarianzmatrix der latenten Variablen sowie der Ladungen verwendet werden.

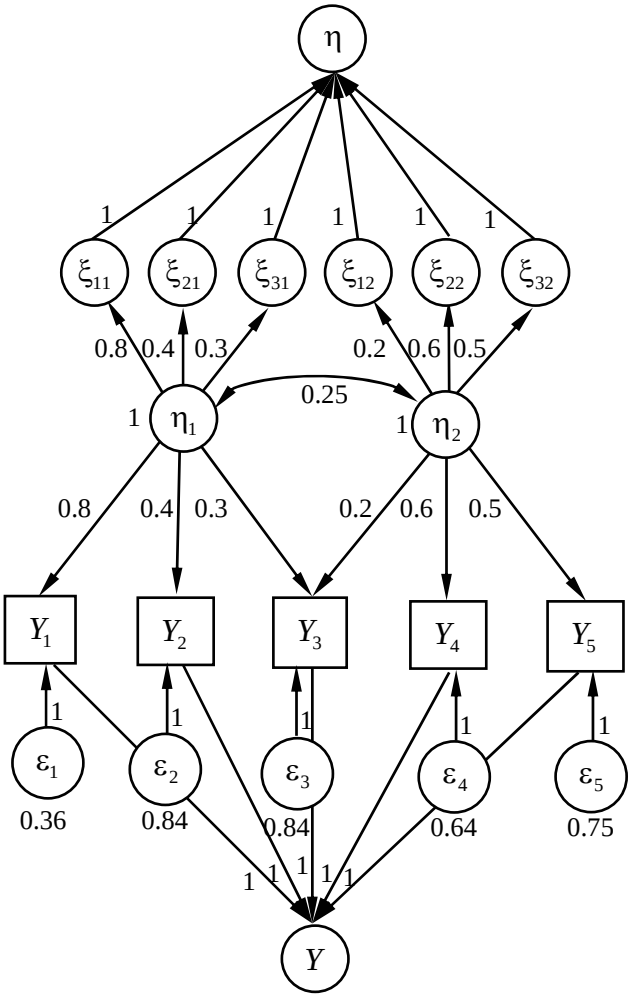


Abb. 2-21: *Das Modell zur Ermittlung der Reliabilität der Summe von fünf Tests.*

Methode 2-6: *Ermittlung der Reliabilität der gewichteten Summe von Testwerten im allgemeinen faktorenanalytischen Modell mittels Matrizen.*

Gegeben:

Das allgemeine faktorenanalytische Modell mit *n* latenten Variablen $\eta_1, \eta_2, \ldots, \eta_n$ und *m* Tests $Y_1, Y_2, \ldots, Y_m$:

Gesucht:

Die Reliabilität der gewichtete Summe:

$$Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + \cdots + w_m \cdot Y_m$$

1. Schätze das Modell und ermittle die drei folgenden Grössen:
   (i) Die ($n \times n$) - Kovarianzmatrix der latenten Konstrukte:

$$\boldsymbol{\Phi} = \begin{array}{c} \\ \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{array} \begin{array}{c} \begin{array}{cccc} \eta_1 & \eta_2 & \cdots & \eta_n \end{array} \\ \begin{bmatrix} \phi_1^2 & \phi_{12} & \cdots & \phi_{1n} \\ \phi_{21} & \phi_2^2 & \cdots & \phi_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n1} & \phi_{n2} & \cdots & \phi_n^2 \end{bmatrix} \end{array},$$

   (ii) Die (n×m) - Matrix der Ladungen:

$$\boldsymbol{\Lambda} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{array} \begin{array}{c} \begin{array}{cccc} \eta_1 & \eta_2 & \cdots & \eta_n \end{array} \\ \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \cdots & \lambda_{mn} \end{bmatrix} \end{array},$$

   (iii) Die durch das Modell vorhergesagte (m×m) - Kovarianzmatrix der Messungen:

$$\hat{\boldsymbol{\Sigma}} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & \cdots & Y_m \end{array} \\ \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1m} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots & \hat{\sigma}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{m1} & \hat{\sigma}_{m2} & \cdots & \hat{\sigma}_m^2 \end{bmatrix} \end{array}$$

2. Berechne die True-Score Varianz (Konzept 2-7) von *Y*:

$$\mathrm{Var}_{\eta}(Y) = \mathbf{w}^{\mathbf{T}} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\Lambda}^{\mathbf{T}} \cdot \mathbf{w} \qquad (2\text{-}22)$$

   Hierbei gilt:
   $\mathbf{w}^{\mathbf{T}} = \begin{bmatrix} w_1 & w_2 & \cdots & w_m \end{bmatrix}$ ist ein Zeilenvektor der Koeffizienten der gewichteten Summe *Y:*
   $Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + \cdots + w_m \cdot Y_m$.
   Das Symbol « $^{\mathbf{T}}$ » repräsentiert die Operation des Transponierens einer Matrix bzw. eines Vektors (Vertauschen des Inhalts der Zeilen und Spalten).
3. Berechne die Gesamtvarianz von *Y*:

$$\mathrm{Var}(Y) = \mathbf{w}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}} \cdot \mathbf{w}$$

4. Die Reliabilität von Y entspricht dem Quotienten der beiden Varianzen:

$$\mathrm{Rel}(Y) = \frac{\mathrm{Var}_{\eta}(Y)}{\mathrm{Var}(Y)}$$

Bsp.2-10 illustriert die Methode.

Bsp.2-10: Ermittlung der Reliabilität der Summe der Testwerte im allgemeinen faktorenanalytischen Modell mit Hilfe von Matrizen (Fortsetzung von Bsp.2-9):

Gegeben: Das Faktorenmodell von Abb. 2-21 zur Modellierung der Kovarianz zwischen den 5 Tests Y_1-Y_5.

Gesucht: Die Reliabilität der Summe der fünf Messungen.

Die einzelnen Matrizen sehen wie folgt aus (siehe Abb. 2-21):

$$\mathbf{\Phi} = \begin{array}{c} \\ \eta_1 \\ \eta_2 \end{array} \begin{array}{c} \begin{array}{cc} \eta_1 & \eta_2 \end{array} \\ \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix} \end{array}$$

$$\mathbf{\Lambda} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{array} \begin{array}{c} \begin{array}{cc} \eta_1 & \eta_2 \end{array} \\ \begin{bmatrix} 0.8 & 0.0 \\ 0.4 & 0.0 \\ 0.3 & 0.2 \\ 0.0 & 0.6 \\ 0.0 & 0.5 \end{bmatrix} \end{array}$$

$$\hat{\mathbf{\Sigma}} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{array} \begin{array}{c} \begin{array}{ccccc} Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \end{array} \\ \begin{bmatrix} 1.000 & 0.320 & 0.220 & 0.120 & 0.100 \\ 0.320 & 1.000 & 0.110 & 0.060 & 0.050 \\ 0.220 & 0.110 & 1.000 & 0.210 & 0.175 \\ 0.120 & 0.060 & 0.210 & 1.000 & 0.300 \\ 0.100 & 0.050 & 0.175 & 0.300 & 1.000 \end{bmatrix} \end{array}$$

Der Gewichtsvektor ist:

$$\mathbf{w} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \text{ bzw. } \mathbf{w}^{\mathrm{T}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Die True-Score Varianz von *Y* beträgt:

$$\begin{aligned}
&\mathrm{Var}_{\eta}(Y) \\
&= \mathbf{w}^{\mathbf{T}} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\Lambda}^{\mathbf{T}} \cdot \mathbf{w} \\
&= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0.8 & 0.0 \\ 0.4 & 0.0 \\ 0.3 & 0.2 \\ 0.0 & 0.6 \\ 0.0 & 0.5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0.8 & 0.4 & 0.3 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.2 & 0.6 & 0.5 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \\
&= 4.9
\end{aligned}$$

Die Gesamtvarianz von *Y* betägt:

$$\begin{aligned}
&\mathrm{Var}(Y) = \mathbf{w}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}} \cdot \mathbf{w} \\
&= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1.000 & 0.320 & 0.220 & 0.120 & 0.100 \\ 0.320 & 1.000 & 0.110 & 0.060 & 0.050 \\ 0.220 & 0.110 & 1.000 & 0.210 & 0.175 \\ 0.120 & 0.060 & 0.210 & 1.000 & 0.300 \\ 0.100 & 0.050 & 0.175 & 0.300 & 1.000 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \\
&= 8.33
\end{aligned}$$

Die Reliabilität von Y ist daher:

$$\begin{aligned}
\mathrm{Rel}(Y) &= \frac{\mathrm{Var}_{\eta}(Y)}{\mathrm{Var}(Y)} \\
&= \frac{4.9}{8.33} \\
&= 0.588
\end{aligned}$$

Das Ergebnis ist identisch zu jenem, welches mit Hilfe von Methode 2-5 ermittelt wurde (vergleiche Bsp.2-9*)*.

Praktischen Durchführung der Berechnung von Matrizenoperationen:

Zur praktischen Durchführung der Berechnung benötigt man ein Programm, das einfache Matrizenoperationen durchführen kann (z.B. SPSS, R, Matlab).

Das Programm Excel besitzt Matrizenfunktionen zur Multiplikation [Funktion: MMULT()] und zum Transponieren [Funktion TRANSPOSE()] von Matrizen.

Hinweis: Matrizenberechnungsausdrücke müssen mit der Tastenkombination *Ctrl-Shift-Enter* abgeschlossen werden (für Details siehe z.B. Macho, 2002).

Nach dieser ausführlichen Behandlung der Schätzung der Reliabilität von Summenwerten wenden wir uns einigen Problemen im Zusammenhang mit der Verwendung von Cronbachs α zur Schätzung der Reliabilität zu.

2.3.6 Über- und Unterschätzung der wahren Reliabilität durch Cronbachs α

In den zuvor gezeigten Beispielen wurde die wahre Reliabilität durch Cronbachs α jeweils unterschätzt. In diesem Falle führt die Verwendung von Cronbachs α zur Abschwächungskorrektur immer zu einer Überschätzung der korrigierten Grösse (vergleiche Abschnitt 2.5).
Die Verwendung von Cronbachs α kann jedoch nicht nur zu einer Über- sondern auch zu einer Unterschätzung der wahren Reliabilität führen.

Prinzip 2-3: *Über- und Unterschätzung der Reliabilität durch Cronbachs α*

Cronbachs α kann die Reliabilität einer Summe von Tests sowohl über- als auch unterschätzen. Hierbei gilt:

5. Bei Vorliegen kongenerischer Test (die nicht τ-äquivalent sind) wird die Reliabilität unterschätzt (Siehe Bsp.2-7).
6. Eine Überschätzung kann der Reliabilität kann bei Vorliegen korrelierten Fehler auftreten. Die Überschätzung ergibt sich dadurch, dass Cronbachs α die erhöhte Kovarianz zwischen den beobachteten Werten irrtümlicherweise als durch die latente Variable verursacht betrachtet (Siehe Bsp.2-11).

Bsp.2-11: Unterschätzung der Reliabilität durch α:
Gegeben: Die Kovarianzmatrix von 5 Tests (Tab. 2-3):

	Y_1	Y_2	Y_3	Y_4	Y_5
Y_1	1.00	0.41	0.28	0.35	0.35
Y_2	0.41	1.00	0.47	0.15	0.49
Y_3	0.28	0.47	1.00	0.20	0.20
Y_4	0.35	0.15	0.20	1.00	0.25
Y_5	0.35	0.49	0.20	0.25	1.00

Tab. 2-3: *Kovarianzmatrix von fünf Testitems.*

Diese Kovarianzmatrix wurde mit Hilfe des Modells von Abb. 2-22 erzeugt (man beachte die Kovarianzen zwischen Fehlertermen).
Die Reliabilität der Summe der 5 Tests ist gemäss Cronbachs $\alpha = .697$.

Der korrekte Wert der Reliabilität der Summe beträgt jedoch: $\mathrm{Rel}(Y) = .510$. Zur Ermittlung der korrekten Reliabilität wurde Methode 2-3 verwendet.

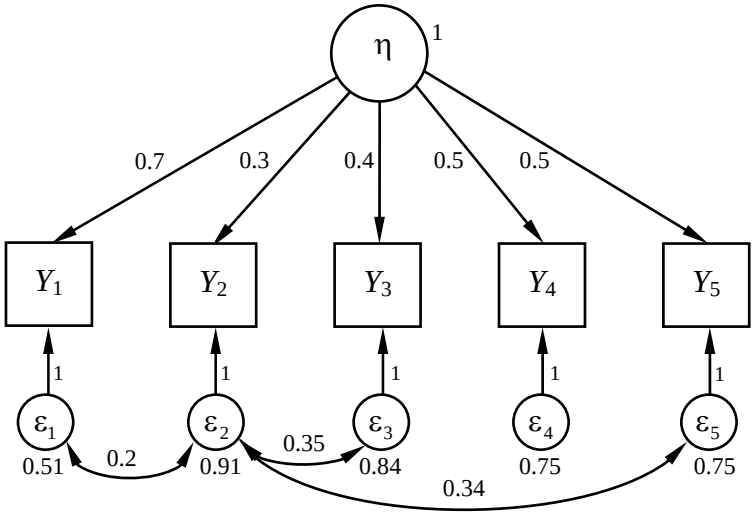


Abb. 2-22 *Modell zur Erzeugung der Daten von Tab. 2-3.*

Bemerkung zur Über- und Unterschätzung der wahren Reliabilität durch Cronbachs α in der Praxis:
Ich vermute, dass die publizierten Werte von Cronbachs α die wahre Reliabilität meist *überschätzen*. Dies lässt sich wie folgt begründen:
Im Allgemeinen besteht ein Itempaket (Summe von Einzeltests), für welches α berechnet wird, aus mindestens 10-15 Tests. Es ist kaum anzunehmen, dass die Kovarianz zwischen diesen Tests zufrieden stellend durch das zu messende latente Zielkonstrukt erklärt werden kann. Es liegt daher eher eine Situation vor, wie sie in Abb. 2-22 dargestellt ist, wonach das latente Konstrukt nur einen Teil der Kovarianz zwischen den Tests erklärt und der Rest auf andere ungeklärte Ursachen zurückzuführen ist.
Die grosse Schwäche von Cronbachs α (und vergleichbarer Koeffizienten) liegt letztendlich darin, dass nur die beobachteten Kovarianzen in die Berechnung einfliessen, während die zugrunde liegenden Ursachen für das Vorhandensein der Kovarianzen unberücksichtigt bleiben.

2.3.7 Eine mögliche Fehlinterpretation: Cronbachs α als Homogenitätskoeffizient

Ein möglicher Fehler hinsichtlich der Verwendung von Cronbachs α besteht darin, den Koeffizienten als Maß für die *Homogenität,* d.h. *Eindimensionalität* der Tests zu betrachten. Dies kann auf einfache

Weise demonstriert werden, indem gezeigt wird, dass α hoch sein kann, obwohl die Test nicht eindimensional sind.

Bsp.2-12: Cronbachs α und Homogenität (Eindimensionalität) von Tests (Green, Lissitz & Mulaik, 1977).

Gegeben: Das Modell von Abb. 2-23.

Die 10 Indikatoren des Modells werden durch insgesamt 5 unkorrelierte (fehlende Kovarianzbögen) Faktoren beeinflusst. Jeder Faktor beeinflusst. Jeder Faktor hat einen Einfluss auf 4 Indikatoren, wobei jedes Paar von Faktoren genau einen gemeinsamen Indikator beeinflusst (z.B. τ_1 und τ_5 beeinflussen beide Y_4 und dies ist der einzige gemeinsame Indikator). Die Tests sind daher in keiner Weise eindimensional.

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
Y_1	1	0.49	0.49	0.49	0.49	0.49	0.49	0	0	0
Y_2	0.49	1	0.49	0.49	0.49	0	0	0.49	0.49	0
Y_3	0.49	0.49	1	0.49	0	0.49	0	0.49	0	0.49
Y_4	0.49	0.49	0.49	1	0	0	0.49	0	0.49	0.49
Y_5	0.49	0.49	0	0	1	0.49	0.49	0.49	0.49	0
Y_6	0.49	0	0.49	0	0.49	1	0.49	0.49	0	0.49
Y_7	0.49	0	0	0.49	0.49	0.49	1	0	0.49	0.49
Y_8	0	0.49	0.49	0	0.49	0.49	0	1	0.49	0.49
Y_9	0	0.49	0	0.49	0.49	0	0.49	0.49	1	0.49
Y_{10}	0	0	0.49	0.49	0	0.49	0.49	0.49	0.49	1

Tab. 2-4: *Die vom Modell in Abb. 2-23 implizierte Kovarianzmatrix.*

Tab. 2-4 zeigt die vom Modell implizierte Kovarianzmatrix. Der aufgrund der Matrix berechnete Koeffizient ist relativ hoch: $\alpha = .829$.

Cronbachs α ist daher ungeeignet als Maß der Eindimensionalität der Tests. Letzteres prüft man durch Testung des kongenerischen Modells: Falls das kongenerische Modell die Daten gut fittet, kann man davon ausgehen, dass die Tests eindimensional sind.

Der fehlerhafte Schluss von einem hohen Wert von α auf Eindimensionalität beruht letztendlich darauf, dass von hohen beobachteten Korrelationen zwischen den Items auf Eindimensionalität geschlossen wird. Nun steigt zwar α monoton mit der Höhe der beobachteten Korrelationen zwischen den Items an (vorausgesetzt die Varianzen der Items bleiben konstant). Dies ergibt sich unmittelbar aus Gleichung (2-18), da höhere Korrelationen zu höheren Kovarianzen führen, was un-

mittelbar in einem höheren Wert von α resultiert. Hohe Korrelationen zwischen den Items implizieren jedoch nicht zwingend Eindimensionalität. Hohe Korrelationen zwischen den Messungen können nämlich auch dann vorliegen, wenn die Items durch mehrere latente Konstrukte beeinflusst werden und diese hoch korreliert sind, oder – wie in Bsp.2-12 – dass mehrere latente Variablen jeweils mehrere Indikatoren beeinflussen. Dadurch ergeben sich hohe Korrelationen zwischen Gruppen von Variablen. Der hohe Wert von Cronbachs α in Bsp.2-12 ergibt sich auch aufgrund der grossen Anzahl von Items (Bei zwei korrelierten Items beträgt α nur noch .658). Man beachte jedoch, dass α = .829 die wahre Reliabilität der einzelnen Items (= .980) deutlich unterschätzt.

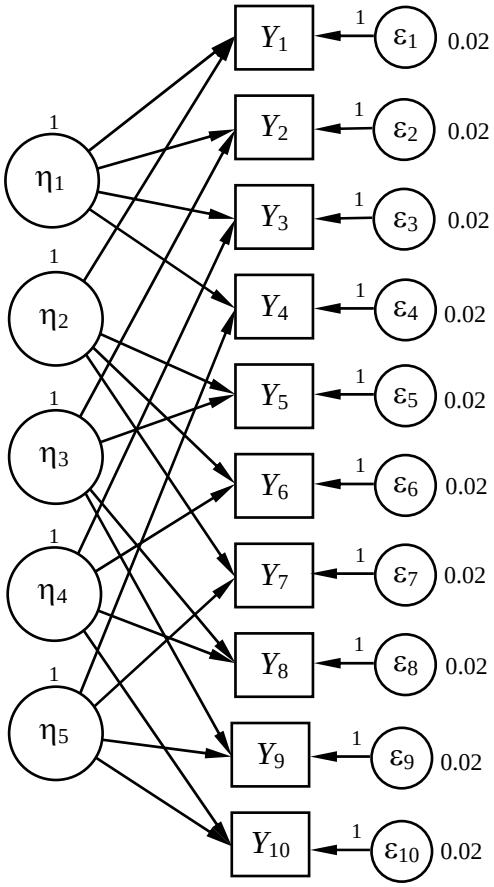


Abb. 2-23: *Modell zur Demonstration der Ungeeignetheit von α als Homogenitätsmaß: Jeder Faktor beeinflusst jeweils vier Indikatoren und zwei Faktoren beeinflussen immer nur einen gemeinsamen Indikator. Die Ladungen sind alle identisch: λ=0.7. Die Faktoren sind unkorreliert.*

2.3.8 Fehlende Monotonieeigenschaften der Reliabilität der Summe von Testwerten

Die Bildung einfacher (ungewichteter) Summen von Testwerte scheint in der täglichen Praxis der Analyse von Testwerten eine gängige Praxis zu sein. Dies ist insofern erstaunlich, als bereits aus der Elementarstatistik bekannt ist, dass das »Poolen« von Mittelwerten und Varianzen nicht ungewichtet erfolgen sollte.
Verwendet man die einfache Summe von Testwerten, so erhält man nicht die maximale Reliabilität. Zusätzlich garantiert eine einfache Summe nicht die Erhaltung von drei wichtigen Anforderungen die Monotonie der Reliabilität betreffend.

Prinzip 2-4: *Monotonieforderungen an die Reliabilität:*

1. Fügt man zu einer bestehenden Menge von Tests reliable Testitems hinzu, so sollte die Reliabilität des resultierenden Summenwertes erhöht werden.
2. Ersetzt man in einer Menge von Tests einen Test durch einen anderen von grösserer Reliabilität, so sollte die Reliabilität des resultierenden Summenwertes jene des alten übersteigen.
3. Wird die Korrelation zwischen den Konstruktwerten zweier Tests verringert, so sollte sich die Reliabilität der Summe der beiden Tests verringern.

Im Folgenden wird – in Anlehnung an Li, Rosenthal und Rubin (1996) – demonstriert, dass einfache Summen von Testitems alle drei Monotonieforderungen verletzen können.

Bsp.2-13: Reliabilität der Summe von Items steigt nicht monoton mit der Testlänge.

Gegeben: Das τ-äquivalente Modell von Abb. 2-24.

Die Testitems bestehen aus zwei Gruppen:

Items Y_1 und Y_2 besitzen eine hohe Reliabilität: $\text{Rel}(Y_i) = .6$.

Die restlichen Items $Z_1,\ldots,Z_5$ weisen hingegen eine geringe Reliabilität: $\text{Rel}(Y_i) = .1$ auf.

Gemäss Spearman-Brown – Formel ergibt sich für die Summe $Y = Y_1 + Y_2$ die Reliabilität:

$$\text{Rel}(Y) = \frac{m \cdot \rho}{1 + (m-1) \cdot \rho}$$

$$= \frac{2 \cdot .6}{1 + (2-1) \cdot .6} = .75 \quad .$$

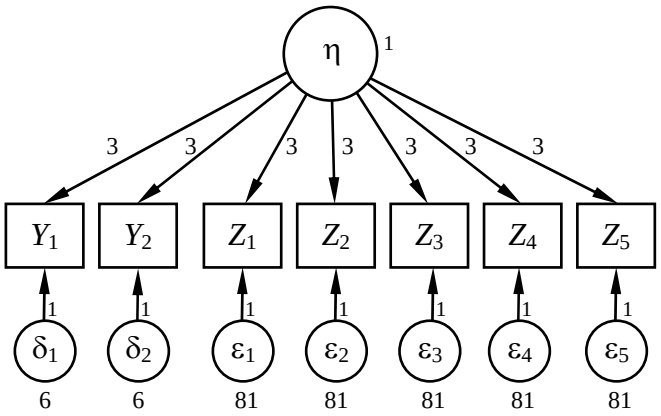


Abb. 2-24: *Modell zur Demonstration fehlender Monotonie der Reliabilität von Summen mit der Testlänge.*

Gemäss Cronbachs α ergibt sich für die Summe aus allen Items $Z = Y_1 + Y_2 + Z_1 + Z_2 + Z_3 + Z_4 + Z_5$ die Reliabilität:

$$\alpha = \text{Rel}(Z) = .514 .$$

Die Hinzufügung der 5 Items $Z_1, Z_2, \ldots, Z_5$ zu Y_1 und Y_2 führt also zu einer Verringerung der Reliabilität der Gesamtsumme.

Das nächste Beispiel demonstriert die Verletzung der zweiten (und auch der ersten) Anforderung an die Monotonie der Reliabilität.

Bsp.2-14: Reliabilität der Summe von Items steigt nicht notwendig mit der Reliabilität der einzelnen Items.

Gegeben: Die beiden Modelle von Abb. 2-25.

Das linke Modell umfasst zwei Testitems mit den Reliabilitäten .6 und .1. Es ergibt sich: $\alpha = \text{Rel}(Y + Z) = .29$.

Im rechten Modell wurde das Item mit Reliabilität .6 durch eines mit Reliabilität .65 ersetzt. Die Varianz des Items wurde jedoch verringert. Für diese Konstellation ergibt sich:
$\text{Rel}(Y + Z) = .23$, $\alpha = .22$.

Aufgrund der verringerten Varianz des Items mit höherer Reliabilität im rechten Modell, ergibt sich eine Verringerung des Gewichts dieses Items relativ zum anderen. Dadurch wird die Reliabilität der Summe abgesenkt.

Man beachte auch, dass in diesem Beispiel das Hinzufügen von Item *Z* zu Item *Y* zur Absenkung der Reliabilität der Summe auf die Hälfte der Reliabilität von *Y* führt.

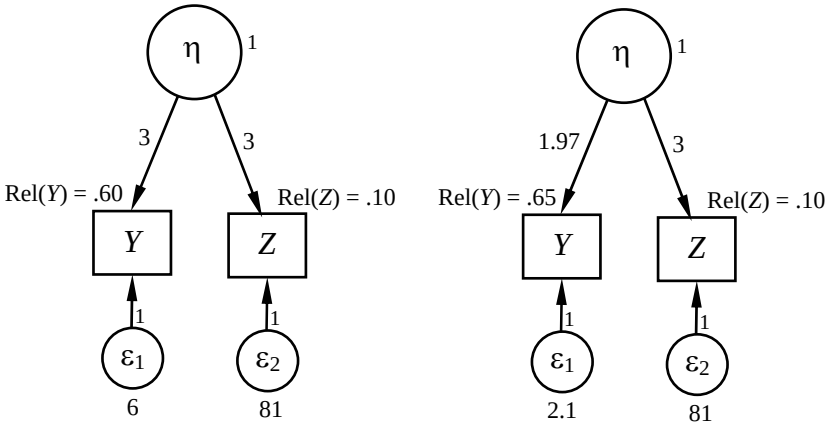


Abb. 2-25: *Zwei Modelle zur Demonstration fehlender Monotonie der Reliabilität der Summe von Testwerten mit der Reliabilität der einzelnen Tests.*

Das letzte Beispiel illustriert die Verletzung der dritten Monotonieforderung (und auch der ersten).

Bsp.2-15: Reliabilität der Summe von Items steigt nicht notwendig mit der Korrelation der latenten Faktoren.

Gegeben: Die beiden Modelle von Abb. 2-26.

Das linke Modell ist identisch zum linken Modell in Abb. 2-25. es gilt daher wiederum: $\alpha = \mathrm{Rel}(Y+Z) = .29$.

Im rechten Modell wurde die Varianz des Items mit Reliabilität .6 erhöht und zugleich die Korrelation zwischen den latenten Variablen auf .95 verringert, unter Beibehaltung der Reliabilität. Für die Summe ergibt sich die Reliabilität:
$\mathrm{Rel}(Y+Z) = .34$, $\alpha = .32$.

Durch die Erhöhung der Varianz des Items mit höherer Reliabilität im rechten Modell, erhält dieses Item relativ zum anderen mehr Gewicht. Dies führt dazu, dass die verringerte Korrelation zwischen den latenten Variablen mehr als ausgeglichen wird.

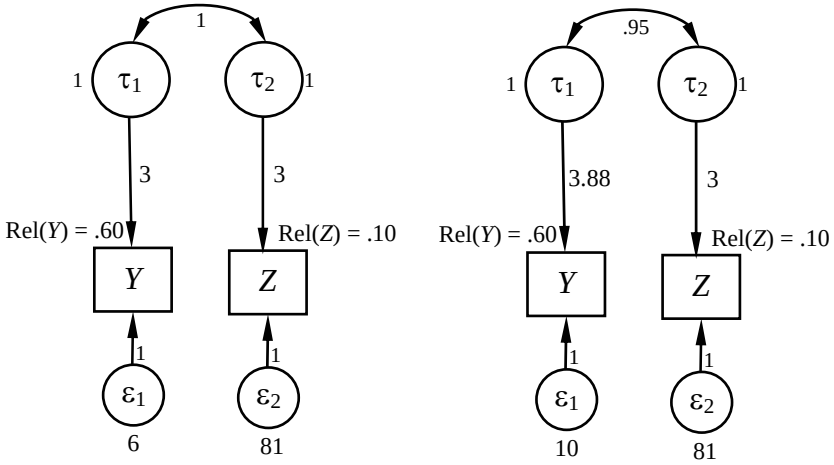


Abb. 2-26: *Zwei Modelle zur Demonstration fehlender Monotonie der Reliabilität der Summe von Testwerten mit der Korrelation der latenten Faktoren.*

Bsp.2-13, Bsp.2-14 und Bsp.2-15 demonstrieren auf eindringliche Weise die Nachteile der Bildung ungewichteter Summen von Testwerten. Aufgrund dieser Probleme stellen sich daher zwei Fragen.

Fragen:

1. *Gibt es ein optimales Maß für die Reliabilität, welches die drei Monotonieeigenschaften erfüllt?*
2. *Wie müssen die einzelnen Testitems gewichtet werden, um die Reliabilität der gewichteten Summe zu maximieren?*

Die Beantwortung dieser beiden Fragen bildet den Gegenstand des nächsten Abschnitts.

2.3.9 Maximale Reliabilität und optimale Gewichtung der Tests

Im Folgenden werden zuerst verschiedene Methoden zur Ermittlung der maximalen Reliabilität, sowie der optimalen Gewichtung der Testitems dargestellt. Dem folgt die Demonstration, dass die maximale Reliabilität die drei oben präsentierten Anforderungen hinsichtlich der Monotonie nicht verletzt.

Methode 2-7: *Die maximale Reliabilität der optimal gewichteten Summe kongenerischer (τ-äquivalenter, bzw. paralleler) Tests und die zugehörigen optimalen Gewichte*

Gegeben: Kongenerisches Testmodell (siehe z.B. Abb. 2-5) für m Tests: $Y_1, Y_2, \ldots, Y_m$.

1. Die *maximale Reliabilität* der Komposition der Tests ergibt sich durch:

$$\mathrm{Rel}_{\max}(Y) = \frac{\frac{\lambda_1^2}{1-\lambda_1^2} + \frac{\lambda_2^2}{1-\lambda_2^2} + \cdots + \frac{\lambda_m^2}{1-\lambda_m^2}}{1 + \frac{\lambda_1^2}{1-\lambda_1^2} + \frac{\lambda_2^2}{1-\lambda_2^2} + \cdots + \frac{\lambda_m^2}{1-\lambda_m^2}} \qquad (2\text{-}23)$$

Hierbei gilt:
Die λ_i symbolisieren die *standardisierten* Ladungskoeffizienten.

2. Die *optimalen* Gewichte $w_1, w_2, \ldots, w_m$ zur Bildung der Linearkombination $Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + \cdots + w_m \cdot Y_m$ mit maximaler Reliabilität sind gegeben durch:

$$w_i = \frac{\lambda_i}{\theta_i^2}\,. \qquad (2\text{-}24)$$

Hierbei gilt:
λ_i symbolsiert den *unstandardisierten* Ladungskoeffizienten von Test Y_i und θ_i^2 repräsentiert die Varianz des zugehörigen Fehlers.

Bsp.2-16: Schätzung der maximalen Reliabilität der Summe der Testwerte im kongenerischen Modell:

Gegeben: Die Kovarianzmatrix von 5 kongenerischen Tests (Tab. 2-2):

Gesucht: Die maximale Reliabiltät der optimal gewichteten Summe der Test und die zugehörigen Gewichte.

Tab. 2-5 zeigt die relevanten Grössen zur Berechnung der maximalen Reliabilität und der optimalen Gewichte:

Test	$\lambda_i^{\text{standardisiert}}$	$\frac{\lambda_i^2}{1-\lambda_i^2}$	θ_i^2	$\lambda_i^{\text{unstandardisiert}}$	w_i
Y_1	0.946	8.526	0.38	1.8	4.737
Y_2	0.632	0.667	1.50	1.0	0.667
Y_3	0.141	0.020	1.98	0.2	0.101
Y_4	0.918	5.333	1.08	2.4	2.222
Y_5	0.333	0.125	2.88	0.6	0.208

Tab. 2-5: *Daten zur Berechnung der maximalen Reliabiltät und der optimalen Gewichte.*

Spalte 2 enthält die standardisierten Ladungen

In Spalte 3 befinden sich die Summanden, welche in die Berechnung eingehen.
Spalte 4 enthält die Varianzen der Fehler und Spalte 5 die unstandardisierten Ladungskoeffizienten. Die Gewichte in Spalte 6 ergeben sich durch Division der Einträge in Spalte 5 und 4.

Die maximale Reliabiltät ergibt sich daher durch:

$$\begin{aligned}\mathrm{Rel}_{\max}(Y) &= \frac{\frac{\lambda_1^2}{1-\lambda_1^2}+\frac{\lambda_2^2}{1-\lambda_2^2}+\cdots+\frac{\lambda_m^2}{1-\lambda_m^2}}{1+\frac{\lambda_1^2}{1-\lambda_1^2}+\frac{\lambda_2^2}{1-\lambda_2^2}+\cdots+\frac{\lambda_m^2}{1-\lambda_m^2}} \\ &= \frac{8.526+0.667+0.020+5.333+0.125}{1+(8.526+0.667+0.020+5.333+0.125)} \\ &= .936\end{aligned}$$

Das optimale Gewicht für den Test Y_1 eribt sich durch:

$$w_1 = \frac{\lambda_1}{\theta_1^2} = \frac{1.8}{0.38} = 4.737\,.$$

Auf analoge Weise lassen sich die Gewichte für die anderen Test berechnen.

Die maximale Reliabilität (.936) fällt deutlich höher aus als die Reliabilität der einfachen Summe (.821) und der durch Cronbachs α gegebene Wert (.730).

Die einfache Formel zur Berechnung der optimalen Reliabilität ist im Falle von Tests, welche durch mehrere latente Konstrukte beeinflusst sind, nicht mehr gültig. In diesem Fall überschätzt die Formel den korrekten Wert. So ergibt sich z.B. für die 4 Tests von Bsp.2-8 gemäss der Formel ein Wert von .911. Die maximale Reliabilität beträgt jedoch lediglich .907.
Meines Wissens gibt es kein Verfahren um die maximale Reliabilität in diesem Falle mit Hilfe des Programms zur Schätzung von Strukturgleichungsmodellen oder mittels elementarer Matrizenoperationen – ähnlich wie im Falle der Reliabilität von Summen – zu berechnen. Im Folgenden werde ich 3 Methoden zur Ermittlung der maximalen Reliabilität präsentieren.

Bemerkung:

Die im Folgenden dargestellte Methode 2-8 erfordert spezielle Kenntnisse über Matrizen. Sie wird für das Verständnis des Nachfolgenden nicht benötigt und kann daher übersprungen werden.

Methode 2-8: *Ermittlung der maximalen Reliabilität und der optimalen gewichteten Summe von Testwerten im allgemeinen faktorenanalytischen Modell.*

Gegeben: Das allgemeine faktorenanalytische Modell (siehe Abb. **xxxx**):

Gesucht: Die maximale Reliabilität der Summe der Indikatoren und die zugehörigen optimalen Gewichte.

Methode I: Direkte Maximierung der Reliabilität:

1. Verwende Methode 2-6 zur Berechnung der Reliabilität der gewichteten Summe im allgemeinen faktorenanalytischen Modell mit Hilfe von Matrizen.
2. Maximiere die berechnete Reliabilität unter Veränderung der Gewichte. Hierbei ist darauf zu achten, dass alle Gewichte das gleiche Vorzeichen aufweisen (entweder alle positiv oder alle negativ sind).

Bemerkung:
Zur Durchführung des zweiten Schrittes benötigt man einen Optimierer. So kann man z.B. bei Verwendung von Excel den mitgelieferten *Solver* zur Optimierung verwenden (siehe z.B. Macho, 2002).

Methode II (Greene & Carmines, 1980):

1. Die maximale Reliabilität entspricht dem grössten Eigenwert der Matrix:

$$\hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\Lambda}^{\mathrm{T}} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2}. \qquad (2\text{-}25)$$

wobei gilt:

$\boldsymbol{\Phi}$ ist die Kovarianzmatrix der latenten Konstrukte,

$\boldsymbol{\Lambda}$ ist die Ladungsmatrix,

$\hat{\boldsymbol{\Sigma}}$ ist die vom Modell implizierte Kovarianzmatrix.

(Zur genauen Form der Matrizen, siehe Methode 2-6).

Die Matrix $\hat{\boldsymbol{\Sigma}}^{-1/2}$ repräsentiert so etwas, wie die inverse Wurzel aus der Matrix $\hat{\boldsymbol{\Sigma}}$. Diese Matrix erhält man am besten mit Hilfe der so genannten *Singulärwertzerlegung* (SVD = singular value decomposition), über die jedes Statistikprogramm – wie SPSS, Systat oder R – verfügt.
Die SVD liefert folgende Matrizenzerlegung:

$$\hat{\boldsymbol{\Sigma}} = \mathbf{V} \cdot \boldsymbol{\Delta} \cdot \mathbf{V}^{\mathrm{T}}$$

Hierbei gilt:

$\mathbf{V}$ ist die Matrix der Eigenvektoren von $\hat{\boldsymbol{\Sigma}}$.

$\mathbf{\Lambda}$ ist eine Diagonalmatrix der Singulärwerte (= Eigenwerte von $\hat{\mathbf{\Sigma}}$) mit allen Einträgen > 0 (falls $\hat{\mathbf{\Sigma}}$ eine sinnvolle Kovarianzmatrix darstellt):

$$\mathbf{\Lambda} = \begin{bmatrix} \delta_1^2 & 0 & \cdots & 0 \\ 0 & \delta_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_m^2 \end{bmatrix}$$

Es gilt nun:

$$\hat{\mathbf{\Sigma}}^{-1/2} = \mathbf{V} \cdot \mathbf{\Lambda}^{-1/2} \cdot \mathbf{V}^{\mathbf{T}},$$

wobei gilt:

$$\mathbf{\Lambda}^{-1/2} = \begin{bmatrix} \frac{1}{\delta_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\delta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\delta_m} \end{bmatrix},$$

d.h. man bildet eine Diagonalmatrix, mit Diagonaleinträgen gleich 1 dividiert durch die Wurzeln aus den Diagonaleinträgen von $\mathbf{\Lambda}$ (Dies ist immer möglich, da die Diagonaleinträge von $\mathbf{\Lambda}$ alle grösser 0 sind).

2. Die optimalen Gewichte ergeben sich durch:

$$\mathbf{w}_0 = \hat{\mathbf{\Sigma}}^{-1/2} \cdot \mathbf{u}_0 \qquad (2\text{-}26)$$

Hierbei ist $\mathbf{u}_0$ der zum maximalen Eigenwert gehörige Eigenvektor.

Begründung:

Die Methode basiert auf folgendem Theorem über Matrizen (siehe z.B. Noble & Daniel, 1988):

Gegeben:

$\mathbf{\Sigma}$ sei eine (n×n)-Kovarianzmatrix,

$\rho = \frac{\mathbf{w}^{\mathbf{T}} \cdot \mathbf{\Sigma} \cdot \mathbf{w}}{\mathbf{w}^{\mathbf{T}} \cdot \mathbf{w}}$ ist der *Rayleigh Quotient*.

Hierbei ist **w** ein beliebiger (n×1)-Vektor.

Gesucht:

1. Der maximale Wert von ρ unter allen möglichen **w** (bei fixem $\boldsymbol{\Sigma}$)
2. Der Gewichtsvektor $\mathbf{w}_0$, welcher ρ maximal macht.

Das Theorem besagt nun:

1. Der maximale Wert von ρ entspricht dem maximalen Eigenwert von **K**.
2. Der Gewichtsvektor $\mathbf{w}_0$, welcher ρ maximiert, ist der zum maximalen Eigenwert gehörige Eigenvektor.

Im aktuellen Fall wollen wir die Reliabilität der gewichteten Summe *Y* maximieren (vergleiche Methode 2-6):

$$\mathrm{Rel}(Y) = \frac{\mathrm{Var}_\eta(Y)}{\mathrm{Var}(Y)} = \frac{\mathbf{w}^\mathbf{T} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\Lambda}^\mathbf{T} \cdot \mathbf{w}}{\mathbf{w}^\mathbf{T} \cdot \hat{\boldsymbol{\Sigma}} \cdot \mathbf{w}} \qquad (2\text{-}27)$$

Leider entspricht der Ausdruck auf der rechten Seite nicht dem Rayleigh-Quotienten. Daher stellen wir **w** als einer (umkehrbar eindeutigen) Funktion einer neuen Variable **u** dar:

$$\mathbf{w} = \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \mathbf{u} \qquad (2\text{-}28)$$

Somit ergibt sich:

$$\begin{aligned} \mathrm{Rel}(Y) &= \frac{\mathbf{w}^\mathbf{T} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\Lambda}^\mathbf{T} \cdot \mathbf{w}}{\mathbf{w}^\mathbf{T} \cdot \hat{\boldsymbol{\Sigma}} \cdot \mathbf{w}} \\ &= \frac{\mathbf{u}^\mathbf{T} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\Lambda}^\mathbf{T} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \mathbf{u}}{\mathbf{u}^\mathbf{T} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \hat{\boldsymbol{\Sigma}} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \mathbf{u}} \\ &= \frac{\mathbf{u}^\mathbf{T} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\Lambda}^\mathbf{T} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \mathbf{u}}{\mathbf{u}^\mathbf{T} \cdot \mathbf{u}} \end{aligned} \qquad (2\text{-}29)$$

Der Übergang von der vorletzten zur letzten Zeile beruht auf der Identität $\hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \hat{\boldsymbol{\Sigma}} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} = \mathbf{I}$, wobei **I** die ($n \times n$)-Identitätsmatrix (Diagonalmatrix mit 1 in Hauptdiagonale und 0 sonst) ist.

Der Ausdruck in der letzten Zeile von Gleichung (2-29) hat die Form des Rayleigh-Quotienten und der maximale Wert dieses Quotienten ist identisch zur maximalen Reliabilität.

Die Anwendung des Theorems besagt, dass dieser Wert der grösste Eigenwert der Matrix $\hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\Lambda} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2}$ ist. Dies entspricht exakt dem obigen Ergebnis von Gleichung (2-25).

Der dem grössten Eigenwert korrespondierenden Eigenvektor $\mathbf{u}_0$ entspricht nicht dem maximalen Gewichtsvektor $\mathbf{w}_0$ in der ursprünglichen Formulierung von $\mathrm{Rel}(Y)$ (siehe erste Zeile in 3-47). Letzterer ergibt sich mit Hilfe der Transformation:

$$\mathbf{w}_0 = \hat{\mathbf{\Sigma}}^{-1/2} \cdot \mathbf{u}_0 ,$$

gemäss Gleichung (2-28).

Methode III (Li, 1997):
1. Ermittle den grössten Eigenwert ρ_{max} der Matrix:

$$\mathbf{\Theta}^{-1/2} \cdot \mathbf{\Lambda} \cdot \mathbf{\Phi} \cdot \mathbf{\Lambda}^{\mathrm{T}} \cdot \mathbf{\Theta}^{-1/2} . \qquad (2\text{-}30)$$

wobei gilt:
$\mathbf{\Theta}$ ist die Kovarianzmatrix der Fehler (die anderen Symbole haben die gleiche Bedeutung wie oben).
2. Die maximale Reliabilität ergibt sich durch:

$$\mathrm{Rel}_{max} = \frac{1}{1+1/\rho_{max}} \qquad (2\text{-}31)$$

3. Die optimalen Gewichte ergeben sich durch:

$$\mathbf{w}_0 = \mathbf{\Theta}^{-1/2} \cdot \mathbf{u}_0 \qquad (2\text{-}32)$$

Hierbei ist $\mathbf{u}_0$ der zum maximalen Eigenwert ρ_{max} gehörige Eigenvektor.

Bemerkung:
Auf den ersten Blick scheint Methode III eher aufwändiger als Methode II, da ein zusätzlicher Schritt zur Berechnung der maximalen Reliabilität benötigt wird.
Ein Vorteil der Methode ergibt sich jedoch, falls die Fehler unkorreliert sind. Denn in diesem Falle ist $\mathbf{\Theta}$ eine Diagonalmatrix mit den Varianzen der Fehler in der Hauptdiagonale und $\mathbf{\Theta}^{-1/2}$ entspricht einer Diagonalmatrix mit den inversen Standardabweichungen der Fehler als Diagonaleinträge. Man erspart sich damit die aufwändige Matrizenzerlegung von Methode II.

Bsp.2-17: Maximale Reliabilität im allgemeinen Testmodell (Fortsetzung von Bsp.2-9)

Gegeben:
Das Modell von Abb. 2-27 (Diese entspricht dem Modell von Abb. 2-21, nach Entfernung der Phantomvariablen und der zugehörigen Pfade).

Die maximale Reliabiltät der Summe der 5 Tests beträgt .687 (die Reliabilität der einfachen Summe beträgt .588 und Cronbachs α ergibt .500).
Die optimalen Gewichte sind:
$w_1 = 0.934$, $w_2 = 0.200$, $w_3 = 0.174$, $w_4 = 0.194$ und $w_5 = 0.138$.
Jede der drei oben dargestellten Methoden führt zum gleichen Ergebnis (Übung 2-17).

Wir wollen als nächstes zeigen, dass die maximale Reliabilität die drei Anforderungen bezüglich der Monotonie erfüllt.

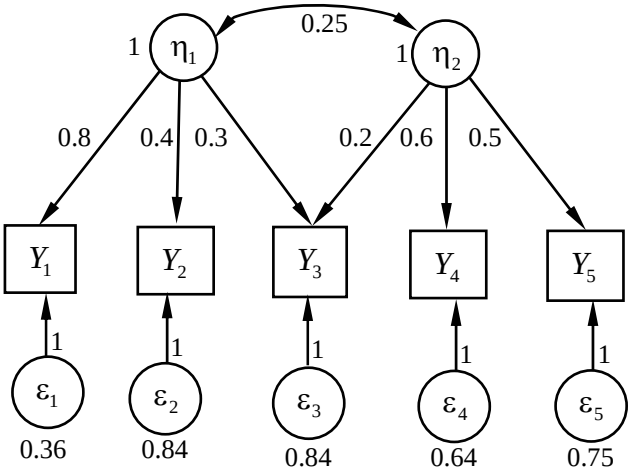


Abb. 2-27: *Faktorenanalytische Modell von fünf Tests.*

Bsp.2-18: Die maximale Reliabilität erfüllt die Anforderungen bezüglich der Monotonie von Prinzip 2-4 (Fortsetzung von Bsp.2-13, Bsp.2-14, Bsp.2-15).

1. Für Bsp.2-13 ergibt sich die maximale Reliabilität durch:

$$\text{Rel}_{\max} = \frac{2 \cdot \frac{.60}{1-.60} + 5 \cdot \frac{.10}{1-.10}}{1 + 2 \cdot \frac{.60}{1-.60} + 5 \cdot \frac{.10}{1-.10}} = .78$$

   Dieser Wert ist grösser als die Reliabilität der Summe aus den beiden Items mit höchster Reliabilität (Letztere beträgt .75)

2. Für Bsp.2-14 betragen die maximalen Reliabilitäten:

$$\text{Rel}_{\max} = \frac{\frac{.60}{1-.60} + \frac{.10}{1-.10}}{1 + \frac{.60}{1-.60} + \frac{.10}{1-.10}} = .62$$

bzw.

$$\text{Rel}_{\max} = \frac{\frac{.65}{1-.65} + \frac{.10}{1-.10}}{1 + \frac{.65}{1-.65} + \frac{.10}{1-.10}} = .66$$

Damit erhält die Kombination mit den Testitems mit grösserer Reliabilität auch insgesamt die höhere Reliabilität. Zusätzlich ist die Reliabilität für beide Items höher als für jedes Item für sich alleine.

3. Für Bsp.2-15 ergibt sich eine höhere maximale Reliabilität $\text{Rel}_{\max} = .617$ für das kongenerische Modell im Vergleich zum Modell mit Korrelation von .95 zwischen den Faktoren: $\text{Rel}_{\max} = .616$ (vergleiche Abb. 2-26).

Die hier dargestellte Methode eignet sich auch bestens zur Bildung von Itempaketen mit optimaler Reliabilität, indem anstelle einfacher Summen die optimalen Gewichte verwendet werden.

Nach dieser ausführlichen Behandlung des Konzepts der Reliabilität wenden wir uns dem zweiten wichtigen Begriff zur Beurteilung von Test zu.

2.4 Validität: Konzept und Schätzung

Das zweite zentrale Konzept im Kontext der Messung latenter Konstrukte ist die *Validität*. Wir beginnen mit einer kleinen Übersicht über die klassischen Ansätze zur Erfassung des Konzepts. Dem folgt die Behandlung des Konzepts im Rahmen latenter Variablenmodelle.

2.4.1 Klassische Konzeptionen von Validität

Fast jedes Lehrbuch, welches sich mit dem Problem der Validität von Messungen befasst, präsentiert das Diktum von Kelley (1927), wonach ein Test *valide* ist, *falls er das misst, was er zu messen vorgibt*. Gemäss dieser Forderung sollte ein Test zur Messung von emotionaler Intelligenz, möglichst genau dieses Konzept messen und nicht andere, wie z.B. verbale Intelligenz oder soziale Kompetenz (vorausgesetzt, dass emotionale Intelligenz von diesen Konstrukten getrennt werden kann). Das Diktum von Kelley ist einleuchtend und wir werden später

eine Definition von *Validität eines Test* im Rahmen latenter Variablenmodelle präsentieren, welches diese Forderung präzisiert.

Einen Meilenstein auf dem Weg zur Präzisierung des theoretischen Begriffs der Validität bildet ein Aufsatz von Cronbach und Meehl (1955), in welchem der Begriff der *Konstruktvalidität* präzisiert wird. Gemäss dieser Konzeption ist es sinnvoller, von der *Validität einer Schlussfolgerung* aufgrund der Anwendung eines Tests zu sprechen als von der Validität eines Tests. In moderner Terminologie lässt sich das Konzept wie folgt ausdrücken: *Schlussfolgerungen basierend auf einem Test sind valide, wenn sie auf einem korrekten Messmodell des Test beruhen.* Die Autoren fokussieren hierbei auf drei Aspekte von Messmodellen (vergleiche hierzu Konzept 2-1):

1. Die Beziehungen zwischen den zu messenden latenten Konstrukten sollte korrekt spezifiziert sein;
2. Die Relationen zwischen den Konstrukten und Messungen sollten im Modell korrekt repräsentiert sein;
3. Die Beziehungen zwischen den Messungen sollten im Modell korrekt widergespiegelt werden.

Betrachtet man Messmodelle als Kausalmodelle, so fällt das Konzept der *Konstruktvalidität* mit jenem der *internen Validität* zusammen. Gemäss letzterem ist *eine Schlussfolgerung oder Erklärung valide, sofern sie auf einem korrekten Kausalmodell basiert.*

Für eine lange Zeit existierte eine *Trias der Validitäten* (siehe z.B. Angoff, 1988):

1. *Konstruktvalidität*: Hierbei handelt es sich um die von Cronbach und Meehl (1955) präsentierte Konzeption.
2. *Inhaltsvalidität:* Diese betrifft die Frage, inwieweit ein Konstrukt durch die Indikatoren adäquat abgedeckt wird, d.h. inwieweit die Indikatoren alle Facetten eines Konstrukts erfassen. So wäre es z.B. möglich, ein Konstrukt mit Hilfe einer Menge von sehr ähnlichen Tests zu messen. Als Folge ergäbe sich eine hohe Reliabilität, da die Tests vermutlich hoch korreliert sind. Das Kriterium der Inhaltsvalidität wäre jedoch nicht erfüllt, da diese Tests die Bedeutung des Konstrukts nicht voll ausschöpfen.
3. *Prädiktive Validität / Kriteriumsvalidität:* Diese Art von Validität betrifft die Vorhersage eines vorgegebenen Kriteriums *K* aufgrund der Testwerte. Ein Test ist demnach umso valider, je besser er *K* vorhersagt, bzw. je gösser der statistische Zusammenhang zwischen den Testwerten und dem Kriterium *K* ist.

Konstrukt- und Inhaltsvalidität sind bis heute von Bedeutung. Prädiktive Validität wird hingegen in modernen Texten zur Testtheorie nicht mehr als sinnvolles testtheoretisches Konzept betrachtet (siehe z.B. Bollen, 1989; McDonald, 1999).

Auch das Konzept der Konstruktvalidität geriet zunehmend unter Kritik. Ein Kritikpunkte lautet, dass das Konzept in der Praxis nur schwer anwendbar sei, da es eine detaillierte Spezifikation der Beziehungen zwischen den latenten Konstrukten (im Fachjargon: ein engmaschiges *nomologisches Netzwerk von Konstrukten*) voraussetzt. Dieses sei jedoch in der psychologischen Praxis nicht vorhanden (Borsboom et al. 2004; Kane, 2001).
Ein zweiter Kritikpunkt besagt, dass das Konzept vor allem die Beziehungen zwischen abstrakten Konstrukten betont und dabei das oben genannte Diktum von Kelley (1927), wonach ein Test valide ist, falls es das misst, was er zu messen vorgibt, aus dem Auge verliert (Borsboom, et al., 2004).
Diese Kritik am Konzept der Konstruktvalidität ist nicht völlig gerechtfertigt. Denn – wie oben ausgeführt – umfasst das nomologische Netzwerk nicht nur die Beziehung zwischen den latenten Konstrukten, sondern auch jene zwischen Konstrukt und Messung, sowie zwischen den Messungen (vergleiche Cronbach und Meehl, 1955, Seite 290).
Das zentrale Problem des Konzepts der Konstruktvalidität bestand meines Erachtens darin, dass zum Zeitpunkt der Formulierung keine praktisch anwendbaren Methoden zur statistischen Testung latenter Variablenmodelle vorhanden waren. Daher wurden Indizes entwickelt, welche rein auf beobachteten Daten beruhten. Ähnlich wie im Falle der Reliabilität sind derartige Maße problematisch. Dies soll im Folgenden demonstriert werden.

2.4.2 Klassische Ansätze zur Erfassung der Validität

Wir betrachten zuerst das Problem der Erfassung der Kriteriumsvalidität.

2.4.2.1 Messung der Kriteriumsvalidität

Ein einfaches Maß zur Erfassung des statistischen Zusammenhangs zwischen einem Test *Y* und dem Kriterium *K* ergibt sich durch die Korrelation $\mathrm{Korr}(Y,K)$ zwischen Test und Kriterium. Der Korrelationskoeffizient wird auch als *Validitätskoeffizient* bezeichnet.
Die Sinnhaftigkeit dieses Maßes hängt jedoch von der Struktur des zugrunde liegenden Messmodells ab. Betrachten wir hierzu das Messmodell von Abb. 2-28. Wir nehmen an, dass dieses Modell die Beziehung zwischen Test *Y* und Kriterium *K* korrekt beschreibt.

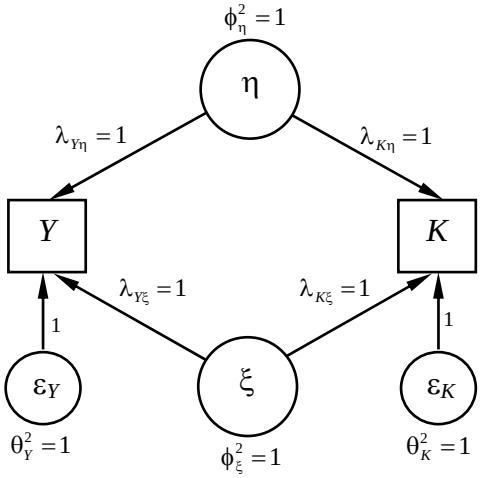


Abb. 2-28: *Messmodell von Indikator Y mit Kriterium K. η repräsentiert das Zielkonstrukt, welches durch Y und K gemessen werden soll und ξ ein weiteres latentes Konstrukt, welches beide Messungen beeinflusst.*

Hierbei repräsentiert η das zu messende Zielkonstrukt und ξ ein weiteres Konstrukt, welches einen Einfluss auf die beiden Tests ausübt (z.B. sei η = soziale Intelligenz und ξ = verbale Fähigkeiten). Mittels Kovarianzrechnung kann gezeigt werden, dass für die gegebenen Werte von Abb. 2-28 die Korrelation $\text{Korr}(Y,K) = 2/3$ beträgt. Der Anteil der Kovarianz, welcher dadurch zustande kommt, dass *Y* und *K* beide mit dem Zielkonstrukt η verbunden sind, beträgt jedoch nur die Hälfte der gesamten Kovarianz zwischen *Y* und *K* (der Rest wird durch das Konstrukt ξ erklärt).
Dieses Beispiel demonstriert zwei gravierende Probleme des Validitätskoeffizienten:

1. Der Validitätskoeffizient gibt keinen direkten Aufschluss darüber, inwieweit *Y* und *K* das gleiche Konstrukt messen.
2. Der Koeffizient gibt auch keinen direkte Information über den Zusammenhang von Zielkonstrukt η und *Y* (Die Korrelation zwischen beiden beträgt im aktuellen Fall $1/\sqrt{3}$).

Wenden wir uns nun dem Problem der Erfassung der Konstruktvalidität zu.

2.4.2.2 Ermittlung der Konstruktvalidität

Nehmen wir an, das Ziel bestünde im Folgenden darin, einen Aspekt des nomologischen Netzwerkes zu messen, nämlich, die Beziehung zwischen zwei abstrakten Konstrukten. Nehmen wir der Konkretheit wegen an, dass jemand den Grad des Zusammenhanges zwischen *emotionaler Intelligenz* (η_1) und *sozialer Kompetenz* (η_2) erfassen möchte.

Wir nehmen an, dass die Realität durch das Messmodell in Abb. 2-29 korrekt abgebildet wird. Das Modell nimmt an, dass die beiden Messungen Y_1 und Y_2 nicht nur durch die beiden relevanten Konstrukte η_1 und η_2 beeinflusst werden, sondern auch durch die verbalen Fähigkeiten der Personen (ξ).
Es ist unmittelbar einleuchtend, dass in der durch das Modell beschriebenen Messsituation die Verwendung der Korrelation zwischen Y_1 und Y_2 als Maß für den Grad des Zusammenhanges zwischen den beiden Konstrukten η_1 und η_2 ungeeignet ist. Denn – ähnlich wie im vorangegangenen Beispiel – ist ein Teil des Zusammenhanges zwischen den beiden Test durch ein drittes latentes Konstrukt bedingt.

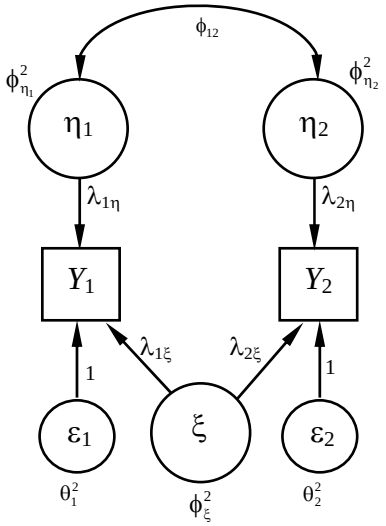


Abb. 2-29: *Messmodell zweier Messungen: η_1 = emotionale Intelligenz, η_2 = soziale Kompetenz und ξ = verbale Fähigkeiten.*

Die beiden Beispiele zur Erfassung der Validität mit Hilfe von Korrelationen sollten vor allem eines deutlich machen:

> *Eine adäquate Beurteilung der Validität von Tests zur Messung abstrakter Konstrukte und deren Relationen ist ohne Berücksichtigung des zugrunde liegenden Messmodells nicht möglich.*

Wie aus Abschnitt 2.3 hervorgeht, bleibt die Aussage auch dann korrekt bleibt, wenn man im letzten Satz das Wort *Validität* durch *Reliabilität* ersetzt.
Wir wenden und nun dem Problem der Beurteilung der Validität eines Test im Rahmen latenter Variablenmodelle zu.

2.4.3 Erfassung der Validität von Tests im Kontext latenter Variablenmodelle

Wir beginnen mit Definition des Konzepts der Validität eines Tests.

Konzept 2-11 *Validität eines Tests (Indikators, Messung):* (Bollen, 1989; Borsboom, Mellenbergh, & Van Heerden, 2004):

Ein Test ist *valide*, falls die systematischen Variationen der Testwerten durch Variationen des zu messenden Konstrukts verursacht werden.
Die *Validität einer Test* entspricht daher der Stärke der *direkten kausalen* Relation zwischen der latenten Variablen und dem Indikator.

Zum besseren Verständnis sei das Konzept an Hand eines Beispiels illustriert.

Bsp.2-19: Validität eines Tests

Gegeben:

1. Das zu messende Zielkonstrukt η: Emotionale Intelligenz;
2. Ein Test *Y* zur Messung von η;
3. Ein weiteres Konstrukt ξ: Soziale Kompetenz.

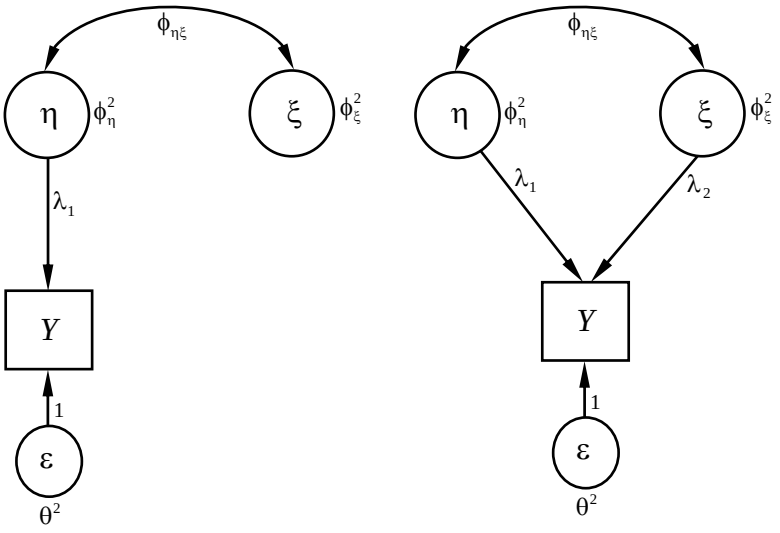


Abb. 2-30: *Zwei Messmodelle: Gemäss Modell (a) ist Y ein valider Test des Konstrukts η, gemäss Modell (b) ist Y nicht oder nur teilweise valide.*

Abb. 2-30 zeigt zwei mögliche Messmodelle zur Repräsentation der Beziehungen zwischen den involvierten Grössen.
Im linken Modell (a) wird der Test nur durch das zu messende Zielkonzept beeinflusst. Jede systematische Variation von *Y* ist daher auf Variationen des zugrunde liegenden latenten Konstrukts zurückzuführen. Daher ist *Y* ein valider Test für η.

Im rechten Modell (b) wird der Test durch beide Konstrukte beeinflusst. Dies schränkt die Validität von *Y* als Maß für η ein.
Man beachte, dass die Reliabilität von *Y* in Modell (b) höher sein kann als im Modell (a). Dies ergibt sich dadurch, dass die Reliabilität die gesamte systematische Varianz im Verhältnis zur Gesamtvarianz betrachtet und keine weiteren Unterscheidungen macht, ob diese Varianz durch das zu messende Zielkonstrukt oder durch andere Konstrukte verursacht wird.

Die gegebene Konzeption von Validität präzisiert die Bedeutung von Kelleys (1927) Diktum, wonach ein Test valide ist, wenn er das misst, was er zu messen vorgibt:

Ein Test misst genau das, was er zu messen vorgibt, wenn die systematischen Variationen der Testwerte ausschliesslich durch Unterschiede im zugrunde liegenden Zielkonstrukt verursacht sind.

Die gegebene Definition legt nahe, dass die Beziehung zwischen den Konstrukten für die Validität von Tests irrelevant ist. Demzufolge wäre auch dieser Aspekt der Konstruktvalidität von Cronbach und Meehl (1955) völlig irrelevant für die Beurteilung der Validität eines Tests.
Diese Schlussfolgerung ist nicht korrekt. Falls z.B. in Bsp.2-19 die latenten Konstrukte perfekt korreliert sind, so ist *Y* auch unter Modell (a) nicht länger ein valider Test für η.
Es ist weiters zu beachten, dass die korrekte Beurteilung der Validität eines Tests ein (einigermassen) korrektes Messmodell voraussetzt und dieses muss auch die Beziehungen zwischen den latenten Konstrukten beinhalten. Daher scheint mir das Konzept der Konstruktvalidität von Cronbach und Meehl auch heute noch wichtig.

Nach dieser ausführlichen Diskussion des Validitätskonzepts wenden wir uns dem Problem der Schätzung der Validität zu.

2.4.4 Schätzung der Validität eines Tests in latenten Variablenmodellen

Bollen (1989) schlägt zwei Validitätsmaße, die im Folgenden vorgestellt werden.

2.4.4.1 Der Ladungskoeffizient als Maß der Validität

Aufgrund der in Konzept 2-11 gegebenen Definition der Validität eines Test *Y* als Maß für das Konstrukt η wird diese in idealer Weise durch den Ladungskoeffizienten λ_Y^{η} repräsentiert, da letzterer ein direktes Maß für die strukturelle Relation zwischen Indikator und latente Variable ist. Wird die Skala des latenten Konstrukts mittels eines Skalierungsindikators festgelegt, so bestimmt sich die Grösse der Ladungskoeffizienten immer relativ zur Ladung des Skalierungsindikators. Die

Validität eines Indikators ist daher immer nur relativ (d.h. im Verhältnis) zu jener des zur Skalierung verwendeten Indikators bestimmbar.
Bei Verwendung der unstandardisierten Koeffizienten ergibt sich das zusätzliche Problem, dass die Varianzen von latenter Variable und Indikator einen Einfluss auf die Grösse des Koeffizienten nehmen. Falls daher nur eine Population untersucht wird, ist im Normalfall der standardisierte Koeffizient als Maß für die Validität zu bevorzugen.
Falls ein Test *Y* einzig durch das Zielkonstrukt kausal beeinflusst wird und damit in optimaler Weise valide ist, so gilt:

$$\lambda_Y^{\eta} = \sqrt{\mathrm{Rel}(Y)},$$

wobei λ_Y^{η} den standardisierten Ladungskoeffizienten symbolisiert. Dies entspricht einem Grundsatz, wonach die maximale Validität gleich der Wurzel aus der Reliabilität ist (siehe z.B. Angoff, 1988).

2.4.4.2 EINDEUTIGE VALIDITÄTSVARIANZ

Bollen (1989) präsentiert ein zweites Maß für die Validität.

Konzept 2-12: *Eindeutige Validitätsvarianz:*

Die *eindeutige Validitätsvarianz* $\mathrm{Rel}_{\eta_i}(Y)$ entspricht jenem Anteil der Reliabilität von *Y*, der eindeutig auf das zugehörige latente Konstrukt η_i zurückgeführt werden kann.

$$\mathrm{Rel}_{\eta_i}(Y) = \mathrm{Rel}(Y) - \mathrm{Rel}_{\boldsymbol{\eta}\text{-}\{\eta_i\}}(Y) \qquad (2\text{-}33)$$

wobei gilt:

$$R^2_{\boldsymbol{\eta}\text{-}\{\eta_1\}\rightarrow Y} = \frac{\boldsymbol{\sigma}^{\mathbf{T}}_{Y\boldsymbol{\eta}\text{-}\{\eta_1\}} \cdot \boldsymbol{\Phi}^{-1}_{\boldsymbol{\eta}\text{-}\{\eta_1\}\rightarrow Y} \cdot \boldsymbol{\sigma}_{Y\boldsymbol{\eta}\text{-}\{\eta_1\}}}{Var(Y)} \qquad (2\text{-}34)$$

Die Symbole haben hierbei die folgende Bedeutung:

$R^2_{\boldsymbol{\eta}\rightarrow Y} = \mathrm{Rel}(Y)$	Ist die durch alle auf *Y* einwirkenden latenten Variablen erklärte Varianz von *Y*. Dies entspricht – gemäss der oben gegebenen Definition – der Reliabilität von *Y*. (Diese wird vom Programm ausgegeben)
$R^2_{\boldsymbol{\eta}\text{-}\{\eta_i\}\rightarrow Y}$	Ist die durch alle mit Ausnahme von η_i auf *Y* einwirkenden latenten Variablen erklärte Varianz von *Y*.
$\boldsymbol{\sigma}_{Y\boldsymbol{\eta}}$	Ist ein Spaltenvektor mit den Kovarianzen zwischen *Y* und jenen latenten Variablen, die auf *Y* einwirken.
$\boldsymbol{\Phi}^{-1}_{\boldsymbol{\eta}\rightarrow Y}$	Ist die inverse Kovarianzmatrix zwischen allen latenten Variablen, die auf *Y* einwirken.

$\boldsymbol{\sigma}_{Y\boldsymbol{\eta}\text{-}\{\eta_i\}}$	Ist ein Spaltenvektor mit den Kovarianzen zwischen *Y* und jenen latenten Variablen ohne η_i, die auf *Y* einwirken.
$\boldsymbol{\Phi}^{-1}_{\boldsymbol{\eta}\text{-}\{\eta_i\}\rightarrow Y}$	Ist die inverse Kovarianzmatrix zwischen allen latenten Variablen ohne η_i, die auf *Y* einwirken.

Die eindeutige Validitätsvarianz ergibt sich daher aus der Differenz, der Varianz, welche durch alle auf *Y* einwirkenden Variablen erklärt wird und der Varianz, welche durch alle auf *Y* einwirkenden Variablen, mit Ausnahme jenes latenten Konstrukts, bezüglich dessen die Validität von *Y* bestimmt werden soll.
Die eindeutige Validitätsvarianz steht in enger Verbindung zur Reliabilität. Dies ersieht man schon daraus, dass der Ausdruck zur Berechnung der Reliabilität (Gleichung 2-33) als erster Term in die Berechnung der eindeutigen Validitätsvarianz eingeht. Aufgrund dieser Relation ergibt sich, dass die Validität niemals die Reliabilität eines Indikators überschreiten kann. Weiters ergibt sich, dass die Validität exakt der Reliabilität entspricht, falls nur ein latentes Konstrukt einen Einfluss auf die Messung besitzt.
Bevor wir versuche, die Bedeutung der eindeutigen Validitätsvarianz bzw. der oben gegebenen Gleichungen zu verstehen, sei die numerische Berechnung an einem einfachen Beispiel demonstriert.

Bsp.2-20: Eindeutige Validitätsvarianz

Gegeben: Das Modell von Abb. 2-31.

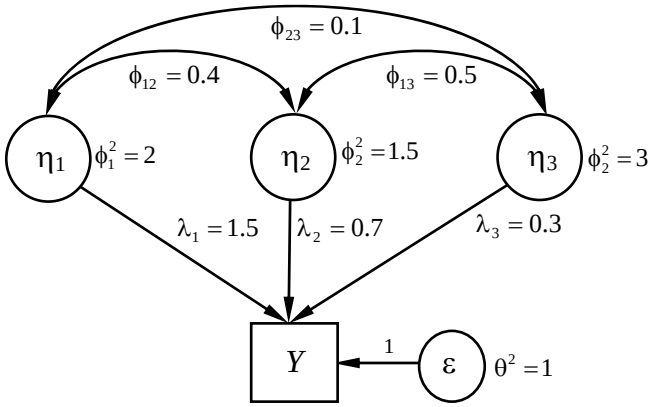


Abb. 2-31: *Strukturelles Modell zur Illustration der eindeutigen Validitätsvarianz.*

Gesucht: Die eindeutig Validitätsvarianzen von η_1, η_2, und η_3.

1. Die Reliabilität von *Y* beträgt: $\text{Rel}(Y) = R^2_{\boldsymbol{\eta}\rightarrow Y} = .869$ (Diesen Wert kann man sich vom Programm ausgeben lassen).

2. Die eindeutigen Validitätsvarianzen für die einzelnen Konstrukte betragen:
   $V_{Y\eta_1} = .557$, $V_{Y\eta_2} = .086$ und $V_{Y\eta_3} = .033$.

Zur Verdeutlichung sei die Berechnung von $V_{Y\eta_1}$ im Detail dargestellt:
Tab. 2-6 zeigt die vom Modell implizierte Kovarianzmatrix:

	η_1	η_2	η_3	Y
η_1	2	0.4	0.1	3.31
η_2	0.4	1.5	0.5	1.8
η_3	0.1	0.5	3	1.4
Y	3.31	1.8	1.4	7.645

Tab. 2-6: *Die vom Modell in Abb. 2-31 implizierte Kovarianzmatrix.*

Wir berechnen nun Gleichung (2-34)

$$R^2_{\boldsymbol{\eta}\text{-}\{\eta_1\}\to Y} = \frac{\boldsymbol{\sigma}^{\mathbf{T}}_{Y\boldsymbol{\eta}\text{-}\{\eta_1\}} \cdot \boldsymbol{\Phi}^{-1}_{\boldsymbol{\eta}\text{-}\{\eta_1\}\to Y} \cdot \boldsymbol{\sigma}_{Y\boldsymbol{\eta}\text{-}\{\eta_1\}}}{Var(Y)}$$

Die Grössen lassen sich aus der Kovarianzmatrix von Tab. 2-6 ermitteln:

$$\boldsymbol{\sigma}^{\mathbf{T}}_{Y\boldsymbol{\eta}\text{-}\{\eta_1\}} = \begin{bmatrix}1.8 & 1.4\end{bmatrix}, \text{ bzw. } \boldsymbol{\sigma}_{Y\boldsymbol{\eta}\text{-}\{\eta_1\}} = \begin{bmatrix}1.8\\1.4\end{bmatrix}$$

Weiters gilt.

$$\begin{aligned}\boldsymbol{\Phi}^{-1}_{\boldsymbol{\eta}\text{-}\{\eta_1\}\to Y} &= \begin{bmatrix}1.5 & 0.5\\0.5 & 3\end{bmatrix}^{-1}\\ &= \frac{1}{4.25}\cdot\begin{bmatrix}3 & -0.5\\-0.5 & 1.5\end{bmatrix}\\ &= \begin{bmatrix}0.706 & -0.118\\-0.118 & 0.353\end{bmatrix}\end{aligned}$$

Somit:

$$\frac{\boldsymbol{\sigma}^{\mathbf{T}}_{Y\boldsymbol{\eta}\text{-}\{\eta_1\}} \cdot \boldsymbol{\Phi}^{-1}_{\boldsymbol{\eta}\text{-}\{\eta_1\}\rightarrow Y} \cdot \boldsymbol{\sigma}_{Y\boldsymbol{\eta}\text{-}\{\eta_1\}}}{Var(Y)} = \frac{\begin{bmatrix}1.8 & 1.4\end{bmatrix} \cdot \begin{bmatrix}1.5 & 0.5\\ 0.5 & 3\end{bmatrix}^{-1} \cdot \begin{bmatrix}1.8\\ 1.4\end{bmatrix}}{7.645}$$

$$= \frac{2.386}{7.645}$$

$$= 0.312$$

Zieht man diesen Wert von $\text{Rel}(Y) = R^2_{\boldsymbol{\eta}\rightarrow Y} = .869$ ab, so ergibt sich der ermittelte Wert von .557.

Glücklicherweise kann man die eindeutige Validitätsvarianz auch mit Hilfe von SEM Programmen ermitteln (SEM = structural equation models). Um dies zu verstehen, wie diese Methode funktioniert, müssen wir aber noch genauer verstehen, was die Gleichung zur Bestimmung der eindeutigen Validitätsvarianz eigentlich macht.
Wir stellen uns zuerst auf den naiven Standpunkt und stellen uns die folgende einfache Frage:

Frage:
Warum kann man zur Bestimmung der eindeutigen Validitätsvarianz von η_1 nicht einfach das Modell von Abb. 2-32 verwenden?

Man beachte dass das Modell von Abb. 2-32 ein Teilmodell des Modells von Abb. 2-31 darstellt, wobei der Ladungskoeffizient, der sich durch die Schätzung des vollen Modells ergab, verwendet wurde.

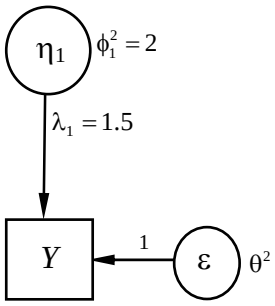


***Abb. 2-32**: Teilmodell des strukturellen Modell von Abb. 2-31.*

Das Problem dieses Modells besteht darin, dass die Varianz des latenten Konstrukts $\phi^2_1 = 2$ zu gross ist, da diese auch jenen Varianzanteil enthält, der sich durch die Kovarianz mit den beiden anderen latenten Konstrukten ergibt. Diese Varianz muss aus der Gesamtvarianz von η_1 »heraus partialisiert« werden. Dies kann mit Hilfe von Strukturgleichungsmodellen geschehen.
Im aktuellen Fall wird dies mit Hilfe des Modells von Abb. 2-33 erreicht. Die Variable ζ (zeta) repräsentiert das Residuum, d.h. denjenigen Varianzanteil in η_1, der nicht durch η_2 und η_3 erklärt wird. Die

residuale Varianz wird in Abb. 2-33 durch das Symbol ψ^2 repräsentiert.

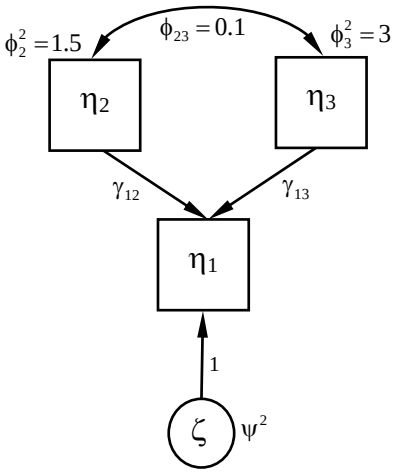


Abb. 2-33: *Strukturgleichungsmodell zur Bestimmung der residualen Varianz von η_1.*

Bemerkung:

In Abb. 2-33 wurden die latenten Variablen η_1, η_2, und η_3 durch Kästchen repräsentiert. Es muss daher die (geschätzte) Kovarianzmatrix dieser 3 Variablen vorliegen. Diese Matrix kann man sich vom SEM Programm ausgeben lassen.

Die Schätzung der residualen Varianz von η_1 erfolgt daher aufgrund der vom Programm geschätzten Kovarianzmatrix zwischen den latenten Variablen.

Im aktuellen Beispiel ergibt sich (bei Verwendung des Programms AMOS) als residuale Varianz der Wert $\psi^2 = 1.874$ (Der korrekte Wert ist jedoch 1.893). Setzt man diesen Wert für die Varianz von η_1 in das Modell von (anstelle von $\phi^2 = 2.0$), so erhält man als Reliabilität von Y den Wert $\text{Rel}(Y) = .577$ (identisch zu jenem Wert, welcher sich mit Hilfe der Matrizengleichung ergibt).

Das Programm AMOS begeht bei der Verwendung dieser Methode zur Berechnung der Reliabilität zwei Ungenauigkeiten, die sich jedoch gegenseitig aufheben:

1. Die residuale Varianz von η_1 (Abb. 2-33) wird nicht korrekt berechnet: $\psi^2 = 1.874$ anstelle von $\psi^2 = 1.893$.

   Verwendet man den von AMOS berechneten Wert zu Ermittlung der Reliabilität, so erhält man als Resultat:

   $$\text{Rel}(Y) = \frac{\lambda_1^2 \cdot \psi^2}{\text{Var}(Y)} = \frac{1.5^2 \cdot 1.874}{7.645} = .552$$

   also einen leicht verringerten Wert.

2. Die geschätzte Varianz von Y im Modell von Abb. 2-32 wird leicht unterschätzt:
   $\mathrm{Var}(Y) = 7.569$ anstelle von $\mathrm{Var}(Y) = 7.645$.

   Zusammen führen die beiden Ungenauigkeiten wieder zu einem korrekten Resultat:

$$\mathrm{Rel}(Y) = \frac{\lambda_1^2 \cdot \psi^2}{\mathrm{Var}(Y)} = \frac{1.5^2 \cdot 1.874}{7.569} = .557$$

   Verwendet man daher AMOS zur Schätzung der residualen Varianz, so sollte man auch die Reliabilität mittels AMOS berechnen und nicht mit Hilfe der Formel.

Die eindeutige Validitätsvarianz als Maß für die Validität hat den Nachteil, dass sie von der Korrelation zwischen der latenten Variable η_i und den anderen auf *Y* einwirkenden Variablen abhängt. Je grösser diese Korrelation, desto kleiner die eindeutige Validitätsvarianz.
So summieren sich in Bsp.2-20 die eindeutigen Validitätsvarianzen nicht zum Wert der Reliabilität (von .869) auf, sondern zu einem geringeren Wert von .677. Falls die Konstrukte unkorreliert sind, so entspricht die Summe der eindeutigen Validitätsvarianzen der Reliabilität.

Bevor wir die klassische Testtheorie endgültig verlassen, wollen wir noch ein Verfahren betrachten, welches in der Praxis häufig angewendet und von Vertretern der klassischen Testtheorie empfohlen wird (siehe z.B. Schmidt & Hunter, 1996). Hierbei handelt es sich um die Korrektur des so genanten Abschwächungseffekts (oder Ausdünnungseffekts).

2.5 Der Abschwächungseffekt (Ausdünnungseffekt) und dessen Korrektur

Rushton, Brainerd und Pressley (1983) demonstrieren an 12 Beispielen (vorwiegend) aus dem Bereich der Entwicklungspsychologie, auf welche Art und Weise die Verwendung von wenig Fehler behafteten Maßen zu theoretisch bedeutsamen Fehlinterpretationen führen kann. Eines ihrer Beispiele betrifft das Problem der Stabilität von Persönlichkeitseigenschaften. So wurde in der Sozialpsychologie die Frage diskutiert, ob es so etwas wie eine moralische Persönlichkeit gibt oder ob das moralische Verhalten vorwiegend durch die Situation beeinflusst wird. Hierbei hat sich die Ansicht durchgesetzt, dass das moralische Verhalten vorwiegend durch die Situation beeinflusst wird (siehe z.B. Nisbett und Ross, 1980). Rushton, Brainerd und Pressley (1983) wiesen nun darauf hin, dass die fehlenden Hinweise auf eine stabile moralische Persönlichkeitseigenschaft darauf zurückzuführen sei, dass die Messungen einerseits eine geringe Reliabilität aufweisen und andererseits meist nur wenige Messungen eines Konstrukts vorgenommen

wurden. Hieraus resultieren geringe Korrelationen. Das folgende Beispiel verdeutlicht dies.

Bsp.2-21: Abschwächungseffekt / Ausdünnungseffekt:

Gegeben: Ein Persönlichkeitsmerkmal η, das zu zwei Zeitpunkten gemessen wird.

Die Beziehung zwischen den beiden Konstrukten zu den beiden Zeitpunkten und die Messsituation ist durch das Modell in Abb. 2-34 repräsentiert

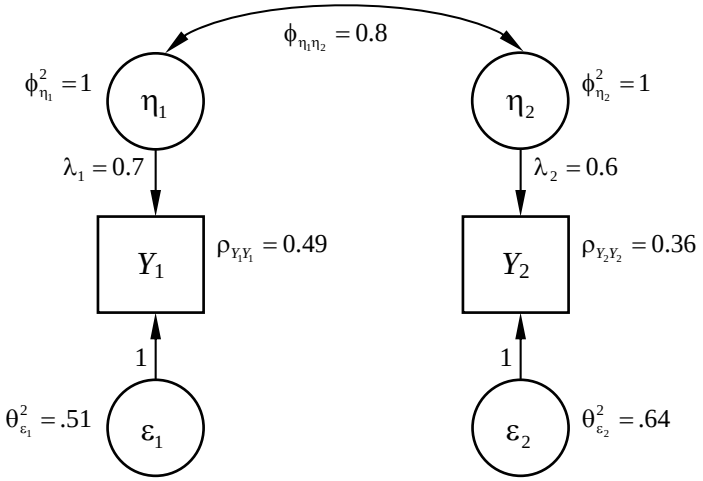


Abb. 2-34: *Messmodell für die Messung eines Konstrukts zu zwei Zeitpunkten.*

Die Symbole in Abb. 2-34 haben die folgende Bedeutung:

η_1, η_2	repräsentieren die zu messende Eigenschaft zu den beiden Zeitpunkten.
$\phi^2_{\eta_1}$, $\phi^2_{\eta_2}$	bezeichnen die Varianzen von η_1 und η_2.
$\phi_{\eta_1\eta_2}$	bezeichnet die Kovarianz zwischen η_1 und η_2. Diese ist im aktuellen Fall identisch zur Korrelation
Y_1, Y_2	bezeichnen die gemessenen Werte.
λ_1, λ_2	bezeichnen den direkten kausalen Einfluss von η_1 und η_2 auf Man spricht hierbei von den Ladungen von Y_1 und Y_2 auf η_1 und η_2.
$\rho_{Y_1Y_1}$, $\rho_{Y_2Y_2}$	bezeichnen die Reliabilitäten von Y_1 und Y_2.
ε_1, ε_2	bezeichnen die Messfehler.
$\theta^2_{\varepsilon_1}$, $\theta^2_{\varepsilon_2}$	bezeichnen die Varianzen der Messfehler.

Bemerkung zu den in Abb. 2-34 gezeigten Werten:

1. Die Reliabilitäten entsprechen den quadrierten Ladungskoeffizeneten. Dies darauf zurückzuführen, dass die Varianzen sowohl der latenten Konstrukte η_1 und η_2, wie auch der gemessenen Variablen Y_1, Y_2 alle gleich 1 sind.
2. Die Varianzen der gemessenen Variablen entsprechen jeweils der Summe aus Reliabilität und Fehlervarianz.
3. Da die Varianzen von η_1 und η_2 beide gleich 1 sind, ist die Kovarianz $\phi_{\eta_1\eta_2}$ zwischen den beiden Grössen identisch zur Korrelation.

Man beachte, dass die Korrelation zwischen den beiden Messzeitpunkten des Konstrukts relativ hoch ist: $\phi_{\eta_1\eta_2} = .8$. Dies weist auf eine hohe Stabilität des Konstrukts hin. Die Korrelation zwischen den gemessenen Werten ist hingegen deutlich geringer:

$$\rho_{Y_1Y_2} = .7 \cdot .8 \cdot .6 = .336 .$$

Diese verringerte Korrelation zwischen den gemessenen Grössen gegenüber wahren Konstrukten nennt sich *Abschwächungseffekt*.
Falls – wie im aktuellen Beispiel – die Reliabilitäten bekannt sind, kann der Abschwächungseffekt korrigiert werden, indem die Korrelation zwischen den beobachteten Werten durch die Wurzel des Produkts der Reliabilitäten dividiert wird:

$$\begin{aligned}\rho_{Y_1Y_2}^{\langle korrigiert\rangle} &= \frac{\rho_{Y_1Y_2}}{\sqrt{\rho_{Y_1Y_1} \cdot \rho_{Y_2Y_2}}} \\ &= \frac{.336}{\sqrt{.49 \cdot .36}} \\ &= \underline{\underline{0.8}}\end{aligned}$$

Diese Korrektur liefert im aktuellen Fall den exakten Wert. Sie basiert auf der Tatsache, dass die Reliabilitäten den Varianzanteil der latenten Konstrukte in den beobachteten Variablen wiedergeben.

2.5.1 Messfehler und die Aggregation von Daten

Rushton, Brainerd und Pressley (1983) schlagen zur Lösung des Problems die Verwendung von aggregierten Daten vor, d.h. der Summe der Ergebnisse mehrer Messungen. Sie argumentieren, dass durch Kombination der Daten der Messfehler ausgemittelt wird.
Das Problem dieser Methode besteht darin, dass es den Messfehler negiert und daher zwar zu einer Verbesserung führt, die jedoch in keiner Weise perfekt ist. Dies wird durch Bsp.2-22 demonstriert.

Bsp.2-22: Abschwächungseffekt und Datenaggregation:

Gegeben: Ein Persönlichkeitsmerkmal η wird zu zwei Zeitpunkten mit je drei Tests gemessen wird.

Das Messmodell ist in Abb. 2-35 gezeigt (Die Symbole zur Bezeichnung der Parameter wurden weggelassen).

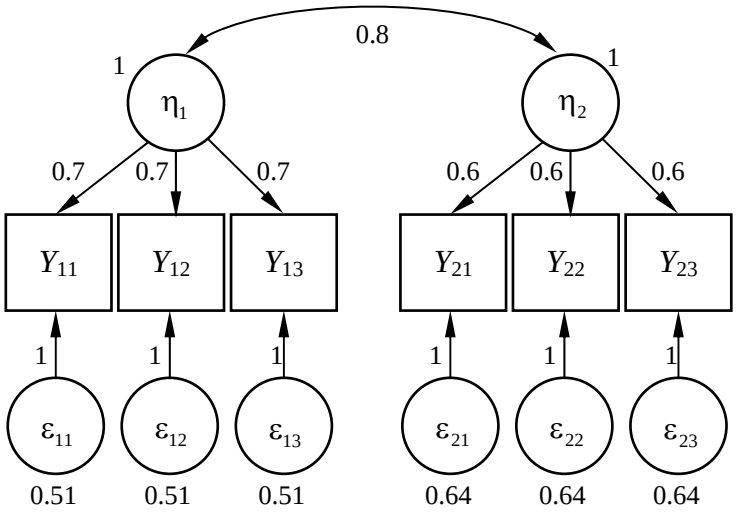


Abb. 2-35: *Messmodell für die Messung eines Konstrukts zu zwei Zeitpunkten mit je drei Indikatoren.*

Die Korrelation der Summen: $Y_1 = Y_{11} + Y_{12} + Y_{13}$ und $Y_2 = Y_{21} + Y_{22} + Y_{23}$ beträgt im aktuellen Fall $\rho_{Y_1Y_2} = .546$. Dies stellt zwar eine klare Verbesserung gegenüber der Situation mit nur einem Indikator dar (dort betrug die Korrelation $\rho_{Y_1Y_2} = .336$). Die korrekte Korrelation wird jedoch noch immer deutlich unterschätzt. Die Verwendung der Reliabilitäten der der Summenvariablen zur Korrektur der Abschwächung führt auch hier wieder zu einem korrekten Ergebnis. Die Reliabilitäten werden im aktuellen Fall von *parallelen Maßen* (siehe Konzept 2-5) durch den *Spearman-Brown* Koeffizienten (Konzept 2-8) bzw. durch *Cronbachs* α (Konzept Konzept ***2-9***) exakt repräsentiert (Beide Koeffizienten sind in diesem Fall identisch; siehe Übung 2-12). Es ergibt sich:

$$\rho_{Y_1Y_2}^{\langle korrigiert \rangle} = \frac{\rho_{Y_1Y_2}}{\sqrt{\rho_{Y_1Y_1} \cdot \rho_{Y_2Y_2}}}$$

$$= \frac{.546}{\sqrt{.742 \cdot .628}}$$

$$= \underline{\underline{0.8}}$$

Bsp.2-22 macht deutlich dass die Aggregation von Daten zwar das Problem der Abschwächung verkleinert, aber keineswegs völlig beseitigt, so lange nicht sehr viele Maße pro Konstrukt verwendet werden, die zudem noch hohe Reliabilitäten aufweisen sollten.

2.5.2 Die Grenzen der Abschwächungskorrektur

Die Korrektur der Abschwächung unter Verwendung von Cronbachs α führt zu Problemen, sobald die Ladungen nicht mehr alle identisch sind. In diesem Falle unterschätzt α die Reliabilitäten. Als Konsequenz wird die Korrelation zwischen den Konstrukten *überschätzt*.
Bsp.2-23 demonstriert dies auf eindrückliche Weise.

Bsp.2-23: Grenzen der Abschwächungskorrektur:

Gegeben: Ein Persönlichkeitsmerkmal η wird zu zwei Zeitpunkten mit je drei Tests gemessen wird. Die Tests besitzen unterschiedliche Ladungen.

Das Messmodell ist in Abb. 2-36 gezeigt (Die Symbole zur Bezeichnung der Parameter wurden weggelassen).

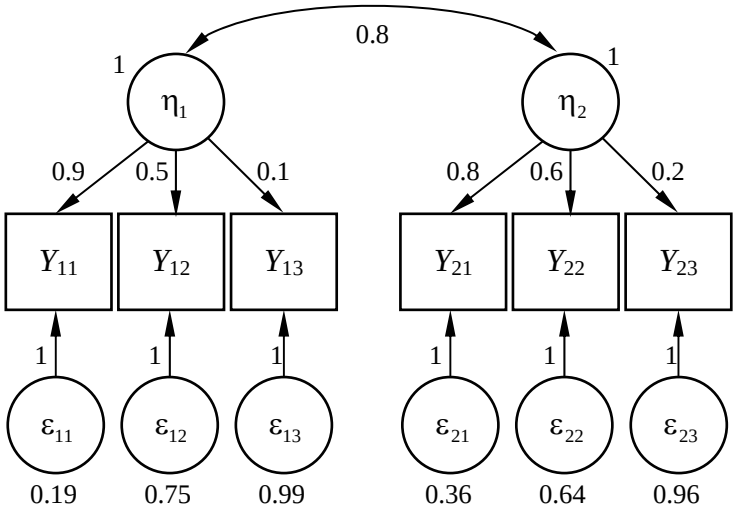


Abb. 2-36: *Messmodell für die Messung eines Konstrukts zu zwei Zeitpunkten mit je drei Indikatoren mit unterschiedlichen Ladungen.*

Die Korrelation der Summen: $Y_1 = Y_{11} + Y_{12} + Y_{13}$ und $Y_2 = Y_{21} + Y_{22} + Y_{23}$ beträgt im aktuellen Fall $\rho_{Y_1 Y_2} = .442$. Die Korrektur unter Verwendung von Cronbachs α ergibt:

$$\rho_{Y_1Y_2}^{\langle korrigiert \rangle} = \frac{\rho_{Y_1Y_2}}{\sqrt{\rho_{Y_1Y_1} \cdot \rho_{Y_2Y_2}}}$$
$$= \frac{.546}{\sqrt{.423 \cdot .504}}$$
$$= \underline{\underline{0.956}}$$

Diese Überschätzung der Korrelation ist auf die Unterschätzung der Reliabilitäten durch Cronbachs α zurückzuführen. Die korrekten Reliabilitäten – welche mit Hilfe von Strukturgleichungsmodellen geschätzt werden können (siehe Abschnitt 2.3.5) betragen: $\rho_{Y_1Y_1} = .538$ und $\rho_{Y_1Y_1} = .566$. Werden diese Werte zur Korrektur verwendet, so ergibt sich wiederum die korrekte Korrelation.
Die gezeigten Beispiele verdeutlichen zwei Dinge:

3. Die Nichtbeachtung von Messfehlern kann zu verzerrten Schätzungen und fehlerhaften Schlussfolgerungen führen.
4. Die Korrektur der Abschwächung unter Verwendung von Cronbachs α stösst an bestimmte Grenzen. Daher ist von einer blinden Anwendung (d.h. ohne Prüfung der Voraussetzungen) dieses Koeffizienten zur Korrektur der Abschwächung abzuraten.

Die Behandlung der klassischen Testmodelle und damit verbundener relevanter Grössen ist somit abgeschlossen. Wir wenden uns im Folgenden komplexeren Messmodellen zu.

2.6 Übungen zur Kapitel 2

Übung 2-1: *Fixieren der Korrelation zwischen zwei latenten Variablen auf den Wert 1.0:*

Gegeben: Das Modell von Abb. 2-7.
Zeige, dass die Korrelation zwischen η_i und η_i im Modell tatsächlich 1.0 ist.

Hinweis: Gehe wie folgt vor:

6. Zeige: $\mathrm{Var}(\eta_i) = \phi_i^2$ und $\mathrm{Var}(\eta_j) = \phi_j^2$.
7. Zeige: $\mathrm{Cov}(\eta_i, \eta_j) = \phi_i \cdot \phi_j$
8. Wende die Formel für die Berechnung der Korrelation an.

Übung 2-2: *Äquivalenz der drei Versionen des Modells kongenerischer Tests:*

Gegeben:
Die 3 Versionen des Modells kongenerischer Tests von Abb. 2-5, Abb. 2-8 und Abb. 2-9.:
Zeige für jedes der Modelle die Gültigkeit der Gleichungen:

$$\text{Kov}(Y_i, Y_j) = \lambda_i \cdot \lambda_j \qquad (2\text{-}35)$$

und

$$\text{Var}(Y_i) = \lambda_i^2 + \theta_{\varepsilon_i}^2 , \qquad (2\text{-}36)$$

woraus sich die Schlussfolgerung ergibt, dass die drei Modelle die gleichen empirisch prüfbaren Vorhersagen machen.

Übung 2-3: *Herleitung der Parameter des Modell kongenerischer Tests:*

Gegeben:
Das Modell dreier kongenerischer Tests (Abb. 2-37):
Leite die Berechnungsausdrücke für die Parameter aus den beobachteten Varianzen und Kovarianzen her.

Hinweis:
Verwende zur Herleitung der Ladungskoeffizienten λ_i die beobachteten Kovarianzen und leite anschliessend aus den Varianzen und den zuvor abgeleiteten Berechnungsausdrucke für λ_i die Varianzen $\text{Var}(\varepsilon_1)$ der Fehlerterme her.

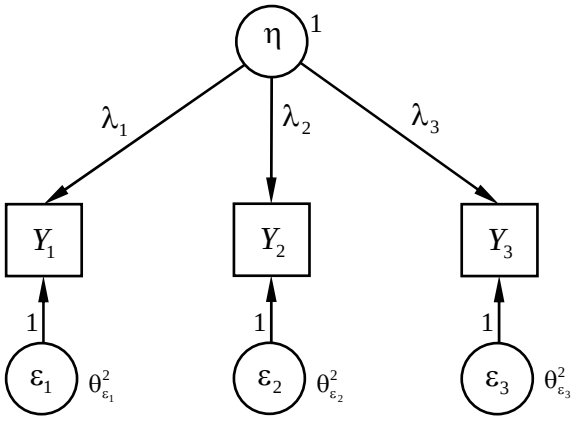


Abb. 2-37: *Modell dreier kongenerischer Tests.*

Übung 2-4*: Prüfung der unterschiedlichen Testmodelle von Jöreskog (1971):*

Prüfe die 4 Hypothesen H_1 – H_4. (siehe Bsp.2-3 und die dort dargestellten Daten).
Welche der vier Hypothesen ist zu bevorzugen. Begründe Deine Wahl.

***Übung 2-5**: Angst als situationsspezifischer Faktor (Steyer, 1989; Steyer & Eid, 1993)*

Gegeben:

Ein Tests zur Testung von Angst wurde zu zwei Zeitpunkten mit zwei Monaten Abstand angewendet.

Der Test wurde in zwei Hälften mit jeweils 10 Testitems eingeteilt und es wurde pro Testhälfte jeweils die Summe der 10 Testwerte gebildet.

Tab. 2-7 zeigt die Kovarianzmatrix für die beiden Testhälften zu den beiden Testzeitpunkten.

Die Stichprobengrösse betrug $N = 179$.

Prüfe die folgende Annahme von Steyer (1989):

1. Die beiden Testhälften sind jeweils parallel.
2. Da Angst eine Zustandsvariable und daher nicht konstant über die Messzeitpunkte hinweg ist, sind die vier Testwerte nicht kongenerisch.

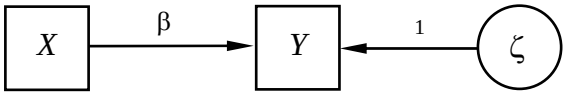
	S_1A_1	S_1A_2	S_2A_1	S_2A_2
S_1A_1	24.670			
S_1A_2	21.895	25.135		
S_2A_1	10.353	10.624	27.239	
S_2A_2	11.665	12.636	25.258	28.683

***Tab. 2-7:** Kovarianzmatrix der Testwerte der beiden Testhälften eines Angsttests, die zu zwei Testzeitpunkten angewendet wurden: S = Situation (d.h. der Testzeitpunkt), A = Testhälfte (Nach Steyer, 1989).*

***Übung 2-6:** Reliabilität und multipler korrelierter Korrelationskoeffizient:*

Gegeben:

Das einfache lineare Regressionsmodell:



$$Y = \beta \cdot X + \zeta$$
$$\mathrm{Kov}(X, \zeta) = 0$$

Zeige, dass die quadrierte Korrelation die Reliabilität der quadrierten Korrelation $R^2_{Y.X}$ zwischen Y und $\hat{Y} = \beta \cdot X$ der Reliabilität von Y entspricht:

$$\mathrm{Rel}(Y) = \frac{\beta^2 \cdot \mathrm{Var}(X)}{\mathrm{Var}(Y)}$$

Übung 2-7: *Fehlervarianz und Reliabilität:*

Gegeben:

Das lineare Messmodell von Abb. 2-11:

Zeige die Gültigkeit von: $\mathrm{Var}(\varepsilon) = \mathrm{Var}(X) \cdot [1 - \mathrm{Rel}(X)]$.

Übung 2-8: *Reliabilität einer Messung mit zwei latenten Variablen:*

Gegeben:

Das Modell von Abb. 2-38:

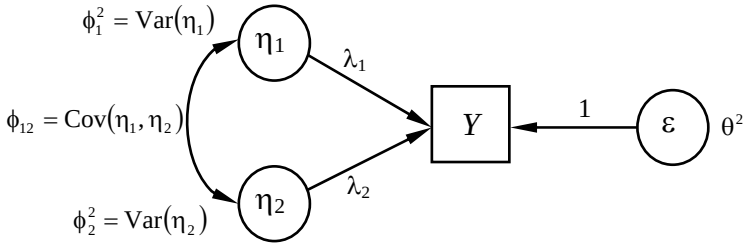


Abb. 2-38: *Kausaldiagramm des Zielmodells.*

1. Leite den Berechnungsausdruck von $\mathrm{Rel}(Y)$ als Funktion der Modellparameter her.

   Hinweis: Die Reliabilität $\mathrm{Rel}(Y)$ von *Y* ist der durch die beiden Faktoren η_1 und η_2 erklärte Anteil an der Gesamtvarianz von *Y*.

2. Verwenden AMOS (oder ein anderes Programm) und zeige, dass der hergeleitete Berechnungsausdruck von $\mathrm{Rel}(Y)$ für das Modell in Abb. 2-39 das gleiche numerische Ergebnis liefert wie das Programm.

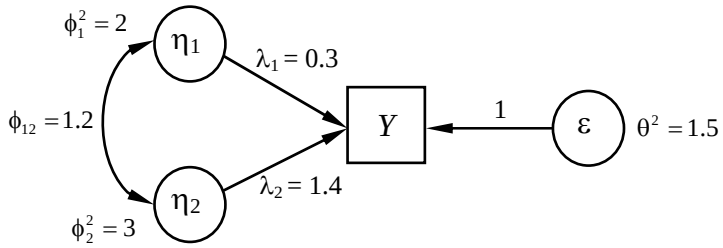


Abb. 2-39: *Kausaldiagramm des Zielmodells mit konkreten numerischen Werten zur Prüfung des Berechnungsausdrucks für die Reliabilität.*

Übung 2-9: *Für parallele Maße entspricht die Korrelation der Maße der Reliabilität der beiden Maße.*

Gegeben: Modell zweier paralleler Tests (Abb. 2-40).

Zeige die Gültigkeit von:

$$\text{Rel}(Y_1) = \text{Rel}(Y_2) = \text{Korr}(Y_1, Y_2),$$

d.h. die Reliabilität zweier paralleler Tests entspricht deren Korrelation.

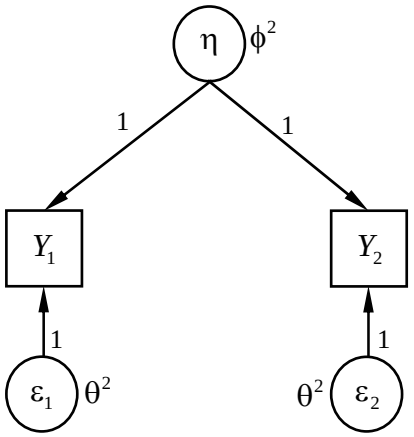


Abb. 2-40: *Das lineare Messmodell zweier paralleler Maße*

Übung 2-10: *Die Reliabilität der Summe zweier bzw. von n τ-äquivalenten Tests entspricht Cronbachs α:*

Gegeben: Modell zweier τ-äquivalenter Tests (Abb. 2-41).

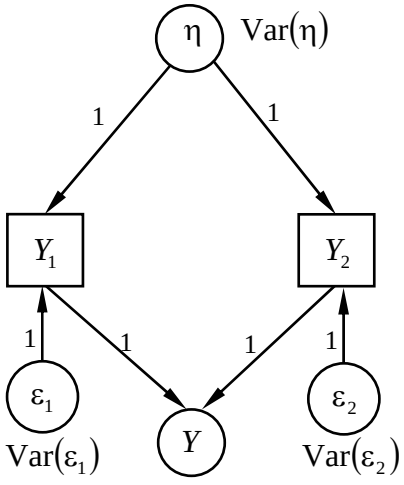


Abb. 2-41: *Modell τ-äquivalenteTests: Y_1 und Y_2 repräsentieren die beiden Tests und Y repäsentiert deren Summe.*

1. Zeige, dass die Reliabilität von *Y* korrekt durch die Formel für Cronbachs α wiedergegeben wird.

Hinweis: Gehe wie folgt vor:

(i) Zeige, dass die True-Score Varianz von Y $4 \cdot \mathrm{Var}(\eta)$ beträgt.

(ii) Zeige, dass gilt: $\mathrm{Var}(\eta) = \mathrm{Cov}(Y_1, Y_2)$

(iii) Einsetzen in die Formel für die Reliabilität:

$$\mathrm{Rel}(Y) = \frac{\text{True - Score Var}(Y)}{\mathrm{Var}(Y)}$$

und Umformung ergibt die Lösung.

2. Verallgemeinere das Ergebnis auf den Fall von *m* τ-äquivalenten Tests und zeige, dass auch in diesem Fall die Reliabilität exakt Cronbachs α entspricht.

Übung 2-11: *Cronbachs α und die Reliabilität einfacher Summen:*

Gegeben: Die Kovarianzmatrix der Testitems $X_1, X_2, \ldots, X_8$ in Tab. 2-8 (N = 165).

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	0.249	0.169	0.146	0.190	0.190	0.186	0.168	-0.001
X_2	0.169	0.251	0.135	0.172	0.172	0.166	0.148	-0.002
X_3	0.146	0.135	0.243	0.153	0.153	0.145	0.158	0.019
X_4	0.190	0.172	0.153	0.247	0.198	0.188	0.177	0.014
X_5	0.190	0.172	0.153	0.198	0.247	0.182	0.177	0.020
X_6	0.186	0.166	0.145	0.188	0.182	0.250	0.159	0.009
X_7	0.168	0.148	0.158	0.177	0.177	0.159	0.251	0.016
X_8	-0.001	-0.002	0.019	0.014	0.020	0.009	0.016	0.250

Tab. 2-8: *Kovarianzmatrix von 8 Testitems.*

1. Berechne die folgenden Grössen:
   (i) Cronbachs α für alle 8 Items.
   (ii) Cronbachs α für die ersten 7 Items.
2. Prüfe:
   (i) Ob die 8 Testitems kongenerisch sind.
   (ii) Ob die 8 Testitems τ-äquivalent sind.
   (iii) Ob die ersten 7 Items τ-äquivalent sind.
3. Berechne die Reliabilität der folgenden Summen mit Hilfe des Strukturgleichungsprogramms:
   (i) Reliabilität der Summe aller 8 Items.

(ii) Reliabilität der Summe aus den ersten 7 Items unter Annahme des kongenerischen Modells.
(iii) Reliabilität der Summe aus den ersten 7 Items unter Annahme des τ-äquivalenten Modells.

Übung 2-12: *Cronbachs α liefert für parallel Messungen eines Konstrukts das gleiche Ergebnis wie die Spearman-Brown-Formel:*

Gegeben: m parallele Messungen eines Konstrukts:

Zeige, dass in diesem Falls die Formel für Cronbachs α:

$$\alpha = \frac{m}{m-1} \cdot \frac{\sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \mathrm{Kov}(Y_i, Y_j)}{\mathrm{Var}(Y)}$$

identisch ist zur Spearman-Brown – Formel:

$$\mathrm{Rel}(Y) = \frac{m \cdot \rho}{1 + (m-1) \cdot \rho}$$

Übung 2-13: *Reliabilität der Summen von Tests, die nicht kongenerisch sind:*

Gegeben: Das Modell von Abb. 2-23 (Bsp.2-12). Alle Ladungen haben der Wert: $\lambda = 0.7$.

Berechne die Reliabilität der Summe der 10 Testwerte.

Übung 2-14: *Reliabilität der gewichteten Summen von Tests, im allgemeinen faktorenanalytischen Modell:*

Gegeben: Das Modell von Abb. 2-21 (Bsp.2-10).

Ermittle die Reliabilität der gewichteten Summe der 5 Tests mit den Reliabilitäten der einzelnen Tests als Gewichte:
(i) Mittels Methode 2-5
(ii) Mittels Methode 2-6.

Übung 2-15: *Maximale Reliabilität I:*
Reliabilität einfacher Summen vs. Reliabilität von Summen mit optimalen Gewichten für kongenerische Tests

Gegeben: Das Modell dreier kongenerischer Tests von Abb. 2-42.

Ermittle:
1. Die Reliabilität $\mathrm{Rel}(Y)$ der einfachen Summe:

$$Y = Y_1 + Y_2 + Y_3 .$$

2. Die optimalen Gewichte w_i zur Maximierung der Reliabilität.
3. Die maximale Reliabilität $\text{Rel}_{\max}$ der optimal gewichteten Summe:

$$Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + w_3 \cdot Y_3$$

mit den optimalen Gewichten w_i (zur Maximierung der Reliabilität).

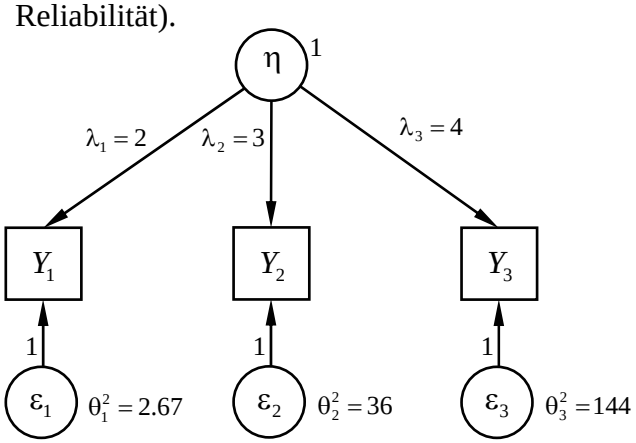


Abb. 2-42: *Modell kongenerischer Tests.*

Übung 2-16: *Maximale Reliabilität II: Reliabilität einfacher Summen vs. Reliabilität von Summen mit optimalen Gewichten für nicht kongenerische Tests:*

Gegeben: Das Modell dreier nicht kongenerischer Tests von Abb. 2-43.

Bemerkung: Abgesehen von der Tatsache, dass die Test nicht kongenerisch sind, ist das Modell identisch zu jenem dreier nicht kongenerischer Tests von Übung 2-15. Daher sind auch die Reliabilitäten der einzelnen Tests identisch.

Ermittle:

1. Die Reliabilität $\text{Rel}(Y)$ der einfachen Summe:

$$Y = Y_1 + Y_2 + Y_3 .$$

2. Die maximale Reliabilität $\text{Rel}_{\max}(Y)$ der optimal gewichteten Summe:

$$Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + w_3 \cdot Y_3$$

mit den optimalen Gewichten w_i (zur Maximierung der Reliabilität).

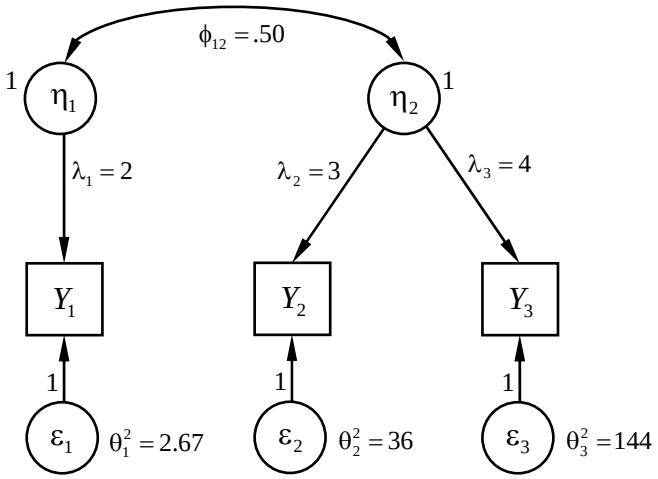


Abb. 2-43: *Modell von nicht kongenerischen Tests.*

Übung 2-17: *Maximale Reliabilität III:*

Ermittle die maximale Reliabilität des faktorenanalytischen Modells von Abb. 2-27 (Bsp.2-17), sowie die zugehörigen optimalen Gewichte.

Übung 2-18: *Eindeutige Validitätsvarianz:*

Gegeben: Das Modell dreier nicht kongenerischer Tests von Abb. 2-44.

Ermittle die eindeutige Validitätsvarianz des Indikators für die drei Faktoren.

Bemerkung:

Das Modell in Abb. 2-44 unterscheidet sich von jenem in Abb. 2-31 nur in den erhöhten Kovarianzen zwischen den latenten Faktoren. Daher sollten die eindeutigen Validitätsvarianzen geringer sein als in Bsp.2-20.

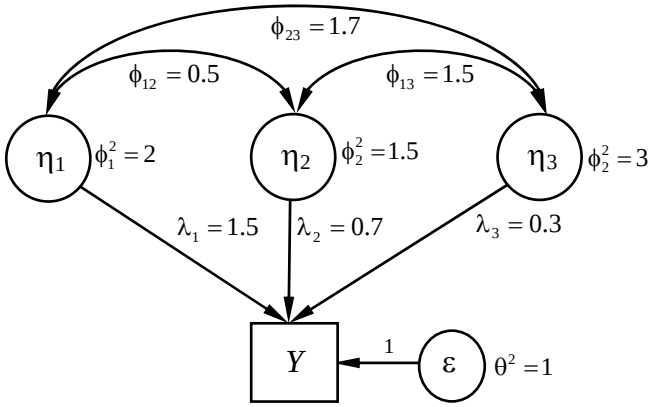


Abb. 2-44: *Strukturelles Modell zur Ermittlung der eindeutigen Validitätsvarianz.*

3. Literatur

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University Press.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71,* 425–440.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203-219.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061-1071.

Camstra, A. & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis. *Sociological Methods & Research, 21*, 89-115.

Fine, A. (1984). The natural ontological attitude. In J. Leplin (Ed.), *Scientific realism* (pp. 83-107). Berkeley: California University Press.

Glymour, C. (1986). Statistics and metaphysics. *Journal of the American Statistical Association, 81,* 964-966.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric measures. *Psychometrika, 36,* 109-133.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319-342.

Kitcher, P., & Salmon, W. C. (1989). *Scientific explanation: Minnesota studies in the philosophy of science* (Vol. 13). Minneapolis: University of Minneapolis Press.

Medin, D. L. & Ortony, A.J. (1989). Psychological essentialism. In: S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 179-195). New York: Cambridge University Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison Wesley.

Pais, A. (1993). *Niels Bohr's times: In physics, philosophy and polity.* Oxford: Clarendon Press.

Popper, K. R. (1959). The propensity interpretation of probability. *British Journal of the Philosophy of Science, 10*, 25-42.

Popper, K. R. (1989). *Die Logik der Forschung* (9. Auflage). Tübingen: Mohr.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199-223.

Singh, S. (2005). *Big Bang: der Ursprung des Kosmos und die Erfindung der modernen Naturwissenschaft.* München: Hanser.

Van Fraassen B. (1980). *The scientific image.* Oxford: Oxford University Press.

Watkins, J. W. N. (1984). *Science and scepticism.* Princeton, NJ: Princeton University Press.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In: C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics, Vol.26: Psychometrics* (pp.45-79). Amsterdam: Elsevier.