Modern Psychometrics

A Modeling Approach

Author: Siegfried Macho

Fribourg, 2019

Contents

1. Introduction: Characterizing Modern Psychometrics

An important activity of psychologists consists in the measurement of mental characteristics. Typical examples of mental characteristics are:

- Intelligence and other cognitive capabilities;
- Personality traits;
- Mood;
- Preferences and opinions;
- Knowledge concerning specific topics;
- Psychic and mental disorders.

A crucial aspect of the measured characteristics consists in the fact that they cannot be observed directly. Rather they have to be inferred on the basis of observed measures. Due to this feature they are called *latent (mental) constructs*. There exists a great variety of psychological measure and measurement methods, respectively. Here are a few examples:

- Classical psychological tests, like intelligence and personality tests;
- Questionnaires;
- Implicit Tests, i.e. tests employing measures (e.g. reaction times) that appear to have no association with the to be measured characteristics (e.g.racist attitude);
- Behavioral measures (e.g. behavioral measures indicating the aggression potential);
- Exams for assessing knowledge and/or abilities;
- Neuropsychological tests (e.g. EEG or fMRI);
- Physiological test (e.g. measuring skin conductance).

Psychometrics is concerned with the measurement of latent constructs. The discipline arose with the development of mental ability tests, pioneered by the French Alfred Binet (1857-1911).

Historical Remark 1-1: *Mental ability testing:*

The history of mental ability tests dates back to the beginning of the 20 century. In 1904 the French minister appointed a commission to recommend procedures for identifying intellectually retarded children.

In 1905, Binet and Simon developed the first intelligence test consisting of 30 Items. Its aim was to identify retarded children requiring specific scholastic training.

In 1916, the Stanford-Binet intelligence test was constructed. This test is based on the Binet-Simon test. In the years 1937, 1960, 1986 and 2003, revisions of the Stanford-Binet tests have been provided.

For further details, cf. Kaplan & Saccuzzo (2017).

The first theoretical latent variable model of mental abilities is due to Charles Spearman (1863-1945) who assumed a general ability factor *g* and different specific factors.
Both types of factors are latent construct. Yet, whereas the general factor *g* is involved in each cognitive task the specific factors are relevant for specific tasks only (e.g. for performing verbal or spatial tasks).

To get a better idea of the importance of psychometrics it is important to illustrate the specific problems that are associated with the measurement of latent constructs.

1.1 Problems Associated with the Measurement of Latent Constructs and the Importance of Psychometrics

All branches of psychology are concerned with the measurement of latent constructs. For example, memory researchers are interested how different manipulations influence memory performance. However, memory and memory performance are theoretical constructs that cannot be observed directly. Thus various measures of memory performance are employed. The most common measures are performance in free recall and in recognition.
Due to the fact that observed scores are employed for measuring underlying constructs (and are relevant only in this respect) the following two considerations are important:

1. The obtained test scores have to be distinguished sharply from the mental constructs the test intends to measure. For example, Examinee 1 might have a higher test score then Examinee 2 despite the fact that the latter is ranked higher with respect to the mental construct.
2. Measurements are error prone. This is due to the fact that the scores are not only influenced by the measured construct but by additional causal factors (e.g. motivation, fatigue, mood, variations in the level of attention, errors of instruments).

Comment 1-1: *Ignoring and minimizing measurement errors:*
The importance of measurement error differs between various branches of psychology. For example, in experimental psychology the problem of measurement errors plays a less prominent role than in other branches (cf. Section 1.2).
This is due to the fact that experiments enable a great deal of control compared to observational studies. Due to this control and repetitions of trials measurement errors may be minimized. For example, with reaction time (RT) experiments trials are repeated many times and the means of the RTs are evaluated.

Summing scores or computing means is a general method for reducing measurement error. This method is also used in the context of the application of psychological tests and of questionnaires.

Psychometrics is concerned with the structure of latent constructs and their measurement. By consequence the following issues have to be addressed:

- *Nature of the latent constructs and their relationship:* Is the measured construct uni- or multidimensional? Which of these dimensions are measured by a specific test? How is the target construct related to other latent constructs?
- *Nature of the response functions:* How are values on the latent constructs mapped onto responses?
- *Quality of the measurement instrument:* Concerns issues like the precision of the instrument and the degree of measurement error, respectively, or the problem of biases.
- *Target of the measurement:* Does the measurement instrument measure the target construct only or are other mental construct also captured?
- *Occasion and method specific influences:* How is the result of measurement influenced by occasion specific influences and by the specific method employed?
- *Fairness of the measurement instrument:* Is the test fair or are different groups (man or women, Black or White) discriminated?
- *Measurement invariance:* Is the measured construct gauged in the same way in different groups of examinees?
- *Conclusions from observed values to values on the latent constructs:* Given a specific value of the measurement. How can one predict the value on the measured latent construct?

1.2 Psychometrics and the two Disciplines of Scientific Psychology

In his presidential address as the president of the American Psychological Association, Cronbach (1957) talks of two disciplines of scientific psychology which he terms »experimental« and »correlational«. According to this view the experimental branch is concerned with variation of treatments (independent variables) whereas the correlational branch is focused on variance due to individual differences which is treated by the experimentalists as a nuisance.

Comment 1-2: *Brain imaging and the neglect of individual differences*

The tendency to ignore individual differences seems to be practiced by brain imaging methodology, too. The mapping of activations is usually performed with respect to a standard brain thus ignored individual differences in brain structure and functioning.

According to Cronbach's distinction psychometrics makes up a part of the correlational discipline. This can be concluded from the fact that psychometrics models are used to assess differences on the latent construct between examinees. In addition, psychometric methods are multivariate, i.e. the latent construct is measured using more than one measure. This is regarded as the only useful way to gauge latent constructs adequately since the latent mental construct results in different realizations in different situations. By contrast the experimental branch is mainly concerned with single outcome variables. Moreover, in case of taking multiple measures, their multivariate structure is frequently not handled adequately.
The separation of the two psychological traditions results in the neglect of interactions between treatments and individual characteristics. Obviously, individual characteristics moderate the effect of treatments manipulated by experimentalist, like instruction, training, or therapeutic interventions.
The separation between the two disciplines is also reflected by differences between cognitive and psychometric models. There have however been some attempts of unifying both types of approaches (cf. the discussion in Section 3.4).

1.3 On the Distinction between »Old« and »New« Psychometrics

According to a common view psychometrics may be classified into »old« and »new« psychometrics with classical test theory (CTT) being regarded as »old« and item response theory (IRT) as »new« psychometrics. For example Embretson and Reise (2009, Chapter 2) make a distinction between »old« and »new« rules of measurement (cf. Table 2.1 on page 15 of Embretson and Reise) exactly along this lines.

Comment 1-3: *Old rules of measurement (Embretson & Reise, 2009):*

Their specification of the old rules by Embretson and Reise is, in part, problematic. Consider, for example, their old Rule 2:

Longer tests are more reliable than shorter tests.

However, nobody would really accept this rule since it is well-known that a shorter test with more reliable items can be more reliable than a test with more but less reliable items.
In addition, adding reliable test items to an existing test can result in a decrease of Cronbach's alpha, a commonly used measure of the reliability of the sum of test items (Li, Rosenthal, and Rubin, 1996).

The identification of CTT with »old« and IRT with »new« psychometrics has become problematic with the representation of the classical test models by means of structural equation models (Jöreskog, 1971) [for details, see below, Chapter 4.3]. In fact, is can be shown that the classical test models can be conceptualized as specific cases of item

response models. The only difference consists in the fact that different observed measures are modeled: Means and (co-) variances in case of classical test models and probabilities of different response categories in case of item response models. By consequence, the main difference of the two types of models consists in the usage of different response functions that map values of latent constructs as well as item characteristics on observed responses (cf. the general psychometric model in Chapter 3).
Thus a better characterization of the difference is that of Embretson (2010a) between procedures based on sum scores and item response models that model single items.

Comment 1-4: *On item response models*
I am not sure whether Embreston considers structural equation models as item response models.

According to the present view, the main difference between »old« and »modern« psychometrics consists in a radical different attitude concerning the measurement model underlying the observed measures, specifically:

- »Old« psychometrics does not care about details of the measurement model and how the psychometric constructs like reliability or validity depend on the underlying model.
- In »modern« psychometrics measurement models and their properties are specified in detail. The dependences of psychometric constructs on the underlying measurement model are made explicit.

The different attitudes with respect to the measurement model representing the measurement process results in an entire different practice, with the »old« psychometrics being characterized by the following features:

- The theoretical status of psychometric constructs, like reliability or validity, is not recognized. Specifically, it is not realized that these quantities can be measured (and get their significance) only with respect to an underlying theoretical model that is assumed to represent the measurement process correctly.
- By consequence, the theoretical constructs are often identified with its measures, e.g. validity is identified with the validity coefficient.
- In general, the practice of the psychometrician consists in computation of coefficients that are computed on the observed responses. Typical examples are Cronbach's α or the Spearman-Brown coefficient using observed variances and covariances or correlations. These coefficients represent reliabilities of sum scores.
- The conditions for application of these coefficients are simply ignored and, in most cases, unknown to their users.

- Latent scores are estimated, and replaced respectively, by the sum or mean of the observed scores.
- Coefficients are used to perform certain adjustments, e.g. using reliability coefficients to correct for attenuation.
- General recommendations are provided that turn out to be wrong in its generality. A typical example is the recommendation: *Cronbachs α (or the Spearman-Brown coefficient) generally underestimate the true reliability.*
- Pseudo-paradoxes have been identified that do not exist under a strict latent variable conception. An example is the *validity-reliabilty paradox* that states that an increase in the reliability can result in a decrease of the validity of a measure.

In contrast, modern psychometric is characterized by the following attributes:

- The psychometric model representing the measurement process makes up the central part in that all relevant quantities (reliability etc.) get their significance with respect to the underlying measurement model only. If the model is not (approximately) correct these quantities are problematic.
- The model provides an analysis of the structure of a test. By consequence, more sophisticated estimators of reliability and validity based on the structural analysis of the test can be computed.
- The theoretical status of concepts like reliability is made obvious. The concept of construct validity (Cronbach & Meehl, 1955) receives its full appreciation.
- Coefficients are computed from model based predictions and not from observed measures resulting in improved estimates of the underlying theoretical constructs (in case of the measurement model being approximately correct).
- The prerequisite of the correct application of coefficients, like Cronbach's α, as well as their limits are made explicit.
- The limits of recommendations, like *Cronbach's α (or the Spearman-Brown coefficient) generally underestimate the true reliability* can be demonstrated.
- Assumed paradoxes can be shown to disappear.
- No adjustments (like correction for attenuation) are required since all relationships are represented within the measurement model and can be estimated in case of the model being correct.
- The computation of latent construct scores of individual examinees depends on the underlying model.
- In general, the empirical adequacy of the measurement model can be tested.

The actual presentation recognizes the superiority of modern psychometrics. Therefore, the following exposition puts measurement models

in the center of the considerations. Derivations of relevant coefficients are based on the specified models.

2. On the Significance of Theoretical Constructs

In this chapter we first consider the function of theoretical constructs in the natural sciences, and why theoretical constructs are required. Next, we examine the significance of psychological constructs in scientific psychology. An important aspect of our discussion concerns the differentiation between constructs and pseudo-constructs.

2.1 On the Function of Theoretical Constructs in the »Hard« Natural Sciences

The »hard« natural sciences, like physics of chemistry, are crowded with theoretical constructs. Typical examples are: electrons, protons, electromagnetic field, molecules, amino acids, double helix, enzymes, etc. The postulation of theoretical constructs seemingly opposes to Occam's razor, a principle of rationality that is of great importance in science (and also in everyday reasoning).

Principle 2-1: *Occam's razor [William of Occam (1288-1347)]*

Entities should not be multiplied without necessity.
»Entia non sunt multiplicanda sine necessitate.«

By consequence there have to be strong reasons for the introduction of new theoretical constructs. In fact there was an influential philosophical tradition that tried to ban theoretical constructs altogether from science. Within psychology, this tradition was associated with the label *behaviorism*. However, finally, it became obvious that a mature science cannot exist without theoretical entities, and that they cannot be reduced to observed quantities.
The main function of theoretical constructs consists in providing explanations of empirical phenomena. Theoretical constructs are especially important in the following two contexts: (a) The explanation of seemingly divergent observations, and (b) the protection of existing well-confirmed theories. However, one of the most important functions of theoretical constructs is cognitive in nature: It enables scientists to draw conclusions and to make new assumptions.
Let us consider these three functions of theoretical constructs in greater detail.

2.1.1 Theoretical constructs and the explanation of divergent observations

Great achievements in science are associated with the detection of a common mechanism that enables an explanation of seemingly divergent observations. Prototypical cases of this sort of *unification* are Isaac Newton's (1643-1727) theory of the gravitational force or James Clerk Maxwell's (1831-1879) theory of the electromagnetic field.

The theoretical construct of (gravitational) force enabled Newton to explain Galilei's (1564-1642) laws of free fall as well as Kepler's (1571-1630) laws of the movements of the planets. In fact, Newton's dynamical theory was able to explain the movement of physical bodies in general. At the heart of Newton's theory we find the theoretical construct of force acting upon physical bodies. The assumption of such a force together with the equations quantifying its effects enable the prediction of the movement of physical bodies under different conditions.

Similarly, Maxwell's assumption (and mathematical description) of an electromagnetic field and the field equations enabled him to explain and describe all sorts of magnetic and electric phenomena discovered by Michael Faraday (1791-1867) and others, like the phenomenon of electromagnetic induction.

Thus, the introduction of theoretical constructs and the associated mechanism and relationships (usually represented by (differential) equations) enable a unified and parsimonious explanation of various seemingly divergent phenomena. One important aspect of theoretical constructs consists in the fact that they enable the *prediction of new phenomena*. For example, Maxwell's theory of the electromagnetic field predicted the existence of electromagnetic waves that have been discovered only years later by Heinrich Hertz (1857-1894). Due to these characteristics of latent constructs (parsimonious unified explanation and the prediction of new phenomena) theoretical constructs have been postulated in the face of Occam's razor.

Let us now take a look at the second important reason for postulating theoretical constructs.

2.1.2 Theoretical constructs and the protection of existing theories

A second situation provoking the postulation of theoretical constructs is given if a newly detected phenomenon seemingly contradicts a well-established theory. A classic example of this sort of situation was the postulation of the existence of the Neutrino by Wolfgang Pauli (1900-1958) in the year 1933. The reason for the introduction of this particle consisted in saving the well-established law of the conservation of energy that was seemingly violated by results concerning the radioactive beta decay. Independent evidence concerning the existence of Neutrinos was provided in 1956, 23 years after its postulation.

Let us now turn to the cognitive function of theoretical constructs.

2.1.3 On the cognitive function of theoretical constructs

If a theory has been completely formalized in terms of mathematical equations the theoretical constructs function as variables (or placeholders) within these equations. Their meaning is no longer required

for making precise predictions. Thus it is completely irrelevant, with respect to prediction, whether a theoretical construct is denoted by the term *force* or simply be the variable name *x*. In fact the predictive power of theoretical construct is contained in its relations to other constructs as well as to observed phenomena. Even if these relations are only qualitative in nature allowing only for categorical or ordinal predictions, the theoretical construct itself is not relevant for the predictions. Accordingly, in well-developed sciences, the theoretical constructs seem to lose their significance as soon as the relations to other constructs and empirical phenomena have been specified precisely.
However, theoretical constructs serve an important function as a vehicle for promoting new ideas and theories. In mature sciences, scientists are thinking and theorizing in terms of theoretical constructs. They communicate, draw conclusions, and use analogical reasoning (and perhaps other sorts of inductive reasoning) in terms of the content of the theoretical constructs. This sort of reasoning results in new theories and new applications of existing theories thus promoting science (see, for example, Dunbar, 1987). The density of the (nomological) network of theoretical constructs as well as the precision of the specified relationships may be important for the successful development of new ideas.

Comment 2-1: *Density and precision of the nomological networks and the restrictions of theories*

The greater the density of the network of theoretical constructs and the higher the precision of the specification of the relationships between constructs the strong the mutual restriction of theories from different branches.

Thus, in a BBC talk on the philosophy of science the philosopher of science, Hillary Putnam, compares science with a jigsaw puzzle where different pieces have to fit together (cf. https://www.youtube.com/watch?v=kH785oawwkk).

The importance of the semantic content and interpretation of theoretical constructs, respectively, is also evidenced by the fact that theories which cannot be given a consistent interpretation, as it is the case with Quantum theory, are experienced as not completely satisfactory. This is true despite of the fact that the theory is extremely successful in making (new) predictions.

To summarize, useful theoretical constructs have three valuable functions: First, they lie at the heart of unifying theories that provide a parsimonious explanation of a great variety of empirical phenomena, and the prediction of new phenomena. Second, they enable the protection of well-established theories in the face of new refuting empirical evidence. Finally, they perform an important cognitive function enabling researchers to communicate and to draw inferences.

Let us now take a look on the function of theoretical constructs in psychology.

2.2 Theoretical Constructs in Psychology

Theoretical constructs are found in everyday as well as in scientific psychology. In both cases they have similar functions as in the »hard« sciences: Unified explanation, prediction, and communication. Here is a preliminary definition of the concept:

Concept 2-1: *Psychological constructs:*

Psychological constructs are mental structures and processes that cannot be observed directly. They are postulated in order to provide a parsimonious explanation of observed behavior.
Cf. Principle 2-2 (page 14) for a detailed exposition of the concept of *theoretical constructs in scientific psychology*.

Let us have a look on some examples of theoretical constructs of folk and of scientific psychology.

Ex. 2-1: *Psychological constructs used in folk and scientific psychology*

Theoretical constructs used in everyday psychology refer to internal (mental) states and dispositions of traits, e.g.:

Anger, pleasure, depressiveness, frustration, cleverness, amorousness, aggressiveness, anxiousness, pain, memory etc.

Some of these constructs can be found in scientific psychology, too. In addition, scientific psychology has postulated a number of new and more sophisticated constructs, e.g.:

Fluid Intelligence, working memory, procedural memory, semantic network, judgmental heuristics, cognitive dissonance, executive functions, etc.

Some of these scientific constructs, like fluid intelligence or procedural memory, are developments of constructs of folk psychology. Others, like semantic networks, judgmental heuristics do not possess a counterpart in everyday psychology.

Obviously, theoretical constructs serve a similar function in everyday and in scientific psychology as in the natural science: providing parsimonious explanations of diverse empirical phenomena and predicting future events. In the present case the empirical phenomena consist in peoples' overt behavior. Personal traits and/or internal states are used for accounting this behavior.

The shortcomings of folk psychology consist in the fact that its predictions are, in many cases, either imprecise or not correct. In fact, everyday psychology is well apt to provide post-hoc explanation of behavior thus delivering a feeling of understanding and control. However, the

theory usually fails to provide precise predictions of human behavior under specific conditions.

Ex. 2-2: *Wrong predictions of everyday psychology concerning the outcome of the Milgram experiment*

Previous to his experiments Milgram asked a number of psychiatrists to rate the percentage of subjects who would apply the maximum dosage of shock. The estimates were located around 1% (the observed percentage was about 65%).

There are two main reasons for these shortcomings of everyday psychology:

1. The constructs are pseudo-constructs (cf. below, Section 2.3).
2. The network of relationships involving constructs and observed behavior has not been specified correctly or it has not been specified precisely enough to enable predictions, i.e. theories and psychological laws endorsed by lay persons are frequently too imprecise and not correct, respectively.

In response to the shortcomings of everyday psychology, scientific psychology tries to develop existing constructs further by refining them (i.e. exploring various facets of these constructs), and by embedding them into a dense and precise network of (possible new) latent constructs (cf. Comment 2-2 on p.19). Ideally, this should result in formal models that enable precise prediction of behavior. Most psychometric as well as many cognitive models are *parametric statistical models* and thus formal models that enable precise predicttions (cf. Chapter 3).

Similar to theoretical concepts in physics a new theoretical concept in psychology gains plausibility if it permits the prediction of new surprising phenomena.

Ex. 2-3: *Cognitive dissonance and the prediction of a surprising behavior*

The theory of *cognitive dissonance* assumes that people try to keep their beliefs, thoughts, and other pieces of knowledge consistent. Otherwise they experience a sort of cognitive tension or cognitive dissonance. This tension is experienced as inconvenient, specifically if the beliefs are closely related to the conception of the self. By consequence people try to reduce the cognitive dissonance.

Peoples' tendency to reduce cognitive dissonance leads to surprising behavior. For example, in an experiment of Festinger and Carlsmith (1959) participants were asked to announce a perfectly stupid and boring experimental task to other subjects as being interesting. Those participants who received no payments for their wrong report assessed the experiment as being more interesting than those subjects that received a payment.

This can be explained by participants' attempts to reduce their cognitive dissonance between their honest self and the fact that they betrayed their colleagues without receiving any reimbursement. By changing their belief that the experiment was not that dull they were able to reduce their cognitive dissonance.

Surprising behavior due to peoples' attempts to reduce cognitive dissonance was observed in different contexts (see, e.g., Aronson, Wilson & Akert, 2010).

One important issue concerns the differentiation between constructs and pseudo-construct. We now turn to this problem.

2.3 On the Nature of Theoretical Constructs and the Problem of Pseudo-Constructs

The *scientific realism debate* in the philosophy of science contrasts two different conceptions concerning the nature of theoretical constructs.

Concept 2-2: *Realistic vs. instrumentalistic conception of theoretical constructs:*

According to the *realistic view* theoretical constructs refer to existing entities that represent the »deep structure« of the observed phenomena.

According to the *instrumentalistic view* constructs are pure inventions that need not have any counterparts in reality. Their function consists in providing precise and correct prediction and explanation of empirical phenomena.

Intuitively, the realistic conception of theoretical constructs seems to be more convincing. However, as noted by the philosopher of science Van Fraassen (1980), in practice the capacity of a theory to predict new and surprising phenomena is one of the most important features of a successful theory (see also Lakatos, 1978). By contrast, whether a theory can be given a consistent interpretation is of secondary importance only. The best example is modern quantum theory that was extremely successful in providing precise predictions. However, up to these days, no consistent interpretation of quantum theory has been provided. Note also that the existence of entities underlying theoretical

constructs cannot be proved since correct predictions of a theory do not necessarily imply that the theory is correct and, thus, the underlying constructs really refer to existing entities.
Within the realm of psychology, the realistic position is the dominant one (Borsboom, Mellenbergh, & Van Heerden, 2003). This raises the following question: What are the entities psychological theoretical constructs are referring to (see also Concept 2-1 on page 11).

Principle 2-2: *Scientific psychological constructs and functional modules:*

Psychological constructs do not refer to neurological structures. Rather, they are conceptualized as *functional* entities or processes, i.e. processing modules that perform certain functions in order to solve specific information processing problems.
These functional models are the building blocks for explaining observed behavior
The functional modules may be composed of more elementary functional modules resulting in a hierarchy of psychological processes and modules that are based on elementary psychological processes that are implemented directly within the brain. The elementary psychological processes make up the *functional architecture* (Pylyshyn, 1984).

Functional modules should not be confused with brain processes. In fact complex functional modules may be distributed over different regions of the brain

The following example illustrates the case.

Bsp.2-1: *The theoretical construct of working memory:*

The working memory is conceived of as a cognitive (functional) unit that enables the temporal maintenance and manipulation of information under interference (see e.g. Miyake & Shah, 1999).
The working memory consists of various components: executive control and memory puffers for different types of information: the phonological loop, the visual sketchpad, as well as an episodic puffer (Baddeley, 2000).
The working memory is not located in a certain region of the brain. Rather subunits are distributed over different locations.

The existence of theories involving theoretical constructs raises an important issue:

Issue 2-1:
How can we discriminate between significant theoretical constructs and nonsensical ***pseudo-constructs****?*

The history of the natural sciences reveals that a number of constructs have been postulated that turned out later to be nonsensical. The most famous example is the *luminiferous aether* that was assumed as the medium within which electromagnetic waves spread out. Another example is the hypothetical *phlogiston* that was assumed to dissipate in the curse of burning. In Biology, *vital forces* specific to living things has been postulated.

Modern physics postulates theoretical constructs whose status is, at present, unclear. Well-known examples are *dark matter* and *dark energy*. The same was true for the *Higgs Boson* whose existence has been confirmed on July 2012 by the Large Hadron Collider at CERN.

The history of psychology is full of pseudo constructs. An early example provides the Greek concept of the four fundamental personality types: *sanguine*, *choleric*, *melancholic*, and *phlegmatic*. It has been assumed that these types are associated with the prevalence of specific body fluids: *blood*, *yellow bile*, *black bile*, and *phlegm*.

Another source of pseudo constructs provides psychoanalysis, e.g. the concept of different stages of development: *oral*, *anal* and *oedipal*. Similarly, the dynamic theory of the human personality with different forces acting in opposite directions may be conceived of as a pseudo construct.

A more modern example of questionable constructs may be found in intelligence research.

Ex. 2-4: *Possible pseudo constructs in intelligence research:*

The following intelligence constructs are assumed to be questionable (see, e.g., Flynn, 2009; Rost, 2013):

- ❑ Multiple intelligences
- ❑ Emotional intelligence
- ❑ Spiritual intelligence

Let us now return to Issue 2-1 and the problem how one can discriminate between useful psychological constructs and pseudo constructs. Previously to considering useful criteria of discrimination it is useful to examine how the difference between useful constructs and pseudo constructs is conceptualized within the realistic and instrumentalistic position:

- ❑ For a proponent of a realistic position, a theoretical construct is useful if it refers to an existing entity otherwise it has to be conceived of as a pseudo construct.
- ❑ For a proponent of an instrumentalistic view, a theoretical construct is useful if it enables parsimonious explanations of existing and the prediction of new surprising phenomena.

The conception of the realistic view does not provide a workable criterion since there is no way to decide definitively whether an entity a construct refers to really exists. By contrast, the conception related to

the instrumentalistic view is more fertile for finding a criterion of discrimination.

Principle 2-3: *Criteria for assessing the usefulness of theoretical constructs*

Theoretical constructs are useful if they are associated with theories that enable the parsimonious explanation and the prediction of empirical phenomena. The degree of the usefulness depends on the following characteristics:

1. The breadth of the range of application, and the degree of parsimony of the explanation, respectively;
2. The degree of surprise of correct predictions;
3. The precision of the prediction which relates to the difficulty of the possibility of refuting the theory;
4. The degree of uniqueness of an explanation, i.e. there do not exist other well established constructs and theories that are able to explain or predict the phenomena in question.
5. Constructs *cut the nature at its joints*. This means they refer to a original realm (realistic position) or the allow for parsimonious and convenient explanations as well as for precise predictions of a specific set of phenomena (instrumentalistic view). By contrast pseudo-constructs *cut across different realms*.

In general, theoretical constructs that are embedded in a dense web of theoretical constructs with precisely specified relations between theoretical constructs and observations are the candidates of useful constructs.

By contrast, pseudo-constructs are irrelevant for the explanation of empirical phenomena since these are explained in a superior way by theories that are associated with other constructs. In addition, these constructs are not or only weakly linked to other generally accepted theoretical constructs.

The specified criteria are not *sufficient* for establishing the usefulness of theoretical constructs. However, they provide a good indication that a scientific construct may be useful. It is always possible that with the development of a new improved theory a construct considered as useful may turn out to be a pseudo-construct.

Ex. 2-5: *Identification of pseudo-constructs in psychology:*

1. The main problem concerning multiple intelligences consists in the fact that they refer to abilities (like musical abilities or motor abilities) that are not part of the intellectual capabilities that are considered as part of intelligence (Flynn, 2009, Rost, 2013).

Thus, these concepts do not *cut nature at ist joints*. They refer to quite divergent phenomena that do not enable for parsimonious explanations.

The same argument applies to other intelligence concepts like emotional or spiritual intelligence. These are simple examples of the inflationary misuse of the intelligence concept.

2. Modern social psychology, specifically the theory of automatisms (cf. Bargh, 1994, 1997), provides an alternative and more stringent explanation to phenomena accounted for by Freud‘s dynamic unconsciousness.

It has been argued above that the plausibility and usefulness of a theoretical construct increase if it is part of a theory that enables the unification of partial theories from different domains thus permitting the parsimonious explanation of a number of phenomena (like Newton’s concept of force or Maxwell’s electromagnetic field). This raises the issue of the existence of similar concepts in psychology enabling the explanation of phenomena from different domains.

It is fair to say that, contrary to physics, psychology does not dispose of high level unified theories. In fact psychology is populated with mini-theories associated with different domains (cf. the complaint of Cronbach, 1957). By the beginning of the nineties a sort of unification has been achieved in the realm of intelligence research.

Ex. 2-6: *Fluid intelligence und working memory capacity:*

In a detailed and sophisticated theoretical analysis Carpenter, Just, and Shell (1990) elucidated that working memory capacity is an important factor in the successful solution of Raven’s progressive matrices.

Subsequently, numerous empirical studies revealed a close relationship between the theoretical construct of fluid intelligence (Ackerman, Beier, & Boyle, 2005; Beier & Ackerman, 2005; Kane, Hambrick & Conway, 2005; Oberauer, Schulze, Wilhelm, & Süss, 2005).

This finding may be conceived of as a successful unification of results from two different domains (the correlational and the experimental domain).

Comment:

There have been previous attempts to bring the concept of general intelligence into alignment with results from experimental psychology, for example:

- General intelligence as the ability to solve complex problems (Dörner, 1983).

- General intelligence and reaction times for elementary decision tasks (Jensen, 1978).

These attempts were less successful since the observed relationships were not particularly strong.

As noted, theoretical constructs that are associated with theories permitting the prediction of a great variety of observations are of great significance. In psychology theoretical constructs are sometimes justified by the fact that they enable the prediction of crucial life events.

Ex. 2-7: *Theoretical constructs and the prediction of crucial life events:*

1. The construct of *general intelligence* turned out to be a predictor of success in various domains of life (Herrnstein & Murray, 1996), for example:
   - Success in school, academic studies, and profession;
   - Health status and life span;
   - Probability of unwanted pregnancy;
   - Probability of successes of psychotherapy.
2. Various personality traits, like low level of *neuroticism*, and high levels of *agreeableness* or *conscientiousness* are associated with satisfaction in the partnership of the partners (Dyrenforth, Kashy, Donnellan & Lucas, 2010).
3. The concept of will power (self-control) exhibits a relationship (positive or negative) with the following relevant aspects of life (Mischel, 2015):
   - Drug abuse;
   - Financial problems;
   - Obesity, high blood pressure, high level of cholesterine;
   - Antisocial and criminal behavior;
   - Aggression, impulsivity, hyperactivity;
   - Power of concentration;
   - Persistence in the prosecution of objectives.

These relationships are of practical interest. However, they do not necessarily indicate the existence of a significant theoretical construct. Specifically, in most cases simple correlations between measures of the constructs and the different life events are computed without specifying in detail the network of involved constructs and the relation of the latter to observed measures.

Comment 2-2: *Theoretical constructs and construct validity:*

Lee Cronbach and Paul Meehl introduced the concept of *construct validity* (Cronbach & Meehl, 1955). They make clear that the evaluation of a theoretical construct comprises a specification of the measures used but also a determination of the *nomological network* of constructs their relationship to the observed measures as well as possible relationships between measures.

Due to a lack of understanding, the concept as well as the associated ideas received little resonance within the scientific community (Kane, 2001). Even today the concept is frequently used erroneously (cf. the discussion in Chapter 4.5).

Let us summarize the basic considerations concerning theoretical constructs and the associated problems:

1. Theoretical constructs are at the heart of theories that enable one to explain and predict empirical phenomena. The theories describe relationships between theoretical constructs and between constructs and observational entities.
2. Theoretical constructs have an important cognitive function: Researchers communicate, think, and reason in terms of theoretical constructs.
3. A theoretical construct is the more versatile the more relationships it enters with other constructs and empirical phenomena, i.e. the denser the nomological network the construct is involved. As a result, more empirical phenomena can be explained.
4. Another characteristic of a useful theoretical construct consist in its potential to predict new surprising empirical phenomena.
5. An important scientific progress consists in unification, i.e. if theoretical constructs from different domains of the discipline turn out to be the same. Unification is associated with a connection of different nomological networks.
6. The precision of the characterization of a theoretical construct as well as the precision of the specification of its relationship are crucial as well since the more precise these specifications the more exact the predictions based on the theoretical construct.
7. The quality of the nomological network, i.e. the number and precision of the relationships of the constructs to other constructs and to empirical phenomena as well as the resulting breadth and precision of explanations are a good indicator of the presence of a significant theoretical construct compared to a pseudo-construct.
8. The criterion separating constructs from pseudo-constructs is not a strict one. In fact, no such criterion does exist since a theory can always turn out to be wrong and may be replaced by a new and

better one. As a consequence, theoretical constructs associated with the old theory may turn out to be wrong.

This ends our discussion of theoretical constructs that play a crucial role in science in general and specifically in modern psychometrics. Previously to discussing specific models, we next consider the general structure of a psychometric model.

3. The Structure of Psychometric Models

The present chapter presents the general structure that is common to most psychometric models. The psychometric model is at the center of modern psychometrics since each of the tasks concerning psychometrics is related to a the psychometric model. Previously to explicating of the model structure of the general psychometric model it is useful to first delineate the basic components of the practice of (modern) psychometrics.

3.1 On the Practice of Psychometrics

A psychometric analysis comprises the following components:

1. Construction of a measurement model that represents the important factors exerting an influence on the measure, and estimating the parameters of the models by fitting the model to data.
2. Test of the measurement model;
3. Evaluation of the characteristics of the test items used for measureing psychological constructs;
4. Prediction of latent construct scores;
5. Comparison of different groups.

By consequence, the presentation of classical test theory in Chapter 4 as well as the exposition of probabilitistic test theory in Chapter 5 comprises these five components of the practice of modern psychometrics. Let us take a short look at these 5 aspects of Psychometrics:

3.1.1 Construction of a Measurement Model and Estimation of Model Parameters

A formal definition of the concept of a measument model is given below in Concept 4-12, on page 81. For the moment it suffices to know that a measurement model (test model , psychometric model) models all relevant factors influencing the test and measurement, respectively. The model contains free parameters (cf. Notation 3-1, on p. 27). They represent predominantly the following three aspects:

1. Properties of the distribution of the latent constructs;
2. The strength of the influences of different factors on the latent variables;
3. Characteristics of test items;

The parameters are estimated from the data. The esrtimation procedure finds those parameter values that best describe the given data.

3.1.2 Evaluation of the Measurement Model

The evaluation of the measurement model comprises three aspects:

1. Assessment regarding the content of the model;
2. Evaluation of how well the model explains the data;
3. Consideration of model variants.

The evalution of the model with respect to the contents is concerned with two issues: First, is the model in accordance with existing accepted theories? For example, does the model represent correctly the relations between latent constructs? Second, are the values of estimated parameters plausible, i.e. do they conform to current knowledge?
The evaluation of how well the model fits the data is ideally performed by means of statistical testing. However, this requires precise distributional assumptions. Another way consists in performing cross-validation: The full dataset is split into different subsets and the model is fit to one of the subsets with model parameters being estimated. Then the model is used to predict the other data sets using the estimated parameters. In case of large deviation between data and model predicitons one may conclude that the model is empirically not adequate. The method of cross validation requires a big data set that can be split into smaller subsets.
It is also useful to consider variants of the basic models. Specifically, it is sensible to take into account simpler model exist that might explain the data equally well.
If the model is regarded as sufficiently acceptable then the next steps may be performed. Otherwise, a new model has to be generated and tested

3.1.3 Evaluation of Item and Test Characteristics

The most important characteristic of test items is concerned with the capability of an item to measure the latent constructs it is intended to measure. In classical test theory an important criterion for evaluating the quality concerns the reliability of a test item or of the sum of test items. In case of probabilistic test theory the information function of test items and tests provides a similar function.
In case of probabilistic test theory the range of latent scores for which an item provides the most information constitutes a second important characteristic. Usually a test items provides information about a latent construct only within a restricted range of latent construct scores. For example, an item may provide information only for high ability participants. Thus, it enables to discriminate between members of a high ability subgroup. For low ability subjects the item may not discriminate since no member of this sub-population may be able to provide a correct answer. By contrast, an item that enables to discriminate between low ability subjects may provide no information in case of high ability participants since the item is too simple and thus answered correctly by each member of this subgroup. With *adaptive testing* items are selected that provide the greatest information according to

the latent construct scores predicted on the basis of the results from items applied so far.

3.1.4 Prediction of Latent Construct Scores

The process of estimating the psychometric model results in estimation of the parameters characterizing the items and the parameters that characterize the distribution of the latent scores. However, the estimation process does not provide information about the value of a specific participant on the latent scores given her test scores. This value has to be predicted after the process of estimation using the estimated parameters of the model. (On the difference between the *estimation* of model parameters and the *prediction* of latent scores cf. Notation 4-10, on p.145).

3.1.5 Comparison of Different Groups

One important objective in the measurement of mental abilities concerns the issue of whether different groups reveal different distributions of the latent construct scores. Conclusions about differences between distributions of latent scores between different populations requires certain presuppositions.
Measurement invariance and Differential item functioning (DIF).

3.2 The General Psychometric Model

Modern psychometric models consist of three main components:
1. Latent constructs and their relationships,
2. Observed measures, test scores etc.,
3. A response model that maps values of latent constructs on values of observed measures.

Figure 3-1 exhibits the basic structure of a psychometric model. The specific models, discussed below, are but specific cases of the general model.

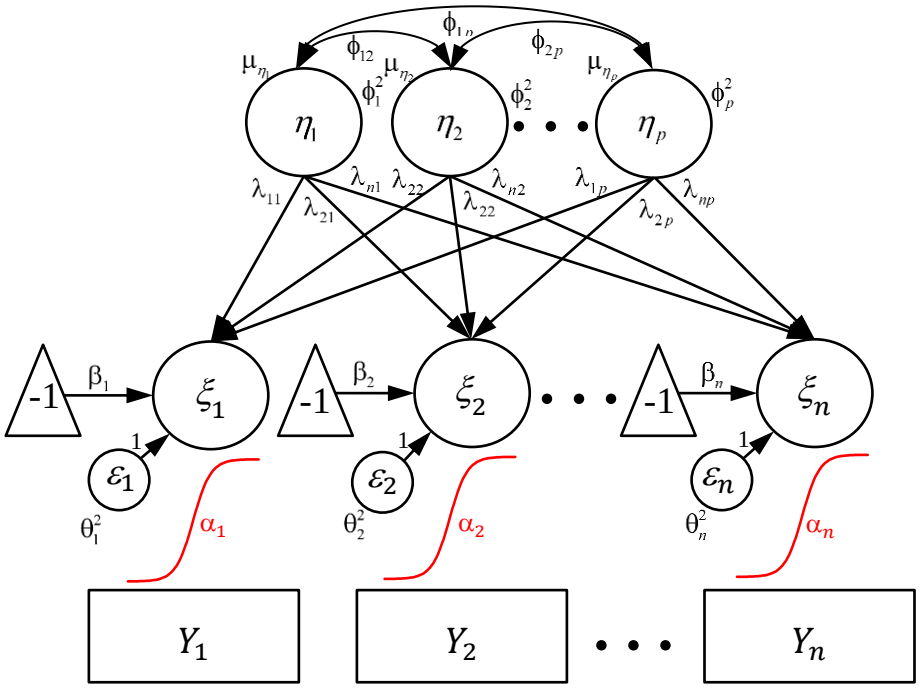


Figure 3-1: *Basic structure of a psychometric model.*

The model comprises the following components:

1. Latent variables are represented by circles and denoted by italic Greek letters. In the model of Figure 3-1 there are three types of latent variables:
   (i) Variables representing latent mental abilities or characteristics are denoted by the letters $\eta_1, \eta_2, \ldots, \eta_p$.
   (ii) Variables called *hidden response processes* are denoted by the letters $\xi_1, \xi_2, \ldots, \xi_n$.
   (iii) Variables denoting residual (or error) terms are symbolized by the letters $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$.

   Psychometric models may comprise other latent variables representing other variables exerting an influence on the test scores, like situational influences, or influences specific to the method used.
2. Observed measures are represented by rectangles and denoted by Latin letters in italic: $Y_1, Y_2, \ldots, Y_n$.
3. Constants are represented by triangles with the value of the constant within the triangle.
4. The double-arrowed arcs between the latent ability constructs represent covariances. The absence of an arrow indicates that there exists no relationship between the latent variables.
5. Arrows represent linear causal relationships between two variables.
6. The red curves represent response functions that transform values of response processes to observed responses.

7. The parameters of the model are denoted by Greek letters (not in italic). The parameters can be divided into the following classes:
   (i) *Variance parameters,* $\phi_1^2, \phi_2^2, \ldots, \phi_p^2$, as well as *covariance parameter,* $\phi_{11}, \phi_{12}, \ldots, \phi_{1p}, \phi_{23}, \phi_{24}, \ldots, \phi_{2p}, \ldots, \phi_{p-1,p}$, of the latent ability constructs.
   (ii) *Variance parameters* of the error terms*:* $\theta_1^2, \theta_2^2, \ldots, \theta_n^2$
   (iii) *Mean parameters* $\mu_1, \mu_2, \ldots, \mu_p$ of the latent ability constructs.
   (iv) *Loading coefficients* $\lambda_{11}, \lambda_{12}, \ldots, \lambda_{1p}, \lambda_{21}, \lambda_{21}, \ldots, \lambda_{2p}, \ldots, \lambda_{np}$ may be conceived of as linear regression coefficients of the regression of the response processes ξ_j on the latent mental ability constructs η_i.
   (v) Parameters $\beta_1, \beta_2, \ldots, \beta_n$ represent *item difficulties* (or thresholds).
   (vi) Parameters $\alpha_1, \alpha_2, \ldots, \alpha_n$ are *discrimination parameters* representing characteristics of the response functions, specifically the slopes of the item response functions.

There are a number of features of the model that should be considered carefully:

1. It is assumed that the latent variables $\eta_1, \eta_2, \ldots, \eta_p$ are multivariate normally distributed *random variables* with mean vector and covariance matrix:

$$\boldsymbol{\mu}_{\boldsymbol{\eta}} = \begin{bmatrix} \mu_{\eta_1} \\ \mu_{\eta_2} \\ \vdots \\ \mu_{\eta_p} \end{bmatrix}, \qquad \boldsymbol{\Phi} = \begin{bmatrix} \phi_1^2 & \phi_{12} & \cdots & \phi_{1p} \\ \phi_{21} & \phi_2^2 & \cdots & \phi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{p1} & \phi_{p2} & \cdots & \phi_p^2 \end{bmatrix}.$$

   (*Comment:* $\boldsymbol{\Phi}$ is symmetric, i.e. the rows and columns are identical; $\phi_{ij} = \phi_{ji}$ $(i, j = 1, 2, \ldots, p)$.

   The assumption of multivariate normality is a convenient assumption that enables an efficient numerical evaluation of integrals in case of non-linear response functions).

2. The error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are also assumed to be are multivariate normal distributed *random variables* with mean vector and covariance matrix:

$$\boldsymbol{\mu}_{\boldsymbol{\varepsilon}} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \boldsymbol{\Theta} = \begin{bmatrix} \theta_1^2 & \theta_{12} & \cdots & \theta_{1n} \\ \theta_{21} & \theta_2^2 & \cdots & \theta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{n1} & \theta_{n2} & \cdots & \theta_n^2 \end{bmatrix}.$$

Note that in Figure 3-1 no covariance arcs between the error terms are shown. This indicates that the covariance between each pair of error terms is zero, i.e. $\theta_{ij} = 0$, for all *i*, and *j* with $i \neq j$. In some of the models presented below the restriction of zero correlation between error terms will be released.

3. The hidden response processes may be interpreted as hidden continuous responses to an item. They are linear functions of the latent ability variables, the item difficulty and the error terms:

$$\xi_j = \lambda_{j1} \cdot \eta_1 + \lambda_{j2} \cdot \eta_2 + \cdots + \lambda_{j1} \cdot \eta_p - 1 \cdot \beta_j + 1 \cdot \varepsilon_j .$$

This equation provides the most general form of the response processes. Usually, they are much simpler. For example, in case of the Rasch model the response processes are given by:

$$\xi_j = \eta - \beta_j .$$

4. The form of the item response functions depends on the framework used:
   (i) In case of the item response version of the classical test theory the item response functions are identical functions. Consequently the hidden response ξ_j corresponds to the observed response Y_j.
   (ii) In case of probabilistic response models the response functions are functions that map the response processes ξ_j into the range [0, 1]. The commonly employed response functions are the logistic (distribution) function:

$$Y_j = \Psi(\xi_j) = \frac{\exp(\xi_j)}{1 + \exp(\xi_j)} ,$$

and the standard normal distribution function (normal ogive):

$$Y_j = \Phi(\xi_j) = \int_{-\infty}^{\xi_j} \frac{1}{\sqrt{2 \cdot \pi}} \cdot \exp\left(-\frac{x^2}{2}\right) dx$$

Both types of response functions exhibit a sigmoidal form as shown by the red curve and blue curves in Figure 3-2. Due to the fact that the logistic distribution function is easier to compute than the normal distribution function, the former has been used in most applications. As also exhibited in Figure 3-2 the logistic distribution function with discrimination parameter $\alpha = 1.7$,

$$Y_j = \frac{\exp(1.7 \cdot \xi_j)}{1 + \exp(1.7 \cdot \xi_j)} ,$$

results in a response curve (also called *item characteristic curve*) that is nearly identical to that of the normal distribution function (cf. the green dashed curve in Figure 3-2). In fact, it can be shown that the absolute difference between both curves is smaller than 0.01 over the whole range of possible values (cf. Birnbaum, 1968), i.e.:

$$\left|\Phi\left(\xi_j\right) - \Psi\left(1.7 \cdot \xi_j\right)\right| < 0.01,$$

and

$$\left|\Phi\left(\frac{\xi_j}{1.7}\right) - \Psi\left(\xi_j\right)\right| < 0.01,$$

where $\Psi\left(\xi_j\right) = \exp\left(\xi_j\right) \Big/ \left[1 + \exp\left(\xi_j\right)\right]$ denotes the logistic (distribution) function.

(iii) For the item response model of the classical test theory the response function is the identical functions, i.e., $\xi_j = Y_j$. This may be interpreted as the absence of a response function.

(iv) In a subsequent chapter (cf. Chapter **xxxx**) more complex item response functions that enable the modeling of ordered responses will be discussed.

Notation 3-1: Variables vs. parameters

Variables have to be strictly differentiated from parameters, even if both may be denoted by Greek letters:

- Variables are quantities that can take on different values.
- Parameters are constants whose values are estimated from the data. They are used to characterize different aspects of the measurement model.

The difference between variables and parameters is reflected by a different notation: Variables are set in italic font whereas parameters and constants are not.

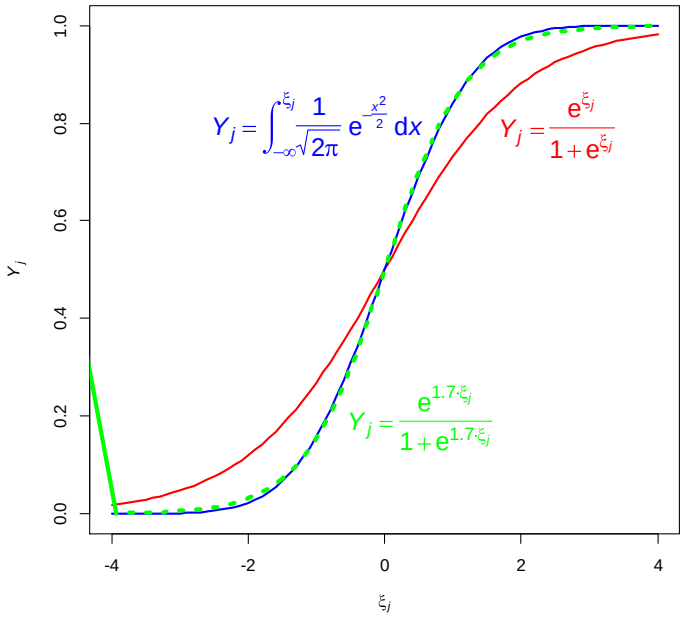


Figure 3-2: *Item response functions of probabilistic item response models: (a) red curve: logistic distribution function; (b) blue curve: standard normal distribution function; (c) green dashed curve: logistic distribution function with discrimination parameter* $\alpha = 1.7$.

Unfortunately, there exists a problem concerning the uniqueness of the interpretation of the nature of the response processes. This will be discussed next.

3.3 Probabilistic vs. Deterministic Thresholds

The item response functions may be interpreted as representing probabilistic decision functions that are based on probabilistic thresholds:

> The participant selects response category c_1 if the hidden response ξ_j of that person surpasses a certain threshold τ_j, otherwise she selects response category c_2.

In case of a probabilistic threshold, the latter is itself a random variable:

$\tau_j = \beta_j + \varepsilon_j$,

with β_j symbolizing a parameter, and ε_j follows a certain distribution. In case of the logistic response function the error term ε_j conforms to a standard logistic distribution and in case of the normal response function the error term conforms to a standard normal distribution. Assuming standard normally distributed threshold noise (i.e.

ε_j follows a standard normal distribution), the probability of response category c_1 is given by:

$$\begin{aligned} P\left(\xi_j > \tau_j\right) &= P\left(\xi_j > \beta_j + \varepsilon_j\right) \\ &= P\left(\xi_j - \beta_j > \varepsilon_j\right). \\ &= \Phi\left(\xi_j - \beta_j\right) \end{aligned}$$

Similarly, in case of a logistic distributed threshold noise, we get:

$$\begin{aligned} P\left(\xi_j > \tau_j\right) &= \Psi\xi\left(\xi_j - \beta_j\right) \\ &= \frac{\exp\left(\xi_j - \beta_j\right)}{1+\exp\left(\xi_j - \beta_j\right)}, \end{aligned}$$

where the symbol $\Psi(x) = \exp(x)/\left[1+\exp(x)\right]$ represents the logistic distribution function, with mean zero and standard deviation $\pi/\sqrt{3}$.

Concerning the alternative interpretation, one might assume a fixed, deterministic threshold:

$$\tau_j = \beta_j,$$

and locate the origin of the probabilistic response process in a noisy hidden response:

$$\zeta_j = \xi_j + \varepsilon_j.$$

This results in the same probability of selecting response category c_1:

$$\begin{aligned} P\left(\zeta_j > \tau_j\right) &= P\left(\xi_j + \varepsilon_j > \beta_j\right) \\ &= P\left(\varepsilon_j > \beta_j - \xi_j\right) \\ &= 1 - \Phi\left(\beta_j - \xi_j\right) \\ &= \Phi\left(\xi_j - \beta_j\right) \end{aligned}.$$

The last line results from the previous one since the standard normal distribution is symmetric about 0, resulting in the equality:

$$1 - \Phi(x) = \Phi(-x).$$

The same is true for the standard logistic distribution (cf. Figure 3-2). Thus a psychometric model with probabilistic response function does not allow for a unique interpretation with respect to the noise associated with the response process. The later can either be located in the hidden response process or in the threshold itself.

To summarize the discussion about the structure of psychometric models: A psychometric model specifies the latent variables and their relations as well as response functions that map the latent response processes to observed responses. The model predicts the distribution of the observed responses and it comprises free parameter. Due to these

features psychometric models are members of the class of *parametric statistical models*.
The psychometric models, to be discussed in the following, are special cases of the model of Figure 3-1 that usually do not contain all the components of the general model.
The general model could be extended further by including measured or unmeasured (latent) variables that exert a direct influence on the latent ability variables, the latent response processes, and/or the observed variables. It is also possible to model direct influences between observed measures. However, these possible extensions will not be considered in the subsequent presentation of specific models.

3.4 Comparing Psychometric and Cognitive Models

In the following we compare psychometric models with cognitive models that are also parametric statistical models. It turns out that important cognitive models are similar in structure to psychometric models despite the fact that cognitive and psychometric model builders are disjoint groups that attend different conferences, publish in different journals, and, do not often cite each other (Batchelder, 2010). The two groups represent the two disciplines alluded to by Cronbach (1957). There are exceptions, however: Darrell Bock who designed the nominal item response model claimed that he was inspired by the Bradly-Terry-Luce model, presented below (cf. Thissen, Cai & Bock, 2010).
Before we discuss the basic differences between both types of models two well-known cognitive models, the already mentioned Bradly-Terry-Luce model and the Gaussian signal detection model are presented. Both types of models comprise a set of latent constructs that are mapped by means of a response function on an observed response.

3.4.1 Bradley-Terry-Luce Model and Thurstone's Model V

The model can be used to explain choice probabilities of pair comparisons where a participant exhibits her preference by choosing one alternative out of a pair of alternatives. For example, a test might consist of different soft drinks (or types of wine, cheese, chocolate, etc.). The participant gets two samples of drinks and has to choose the one she prefers.
The basic idea underlying the model is the following: There exists a latent continuous dimension on which the different drinks (or, in general, the different objects to be evaluated) take on different positions. This latent dimension may be interpreted as a latent preference dimension with the different objects to be evaluated taking different scale values.

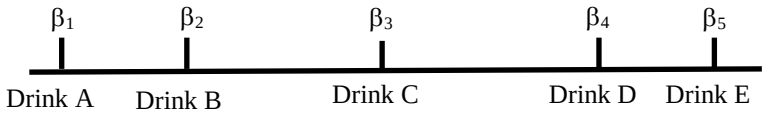


Figure 3-3: *Objects located on a latent preference dimension that constitutes the basis for the decisions. Latent scale values (or latent scores) are represented by the symbols β_1-β_5.*

Figure 3-3 depicts the situation in case of 5 soft drinks. Each drink takes a specific value on the latent scales. These values are denoted by the symbols $\beta_1, \beta_2, \ldots, \beta_5$. The greater the latent scale value of a drink the greater the preference for this drink.

The logistic distribution function (cf. Figure 3-2) maps the difference between two scale values, representing the preferences for the two drinks to be compared into the probability of choosing a drink:

$$P_{ij} = \frac{\exp\left(\beta_i - \beta_j\right)}{1 + \exp\left(\beta_i - \beta_j\right)},$$

where the symbol P_{ij} denotes the probability that drink *i* is chosen in case of drink *i* and *j* being presented. The greater the difference between two scale values, β_i and β_j $\left(\beta_i \geq \beta_j\right)$, the greater P_{ij}.

The model can be conceived of as a special case of the general psychometric model shown in Figure 3-1 (cf. Figure 3-4).

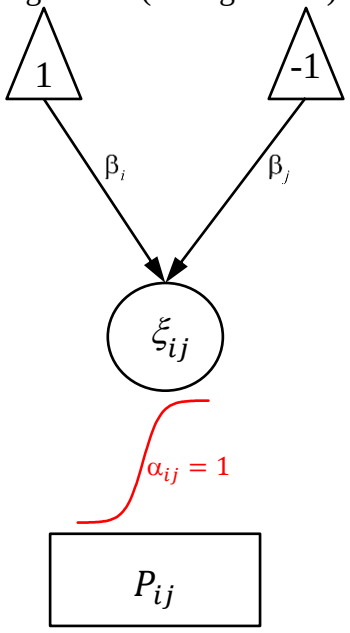


Figure 3-4: *Structure of the Bradley-Terry-Luce model: Latent scores are represented by the parameters β_i and β_j. The response function is the logistic function.*

The hidden response is the difference between the latent scores:

$$\xi_{ij} = \beta_i - \beta_j .$$

These are mapped by the logistic distribution function on the observed responses which are the probabilities P_{ij} of choosing object *i* in case of object *i* and *j* being presented.

Using as a response function the standard normal distribution function:

$$P_{ij} = \Phi\left(\xi_{ij}\right),$$

with the difference between latent scale values as hidden response and a discrimination parameter of $\alpha = 1/\sqrt{2}$, i.e.,

$$\xi_{ij} = \frac{\beta_i - \beta_j}{\sqrt{2}},$$

results in Thurstone's Case V (Thurstone, 1927). In this case the model can be given a different Interpretation: The latent scale values are the means of latent random variables η_i that are independently distributed according to a normal distribution with mean β_i and variance 1.0.

$$\eta_i \sim N\left(\beta_i, 1\right).$$

The symbol $\sim$ indicates that the random variable follows the given distribution. $N\left(\beta_i, 1\right)$ symbolizes a normal distribution with mean β_i and variance 1.0. The fact that η_i is a normally distributed random variable with mean β_i and variance 1.0 can be represented in a slightly different way by means of the equation:

$$\eta_i = \beta_i + \varepsilon_i ,$$

where,

$$\varepsilon_i \sim N\left(0, 1\right).$$

Thus, the random component ε_i conforms to a standard normal distribution. The probability of choosing alternative *i* over *j* corresponds to the probability that the difference between the value of the random variable η_i is greater than (or equal to) the value of η_j:

$$P_{ij} = P\left(\eta_i > \eta_j\right).$$

This probability is equal to the value given by the normal distribution function at $\xi_{ij} = \left(\beta_i - \beta_j\right)/\sqrt{2}$, i.e. (cf. Exercise 3-2):

$$P\left(\eta_i > \eta_j\right) = \Phi\left(\frac{\beta_i - \beta_j}{\sqrt{2}}\right).$$

According to this interpretation, the latent dimension consists of random variables that are normally distributed, with mean β_i and variance equal to 1.0. In this case the threshold is a deterministic one:

Choose option *i*, if $\eta_i > \eta_j$ and option *j* otherwise.

In case of a logistic distribution function an interpretation as in Thurstone's Case V is not possible since, unlike in care of two normally distributed random variables, the difference of two logistically distributed random variables does not conform to a logistic distribution.
Let us now consider another famous cognitive model used to measure cognitive abilities.

3.4.2 The Gaussian Signal Detection (SDT) Model

The signal detection (SDT) model is used to measure participants' ability to discrimination between different stimulus classes. Usually, only two stimulus classes are used. These are called *signal* vs. *signal + noise* in detection experiments or *new* vs. *old* in memory recognition experiments, or H_0 vs. H_A in the context of statistical hypothesis testing. The model makes the following assumptions (cf. Figure 3-5):

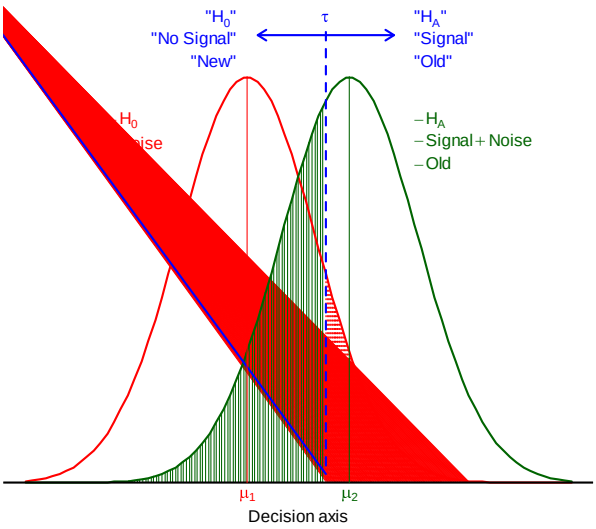


Figure 3-5: *Main components of the Gaussian SDT model: The x-axis constitutes a latent decision axis. The red and green curves represent the distributions of the subjective strength for the two stimulus classes. The blue vertical dashed line represents the decision criterion: If a subjective representation of a stimulus is located to the right of the line it is categorized as »Signal«, »Old « and »H_A«, respectively, otherwise the response category »Noise«, »New« or »H_0« is selected. The red horizontally hatched area represents the probability of a false alarm and the green vertically hatched area the probability of a false rejection.*

1. The item given to the participant is represented on a latent continuous decision axis. This represents the subjective strength of the

signal in case of a detection experiment, the strength of the memory signal in case of a memory experiment or the strength of the evidence in favor of the hypothesis H_A, in case of statistical hypothesis testing.

2. For each stimulus class (or hypothesis) *i* the subjective strengths of the individual items conform to a Gaussian (normal) distribution with a given mean μ_i and variance parameter σ_i^2 ($i = 1, \ldots, n$).
3. To make a decision, the participant sets a decision criterion (or threshold) along the latent dimension. If a subjective strength value is located to the right of the criterion the response »Signal«, »Old«, and »H_A«, respectively, is given. Otherwise the response option »Noise«, »New« or »H_0« is selected.
4. The model enables one to compute the probability of hits and false alarms (as well as those of misses and correct rejections). These are the areas under the normal density curves. Specifically, the red horizontally hatched area in Figure 3-5 represents the probability of a false alarm and the green vertically hatched area the probability of a false rejection. The probability of a hit is one minus the probability of a false rejection and the probability of a correct rejection is one minus the probability of a false alarm. In Figure 3-5 this corresponds to the area under the green density curve right from the criterion and the area left from the criterion under the red curve, respectively.

Knowing the location of the decision criterion τ and the mean and variance of the noise distribution, $\left(\mu_N, \sigma_N^2\right)$, as well as of the signal distribution, $\left(\mu_S, \sigma_S^2\right)$, the relevant probabilities can be computed (Exercise 3-5):

$$P\left(\text{Correct rejection}\right) = P\left(\gg Noise \ll \middle| Noise\right) = \Phi\left(\frac{\tau - \mu_N}{\sigma_N}\right)$$

$$P\left(\text{False alarm}\right) = P\left(\gg Signal \ll \middle| Noise\right) = 1 - \Phi\left(\frac{\tau - \mu_N}{\sigma_N}\right) = \Phi\left(\frac{\mu_N - \tau}{\sigma_N}\right).$$

$$P\left(\text{False rejection}\right) = P\left(\gg Noise \ll \middle| Signal\right) = \Phi\left(\frac{\tau - \mu_S}{\sigma_S}\right)$$

$$P\left(\text{Hit}\right) = P\left(\gg Signal \ll \middle| Signal\right) = 1 - \Phi\left(\frac{\tau - \mu_S}{\sigma_S}\right) = \Phi\left(\frac{\mu_S - \tau}{\sigma_S}\right)$$

Notation 3-2:

In diagnostic contexts the *probability of hits* and *correct rejections* are termed *sensitivity* and *specificity*.

In the context of statistical hypothesis testing *false alarms* and *false rejections* are termed *Type I* and *Type II* errors (or errors of the *first* and *second type*) and the associated probabilities are denoted by α and β. The *probability of a hit* is called the *power* of the test $(=1-\beta)$.

Obviously, the participant's ability to discriminate between different types of signals depends on the overlap of the density functions representing the distributions of the subjective signal strengths: The greater the overlap between two density curves the lower the ability to discriminate between the two types of signals. In the extreme case of total overlap the discrimination capability is zero.

It is important that note that the *observed* discrimination performance depends on the criterion setting. Assuming equal base rates or noise and signal trials and equal (positive or negative) payoffs for different outcomes the optimal location of the criterion conforms to that point on the decision axis, where the two density curves cross. Thus, the discrimination performance is influenced by both the capability to discriminate between the stimuli as well as the criterion setting. By consequence, discrimination performance, as measure for example by the probability of a correct answer, cannot be used as a measure of discrimination ability since it confounds the latter with the decision strategy. The SDT model can be used to disentangle the two aspects, thus providing a *process pure measure* of participant's discrimination capability, provided that the model represents the processes involved approximately correct.

Similar to the Bradley-Terry-Luce model the SDT model can be represented within our modeling framework of psychometric models (cf. Figure 3-1).

3.4.3 On the Difference between Psychometric and Cognitive Models

The previous discussion of the two cognitive models makes it clear that cognitive models have a similar structure as psychometric models. In fact, both models can be represented as specific cases of the general model of Figure 3-1 (on page. 24). Consequently, the difference between psychometric and cognitive models cannot be due to the structure of the models.

In order to grasp the difference between psychometric and cognitive models it is important to understand the different goals of the two modeling approaches.

Cognitive modelers are interested in modeling basic mental processes and how these are influenced by experimental manipulations. They are not (or to a lesser degree) concerned with individual differences within experimental conditions. Rather, it is assumed that participants within an experimental condition constitute a homogenous population. Con-

sequently, the values of the model parameter within an experimental condition are assumed to be (approximately) the same for different participants. Differences between participants within the same experimental condition are conceived of as noise that is not analyzed further.

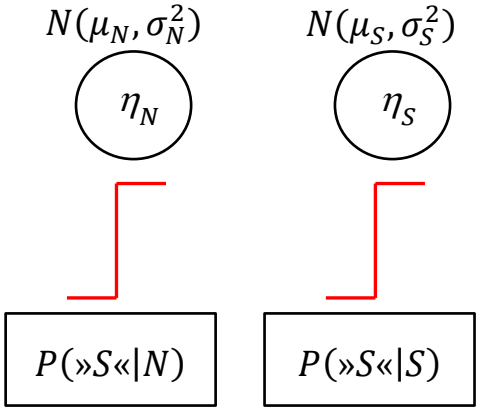


Figure 3-6: *Structure of the Gaussian SDT model for two types of signals (N = noise S = noise + signal): Latent scores are represented are normally distributed random variables. The decision process is based on deterministic thresholds indicated by the red functions.*

By contrast, psychometric modeling is concerned with the measurement of individual differences and with assessing the degree of information that a test item provides for achieving this goal.

Due to these different the parameters of two types of models refer to different entities: In cognitive models, parameters typically refer to cognitive structures or processes or they are measures of the contribution of different processes to overall behavior (cf. Batchelder, 2010). Experimental manipulations are intended to exert an influence on the different cognitive processes and their contribution to overall performance. These manipulations are thus reflected by variations of the model parameters.

By contrast, parameters of psychometric models fall into three classes:

(a) Parameters characterizing the latent ability distribution of participants,
(b) Parameters characterizing various aspects of the test items, and
(c) Parameters representing situations, methods etc.

Comment 3-1: *Parameters in psychometric models:*

In psychometric models, there are no parameters representing the values of single participants on the latent trait variables. This is due to the fact that models represent populations and parameters, thus, refer to the characteristics of the distribution of the latent traits within the population.

As a consequence, latent scores of single members of the population have to be estimated (or predicted) after estimation of the model using the estimated parameters.
There exist different methods for estimating latent scores that do not result in exactly the same estimated score.
The usage of a population of latent scores that is characterized by a distribution serves different functions in different models:

(a) In the context of classical test theory the concept of *reliability*, as well as the *axioms of classical test theory*, are based on the variances and covariances of the population of true scores. Thus the definition of the reliability concept and the specification of the axioms both require the existence of distributions of true scores (cf. Chapter 4).
(b) In probabilistic item response models reliability is replaced by the concept of *information functions*. These do not require population distributions of the latent scores. However, the estimation of item parameters requires the assumption of a population distribution (except for the one-parameter Rasch model).

It may thus be asserted that the *two disciplines*, alluded to by Cronbach (1957) [cf. Section 1.2], are reflected by the difference between cognitive and psychometric models. However, the modeling approach enables a unification of the different approaches since nothing prevents the incorporation of »psychometric characteristics« into cognitive models and/or the incorporation of parameters reflection experimental manipulations into psychometric models (For various attempts in this direction, see different contributions in the book of Embretson (2010)).
For example, a *multilevel* (or *random coefficient version*) of the Bradley-Terry-Luce model in which latent scale values are represented by latent distributions with parameters characterizing these distributions (cf. Figure 3-7) has been proposed by Bökenholt (2001). In the resulting model the population of scores is represented within the model.
On the other hand, Embretson's (1998) *cognitive design system approach* incorporated cognitive design features into psychometric models. The gap between the two disciplines may thus be overcome within the modeling approach.
We, next, turn to the discussion of different types of psychometric models, beginning with the models of classical test theory (CTT).

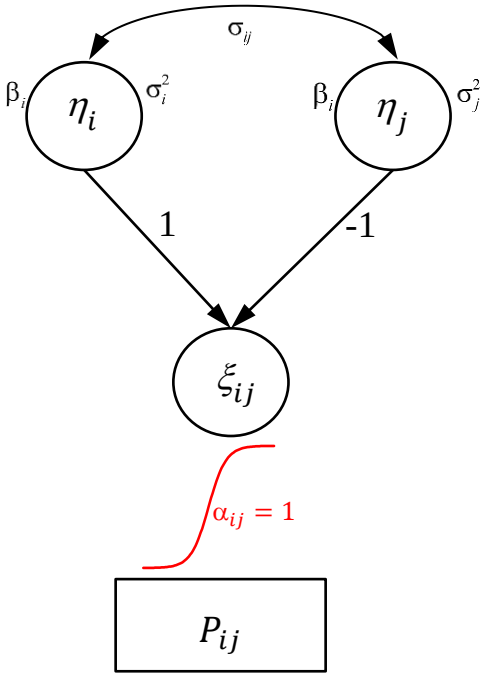


Figure 3-7: *Bradley-Terry-Luce model with a multivariate distribution of the latent scores. The variables η_i and η_j are random variables representing latent scores. The parameters β_i and β_j denote the means of the latent score distribution and the parameter σ_i^2 and σ_j^2 the variances. σ_{ij} denotes the covariance between latent scores.*

3.5 Exercises of Chapter 3

Exercise 3-1: *Generating plots of item response functions*

Generate the plots of the item response functions of Figure 3-2 (p. 28).

Exercise 3-2: *Thurstone's Case V*

Show that in Thurstone's Case V the equation:

$$P\left(\eta_i > \eta_j\right) = \Phi\left(\frac{\beta_i - \beta_j}{\sqrt{2}}\right)$$

holds.

Hints:

1. Note that $P\left(\eta_i > \eta_j\right) = P\left(\eta_i - \eta_j > 0\right)$
2. The distribution of two independently standard normal distributed variables is a normal distributed random variable with mean equal to the difference of the means and variance equal to the sum of the variances of the two variables, in our case:

   $\eta_i - \eta_j \sim N\left(\beta_i - \beta_j, 2\right)$.

Exercise 3-3: *Bradley -Terry-Luce model: Computation of expected frequencies*

Given:

Results of the 1987 Season for the American League Baseball Teams (Agresti, 2002, Table 10.10, on page 437):

Pairing		N_{win}	N_{lose}
Milwaukee	Detroit	7	6
	Toronto	9	4
	New York	7	6
	Boston	7	6
	Cleveland	9	4
	Baltimore	11	2
Detroit	Toronto	7	6
	New York	5	8
	Boston	11	2
	Cleveland	9	4
	Baltimore	9	4
Toronto	New York	7	6
	Boston	7	6
	Cleveland	8	5
	Baltimore	12	1
New York	Boston	6	7
	Cleveland	7	6
	Baltimore	10	3
Boston	Cleveland	7	6
	Baltimore	12	1
Cleveland	Baltimore	6	7

The score value of Baltimore was set to 0 in order to anchor the latent scale. The estimated latent scores of the other teams using the logit and probit models are as follows:

Model	Baltimore	Detroit	Toronto	New York	Boston	Cleveland
Logit	1.581	1.436	1.294	1.248	1.108	0.684
Probit	1.372	1.240	1.128	1.080	0.963	0.588

Use the following design matrix to compute the predicted wins and losses:

Milwaukee	Detroit	Toronto	New York	Boston	Cleveland
1	-1	0	0	0	0
1	0	-1	0	0	0
1	0	0	-1	0	0
1	0	0	0	-1	0
1	0	0	0	0	-1
1	0	0	0	0	0
0	1	-1	0	0	0
0	1	0	-1	0	0
0	1	0	0	-1	0
0	1	0	0	0	-1
0	1	0	0	0	0
0	0	1	-1	0	0
0	0	1	0	-1	0
0	0	1	0	0	-1
0	0	1	0	0	0
0	0	0	1	-1	0
0	0	0	1	0	-1
0	0	0	1	0	0
0	0	0	0	1	-1
0	0	0	0	1	0
0	0	0	0	0	1

Comments:

1. The columns represent the different teams. There is no team representing Baltimore since its scale value was fixed to 0.
2. Each line represents a pairing, e.g. the first line represents the pairing between Milwaukee and Detroit (note that scale values are subtracted).
3. Lines representing parings with Baltimore contain only a 1.

Compute the predicted wins and losses for the logit and the probit model using the scale values given above.

Hint:
In case of the probit model the entries of the design matrix have to be divided by $\sqrt{2}$ (why?).

Exercise 3-4: *Bradley -Terry-Luce model: Estimation of scale values*

Given:
The data and design matrix of Exercise 3-3.
Estimate the scale values from the data using generalized linear models.

Exercise 3-5: *SDT model: Computation of the probability of a correct response*

Given:
The parameters of the SDT model:

$\mu_N = 0,\ \sigma_N = 1$

$\mu_S = 0.8,\ \sigma_S = 1.2$

Compute:

$P_{\text{correct}} = 0.5 \cdot P(\text{Hit}) + 0.5 \cdot P(\text{False alarm}),$

using as decision bounds (thresholds):

$\tau_1 = 0.622$ (=optimal bound in case of equal base rates and payoffs)

$\tau_2 = 1.500$ (=non-optimal criterion).

4. Classical Test Theory (CTT)

Classical test theory (CTT) constitutes one of the approaches to the analysis of tests and questionnaires. Its methods are used commonly. It is thus important to understand the fundamental assumptions and methods of this approach. The following presentation proceeds as follows: We start with the presentation of some elementary facts about covariance matrices (Section 4.1). Then, the traditional characterization of the theory is provided (Section 4.2). Third, the item response version of the classical test models using structural equation models (SEM) is discussed (Section 4.3). The resulting models are but special cases of the general test model described in previous chapter (cf. Figure 3-1, on page 24). Foruth, the concept of reliability (Section 4.4) and validity (Section 4.5) are discussed. It turns out that these concepts are best understood in the context of the modeling approach of modern psychometrics. Finally, problems of modeling mean structures are discussed (Section 4.6). Specifically the problem of estimation latent ability scores and the comparison of different groups are considered.

4.1 Preliminaries: Some elementary facts about covariance matrices

Assume two sets of random variables denoted by $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$. The covariance matrix $\boldsymbol{\Sigma}$ of the whole set of variables is the *mn* × *mn* matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} & X_1 & X_2 & \cdots & X_n & Y_1 & Y_2 & \cdots & Y_m \\ X_1 & \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_n} & \sigma_{X_1Y_1} & \sigma_{X_1Y_2} & \cdots & \sigma_{X_1Y_m} \\ X_2 & \sigma_{X_2X_1} & \sigma^2_{X_2} & \cdots & \sigma_{X_2X_n} & \sigma_{X_2Y_1} & \sigma_{X_2Y_2} & \cdots & \sigma_{X_2Y_m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ X_n & \sigma_{X_nX_1} & \sigma_{X_nX_2} & \cdots & \sigma^2_{X} & \sigma_{X_nY_1} & \sigma_{X_nY_2} & \cdots & \sigma_{X_nY_m} \\ Y_1 & \sigma_{Y_1X_1} & \sigma_{Y_1X_2} & \cdots & \sigma_{Y_1X_n} & \sigma^2_{Y_1} & \sigma_{Y_1Y_2} & \cdots & \sigma_{Y_1Y_m} \\ Y_2 & \sigma_{Y_2X_1} & \sigma_{Y_2X_2} & \cdots & \sigma_{Y_2X_n} & \sigma_{Y_2Y_1} & \sigma^2_{Y_2} & \cdots & \sigma_{Y_2Y_m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_m & \sigma_{Y_mX_1} & \sigma_{Y_mX_2} & \cdots & \sigma_{Y_mX_n} & \sigma_{Y_mY_1} & \sigma_{Y_mY_2} & \cdots & \sigma^2_{Y_m} \end{bmatrix}$$

The matrix contains three different types of entries:

1. Variance parameters in the main diagonal: $\sigma^2_{X_i}$ $(i = 1, 2, \ldots, n)$ and $\sigma^2_{Y_j}$ $(j = 1, 2, \ldots, m)$.
2. Covariance parameters representing the covariance between the same group of variables: $\sigma_{X_iX_{i'}}$ $(i, i' = 1, 2, \ldots, n)$ and $\sigma_{Y_jY_{j'}}$ $(j, j' = 1, 2, \ldots, m)$.

3. Covariance parameters representing the covariance between variables from different groups: $\sigma_{X_iY_j}$ $(i=1,2,\ldots,n)$ $(j=1,2,\ldots,m)$.

The covaince matrix can thus be partitioned into four regions:

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|ccc|ccc} & X_1 & X_2 & \cdots & X_n & Y_1 & Y_2 & \cdots & Y_m \\ X_1 & \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_n} & \sigma_{X_1Y_1} & \sigma_{X_1Y_2} & \cdots & \sigma_{X_1Y_m} \\ X_2 & \sigma_{X_2X_1} & \sigma^2_{X_2} & \cdots & \sigma_{X_2X_n} & \sigma_{X_2Y_1} & \sigma_{X_2Y_2} & \cdots & \sigma_{X_2Y_m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ X_n & \sigma_{X_nX_1} & \sigma_{X_nX_2} & \cdots & \sigma^2_X & \sigma_{X_nY_1} & \sigma_{X_nY_2} & \cdots & \sigma_{X_nY_m} \\ \hline Y_1 & \sigma_{Y_1X_1} & \sigma_{Y_1X_2} & \cdots & \sigma_{Y_1X_n} & \sigma^2_{Y_1} & \sigma_{Y_1Y_2} & \cdots & \sigma_{Y_1Y_m} \\ Y_2 & \sigma_{Y_2X_1} & \sigma_{Y_2X_2} & \cdots & \sigma_{Y_2X_n} & \sigma_{Y_2Y_1} & \sigma^2_{Y_2} & \cdots & \sigma_{Y_2Y_m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_m & \sigma_{Y_mX_1} & \sigma_{Y_mX_2} & \cdots & \sigma_{Y_mX_n} & \sigma_{Y_mY_1} & \sigma_{Y_mY_2} & \cdots & \sigma^2_{Y_m} \end{array}\right] = \begin{bmatrix} \boldsymbol{\Sigma_{XX}} & \boldsymbol{\Sigma_{XY}} \\ \boldsymbol{\Sigma_{YX}} & \boldsymbol{\Sigma_{YY}} \end{bmatrix}.$$

- $\boldsymbol{\Sigma_{XX}}$ represents the $n \times n$ covariance matrix of the variables $X_1, X_2, \ldots, X_n$.
- $\boldsymbol{\Sigma_{YY}}$ represents the $m \times m$ covariance matrix of the variables $Y_1, Y_2, \ldots, Y_m$.
- $\boldsymbol{\Sigma_{XY}}$ represents the $n \times m$ matrix of the covariances between variables from different groups: $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$.
- $\boldsymbol{\Sigma_{YX}}$ represents the $m \times n$ matrix of the covariances between variables from different groups: $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$. It is the transpose of $\boldsymbol{\Sigma_{XY}}$: $\boldsymbol{\Sigma_{YX}} = \boldsymbol{\Sigma_{XY}^T}$. This means that rows and columns are interchanged.

Here are two important results concerning the relationship between the entries of a covariance matrix and the variance and covariance of sums of variables (cf. Exercise 4-1):

1. Let X be the sum $X = X_1 + X_2 + \cdots + X_n$. The variance of the sum variable X, $\mathrm{Var}(X)$, is the sum of all the entries of the covariance matrix $\boldsymbol{\Sigma_{XX}}$ of the variables $X_1, X_2, \ldots, X_n$.
2. Let Y be the sum $Y = Y_1 + Y_2 + \cdots + Y_m$. The covariance between the two sum variables X and Y, $\mathrm{Cov}(X,Y)$, is given by the sum of all the entries of the matrix $\boldsymbol{\Sigma_{XY}}$ (or $\boldsymbol{\Sigma_{YX}}$) of the matrix containing the covariance between the different groups of variables $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$.

Concnering the numbers of different relevant entries the following facts are important. Let $\mathbf{\Sigma}_{xx}$ be the covariance matrix of the variables $X_1, X_2, \ldots, X_n$:

1. $\mathbf{\Sigma}_{xx}$ contains $n \cdot n = n^2$ entries.
2. $\mathbf{\Sigma}_{xx}$ contains n variance terms in the main diagonal.
3. $\mathbf{\Sigma}_{xx}$ contains $n^2 - n = n \cdot (n-1)$ covariance terms: the n^2 entries of the matrix minus the n variances in the main diagonal.
4. $\mathbf{\Sigma}_{xx}$ contains $n \cdot (n-1)/2$ *unique* covariance terms (since $\mathbf{\Sigma}_{xx}$ is symmetric, i.e. the entries in the upper right part are identical to that in the lower left part).
5. $\mathbf{\Sigma}_{xx}$ contains $n \cdot (n+1)/2$ variance plus unique covariance terms: $n \cdot (n-1)/2 + n = n \cdot (n+1)/2$.

4.2 The Basic Concepts of CTT

In the following, a description of CTT is given that is close to explication of the theory in classical texts like Lord and Novick (1968). The presentation of the theory starts with a general characterization (Section 4.2.1). This is followed by an explication of the axioms of CTT (Section 4.2.2) and the classical test models (Section 4.2.3). Finally, we discuss problematic aspects underlying the conception of CTT (Section 4.2.4).

4.2.1 Exposition of CTT

CTT is concerned with the following three types of parameters characterizing tests and their relationships:

1. *Expected values:* Expected test values are used for representing the magnitude of abilities and of test results within the population of examinees. Specific issues concern the difference in expected values between different groups or the bias of tests, that is, whether a test over- or underestimates the latent abilities of different groups.
2. *Variances:* Variances concern the dispersion of test scores, abilities and errors. CTT enables the attribution of the variance of test scores to different sources.
3. *Covariances / Correlations:* These represent the relationships between tests, latent abilities and errors. CTT explains the existence of covariance (or correlations) between observed test scores of different tests.

The expectations constitute the *mean structure* and the variances and covariances the *covariance structure*. CTT introduces a number of assumptions that permit the explanation of the observed mean and covariance structure of the tests. These assumptions can be classified into two groups:

1. *The axioms of CTT:* These constitute the abstract framework of the theory.
2. *The classical test models:* These are based on further assumptions (additionally to the axioms) with respect to the relationships between different tests.

The axioms of CTT provide restrictions on the set of possible models. However, these restrictions are not sufficient to generate predictions that could be testes empirically. Due to this reasons additional assumptions are required that are incorporated into the classical test models. Given enough test items, the classical test models generate empirically testable predictions. They also enable the estimation of the reliability of the test items as well as the size of the measurement error and the latent ability scores.
CTT is concerned with means, variances and covariances (or correlations) only. It makes no assumption about the exact distribution of the test scores. By consequence, the empirical adequacy of the test models cannot be tested statistically.

Concept 4-1: *Weak and strong true score theory:*

Due to the fact that only means, variances and covariances are considered, CTT has also been called *weak true score theory* in contrast to *strong true score theory* that is characterized by the specification of distributions.

Following to this general characterization the axioms of CTT will be described next.

4.2.2 The Axioms of CTT

The core of CTT is constituted by the decomposition of the observed test score into a true score and error score (measurement error):

$$Y_{pi} = \tau_{pi} + \varepsilon_{pi}. \qquad (4\text{-}1)$$

The symbols have the following meaning:

Y_{pi} represents the observed test score of person *p* on test *i*,

τ_{pi} symbolizes the true score of person *p* on test *i*,

ε_{pi} denotes the measurement error of person *p* on test *i*.

Note that in Equation (4-1) the true score is a constant whereas the observed test score and the error are random variables.

Equation (4-1) tells us that the observed test score of person *p* on test *i* is the sum of the true score plus the measurement error. The following additional assumption concerning the expectation of the measurement errors is made:

$$\mathrm{E}\left(\varepsilon_{pi}\right) = 0, \qquad (4\text{-}2)$$

or, equivalently (due to the rules of expectations):

$$\mathrm{E}(Y_{pi}) = \tau_{pi}. \qquad (4\text{-}3)$$

This assumption is based on the idea that *the true score corresponds to the expectation of the test scores in repeated measurements* of person *p* on test *i* (further details on this point are provided below in Section 4.2.4.2).

The decomposition of a test score into true score and error leads to the problem of a *lack of identification* of true score and error: Given a single measured value it is impossible to determine the value of the true score and the error. One possible solution to this problem might consist in the repeated application of the same test item to the same examinee. This would result in a distribution of test scores: The expected value would then serve as an estimator of the true score and the variance of the scores can be used as estimator of the error variance.

Unfortunately this solution is practically infeasible. CTT has thus taken another approach: Different tests (test items) are applied and the covariance structure underlying the observed test scores are assumed to be restricted in specific ways. The restrictions are specified in two different ways:

1. The *axioms of CTT* specify restrictions on the covariances by assuming that certain covariances are zero.
2. The *classical test models* of CTT are specified. These introduce further restrictions on the covariance structure.

Concept 4-2: *Axiom:*

An *axiom* is an statement that is assumed to be true.

Comment: An axiom does *not* represent a necessary or eternal truth.

Here are the axioms of CTT:

$$\mathrm{Cov}(\tau_i, \varepsilon_i) = 0 \qquad (4\text{-}4)$$

$$\mathrm{Cov}(\tau_i, \varepsilon_j) = 0 \qquad (4\text{-}5)$$

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \qquad (4\text{-}6)$$

Previously to discussing the axioms it should be noted that in equation (4-4) to (4-6) the index *p* is missing. By consequence, contrary to τ_{pi}, τ_i represents a random variable (and not a constant). Therefore it is written in italic (cf. Notation 3-1 on page 27). Moreover, ε_i now refers to the variance of the errors within the population of examinees (*between-subjects distribution*), and not, to *within-subjects distribution* like ε_{pi}. Consequently, the axioms refer to populations of examinees (cf. Comment 3-1 on page 36).

The first axiom (4-4) states that the error and the true score of a test *i* are uncorrelated within the population (for all tests $i = 1, 2, \ldots, n$).

The second axiom (4-5) asserts that the measurement error ε_j associated with test *j* is uncorrelated with the true score of on different test *j* $(i \neq j)$.

Together, Axiom 1 and 2 state that the true score of a test is neither correlated with the own error nor with the error of another test. This is true for all tests.

Finally, the third axiom maintains that the errors associated with different tests are uncorrelated. This axiom may be relaxed, and, in fact, models assuming correlated errors will be considered.

Ex. 4-1 illustrates how the axioms of CTT result in the simplification of the covariance structure of true scores and measurement errors.

Ex. 4-1: Simplification of the covariance structure of true scores and errors due to the axioms of CTT

Given:

- 3 random variables Y_1, Y_2, and Y_3 represent the test scores of three tests.
- 3 random variables τ_1, τ_2, and τ_3 represent the true score of the 3 tests.
- 3 random variables ε_1, ε_2, and ε_3, represent the measurement errors of the 3 tests.

The covariance matrix $\boldsymbol{\Sigma}$ constitutes the covariance structure of true scores and errors

$$\boldsymbol{\Sigma} = \begin{array}{c} \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{array} \begin{array}{c} \begin{array}{cccccc} \tau_1 & \tau_2 & \tau_3 & \varepsilon_1 & \varepsilon_2 & \varepsilon_3 \end{array} \\ \begin{bmatrix} \sigma^2_{\tau_1} & \sigma_{\tau_1\tau_2} & \sigma_{\tau_1\tau_3} & \sigma_{\tau_1\varepsilon_1} & \sigma_{\tau_1\varepsilon_2} & \sigma_{\tau_1\varepsilon_3} \\ \sigma_{\tau_2\tau_1} & \sigma^2_{\tau_2} & \sigma_{\tau_2\tau_3} & \sigma_{\tau_2\varepsilon_1} & \sigma_{\tau_2\varepsilon_2} & \sigma_{\tau_2\varepsilon_3} \\ \sigma_{\tau_3\tau_1} & \sigma_{\tau_3\tau_2} & \sigma^2_{\tau_3} & \sigma_{\tau_3\varepsilon_1} & \sigma_{\tau_3\varepsilon_2} & \sigma_{\tau_3\varepsilon_3} \\ \sigma_{\varepsilon_1\tau_1} & \sigma_{\varepsilon_1\tau_2} & \sigma_{\varepsilon_1\tau_3} & \sigma^2_{\varepsilon_1} & \sigma_{\varepsilon_1\varepsilon_2} & \sigma_{\varepsilon_1\varepsilon_3} \\ \sigma_{\varepsilon_2\tau_1} & \sigma_{\varepsilon_2\tau_2} & \sigma_{\varepsilon_2\tau_3} & \sigma_{\varepsilon_2\varepsilon_1} & \sigma^2_{\varepsilon_2} & \sigma_{\varepsilon_2\varepsilon_3} \\ \sigma_{\varepsilon_3\tau_1} & \sigma_{\varepsilon_3\tau_2} & \sigma_{\varepsilon_3\tau_3} & \sigma_{\varepsilon_3\varepsilon_1} & \sigma_{\varepsilon_3\varepsilon_2} & \sigma^2_{\varepsilon_3} \end{bmatrix} \end{array} = \begin{bmatrix} \boldsymbol{\Sigma}_{\tau\tau} & \boldsymbol{\Sigma}_{\tau\varepsilon} \\ \boldsymbol{\Sigma}_{\varepsilon\tau} & \boldsymbol{\Sigma}_{\varepsilon\varepsilon} \end{bmatrix}$$

The symbols have the following meaning:

$\sigma^2_{\tau_i}$ denotes the variance of the true scores of test *i* in der population ($i = 1,2,3$).

$\sigma^2_{\varepsilon_i}$ denotes the variance of the errors of test *i* in der population ($i = 1,2,3$).

$\sigma_{\tau_i\tau_j}$ denotes the covariance between the true scores of test *i* and test *j* ($i, j = 1,2,3$) [$\sigma_{\tau_i\tau_j} = \sigma_{\tau_j\tau_i}$].

$\sigma_{\tau_i\varepsilon_j}$ denotes the covariance between the true scores of test *i* and the errors of test *j* ($i, j = 1,2,3$) [$\sigma_{\tau_i\varepsilon_j} = \sigma_{\varepsilon_j\tau_i}$].

$\sigma_{\varepsilon_i\varepsilon_j}$ denotes the covariance between measurement errors of test *i* and *j* $(i, j = 1,2,3)$ [$\sigma_{\varepsilon_i\varepsilon_j} = \sigma_{\varepsilon_j\varepsilon_i}$].

The axioms of CTT result in a simplification by assuming that the covariance between true scores and errors as well as the covariance between errors is zero: $\sigma_{\tau_i\varepsilon_j} = 0$ and $\sigma_{\varepsilon_i\varepsilon_j} = 0$, respectively $(i, j = 1,2,3)$. This results in the following simplified covariance matrix:

$$\boldsymbol{\Sigma} = \begin{array}{c} \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{array} \begin{array}{c} \begin{array}{cccccc} \tau_1 & \tau_2 & \tau_3 & \varepsilon_1 & \varepsilon_2 & \varepsilon_3 \end{array} \\ \begin{bmatrix} \sigma^2_{\tau_1} & \sigma_{\tau_1\tau_2} & \sigma_{\tau_1\tau_3} & 0 & 0 & 0 \\ \sigma_{\tau_2\tau_1} & \sigma^2_{\tau_2} & \sigma_{\tau_2\tau_3} & 0 & 0 & 0 \\ \sigma_{\tau_3\tau_1} & \sigma_{\tau_3\tau_2} & \sigma^2_{\tau_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{\varepsilon_1} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{\varepsilon_2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{\varepsilon_3} \end{bmatrix} \end{array}$$

The independence assumptions specified by the axioms lead to a simplification of the representation of variances of observed test scores Y_i $(i = 1,2,3)$ as a function of the variances and covariances of true scores and errors.

Due to the decomposition of the observed value into true score and error (cf. Equation 4-1) the variance of the observed score is given by:

$$\mathrm{Var}(Y_i) = \mathrm{Var}(\tau_i + \varepsilon_i) = \mathrm{Var}(\tau_i) + \mathrm{Var}(\varepsilon_i) + 2 \cdot \mathrm{Cov}(\tau_i, \varepsilon_i)$$

Since $\mathrm{Cov}(\tau_i, \varepsilon_i) = 0$ the expression simplifies to:

$$\mathrm{Var}(Y_i) = \mathrm{Var}(\tau_i) + \mathrm{Var}(\varepsilon_i)$$

Consequently, the covariance matrix $\boldsymbol{\Sigma}_y$ of the observed values has the following structure:

$$\boldsymbol{\Sigma}_{\mathbf{y}} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \end{array} \begin{array}{c} \begin{array}{ccc} Y_1 & Y_2 & Y_3 \end{array} \\ \begin{bmatrix} \sigma^2_{\tau_1} + \sigma^2_{\varepsilon_1} & \sigma_{\tau_1\tau_2} & \sigma_{\tau_1\tau_3} \\ \sigma_{\tau_2\tau_1} & \sigma^2_{\tau_2} + \sigma^2_{\varepsilon_2} & \sigma_{\tau_2\tau_3} \\ \sigma_{\tau_3\tau_1} & \sigma_{\tau_3\tau_2} & \sigma^2_{\tau_3} + \sigma^2_{\varepsilon_3} \end{bmatrix} \end{array}$$

The associated *covariance equations* (cf. Concept 4-11, on page 69), representing the variances and covariances of the observed test scores in terms of the variances and covariances of true scores and errors, are thus given by:

$$\begin{aligned} \sigma^2_{Y_1} &= \sigma^2_{\tau_1} + \sigma^2_{\varepsilon_1} \\ \sigma^2_{Y_2} &= \sigma^2_{\tau_2} + \sigma^2_{\varepsilon_2} \\ \sigma^2_{Y_3} &= \sigma^2_{\tau_3} + \sigma^2_{\varepsilon_3} \\ \sigma_{Y_1Y_2} &= \sigma_{\tau_1\tau_2} \\ \sigma_{Y_1Y_3} &= \sigma_{\tau_1\tau_3} \\ \sigma_{Y_2Y_3} &= \sigma_{\tau_2\tau_3} \end{aligned}$$

The observed variances and covariances of the observed test scores are located on left-hand side of the covariance equations whereas the unknown variances and covariances of the true scores and error terms are and right-hand side.

Our objective consists in estimating the unknown parameters representing the variances and covariances of the true scores as well as the variances of the error terms by solving the covariance equations for the unknown quantities. If the unknown parameters can be *uniquely computed* by using the whole set or a subset of the covariance equations then the parameters are *identified*.

Obviously, the variances of the true scores and errors are not identified since there are only three equations involving variance terms whereas there are six unknown variances.

By contrast, the covariances between true scores are (exactly) identified: The observed covariances between test scores may be used as estimates of respective covariances of the involved true scores.

The example illustrates that the axioms of CTT do not specify enough restrictions to identify the variance parameters of the true and error scores.

To enable the identification of all parameters, and, moreover, to generate empirically testable predictions additional restrictions are required leading to a further simplification of the covariance structure of the true scores.

Previously to discussing these additional restrictions let us summarize our considerations arrived at so far:

1. The decomposition of a person's test score into the true score and error score raises the problem of how to determine the person's true and error score.
2. The problem could be solved »in principle« by means of repeated testing of person *p*: Repeated testing results in a *within-subject distribution* of test values whose mean can be used as an estimate of the true score and whose variance as an estimate of the error variance.
   This procedure assumes that such a within-subject distribution exists. Various scientists have cast doubt on the concept of a within-subject distribution of this sort (Borsboom, 2005; Holland, 1990).
3. CTT solves the problem by the specification of axioms and specific test models (to be discussed in the next section) that impose constraints on the covariance structure of true and error scores, thus resulting in models that enable the estimation of the covariance structure of the true and error scores.
4. The axioms do not refer to a *within-subject distribution* but to the distribution of true and error scores within the population of examinees, that is a *between-subjects* distribution.
   Note that the variance of the true scores has no meaning with respect to the *within-subject* perspective since in this case the true score is a constant value. The reliability of a test that is an important quantity in CTT measuring the precision of a test (cf. Section 4.4) is also founded on the between subjects distribution and thus population based (cf. Comment 3-1, page 36).
5. The axioms do not lead to a simplification of the covariance structure that enables the identification of the parameters of the model. Therefore, additional constraints have to be specified. These are implemented in the classical test models to which we turn next.

4.2.3 The Classical Test Models

In the following we discuss three test model of CTT:

1. The congeneric model (also called τ-congeneric model),
2. The τ-equivalent model, and
3. The parallel model.

The congeneric model is the most general one. The imposition of further constraints results in the τ-equivalent model. Specifying additional restrictions lead to the parallel model which is the most specific one. Due to the fact that the more specific model results by imposing restrictions on the superordinate one, the three test models are nested.
For each of the three models the unknown variance parameters can be estimated with enough tests being present.
Let us consider the three models in detail.

4.2.3.1 THE CONGENERIC TEST MODEL

As illustrated above, the general model of CTT given by the decomposition of test scores and application of the axioms of CTT does not impose any restrictions on the covariance structure of the true scores. A simplification of this covariance structure is thus an obvious option for constructing model that can be estimated. For the present models this is achieved by assuming that the true scores of the different tests are linear dependent, i.e. one true score is simply a linear function of the other one.

Concept 4-3: *Congeneric model of mental tests:*

Given:

$Y_1, Y_2, \ldots, Y_n$ observed test scores for *n* tests.

$\tau_1, \tau_2, \ldots, \tau_n$ true scores associated with the *n* tests.

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ measurement errors associated with the *n* tests.

We chose one of the tests, e.g. the first test, as a reference with respect to which the true scores of the other tests are linear functions. The true score of the selected test is denoted by the symbol τ (no subscript).

The *n* tests are *congeneric* if the associated true scores conform to the linear relation:

$$\tau_i = \lambda_i \cdot \tau + \alpha_i \quad (i = 2, \ldots, n) \tag{4-7}$$

The coefficients λ_i are asumed to be positive: $\lambda_i > 0$.

The congeneric model entails the following covariance structure of the true scores:

$$\sigma^2_{\tau_i} = \begin{cases} \sigma^2_\tau & \Leftrightarrow \quad i = 1 \\ \lambda_i^2 \cdot \sigma^2_\tau & \Leftrightarrow \quad i = 2, \ldots, n \end{cases} \tag{4-8}$$

$$\sigma_{\tau_i \tau_j} = \begin{cases} \lambda_j \cdot \sigma^2_\tau & \Leftrightarrow \quad i = 1; j = 2, \ldots, n \\ \lambda_i \cdot \sigma^2_\tau & \Leftrightarrow \quad j = 1; i = 2, \ldots, n \\ \lambda_i \cdot \lambda_j \cdot \sigma^2_\tau & \Leftrightarrow \quad i, j = 2, \ldots, n; i \neq j \end{cases} \tag{4-9}$$

Since the *n* tests are linearly dependent the correlation between all trues scores is 1.0 (cf. Ex. 4-2). Since, by assumption, the coefficients λ_i are positive no negative correlations can result.

Comment:

The intercept parameters $\alpha_i \; (i = 2, \ldots, n)$ are irrelevant as far as the covariance structure is concerned. However they are important for modeling the mean structure (cf. Section 4.6)

Please note that the entailed covariance structure results from the linear constraint by applying simple covariance algebra (cf. Exercise 4-3).

Ex. 4-2: Covariance structure of true scores and of observed test scores of 4 congeneric tests:

Given:

Y_1, Y_2, Y_3, Y_4 the observed test scores of 4 tests.

$\tau, \tau_2, \tau_3, \tau_4$ the associated true scores.

$\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ the respective measurement errors.

The covariance matrix of the true scores looks like this:

$$\begin{bmatrix} \sigma_\tau^2 & \lambda_2 \cdot \sigma_\tau^2 & \lambda_3 \cdot \sigma_\tau^2 & \lambda_4 \cdot \sigma_\tau^2 \\ & \lambda_2^2 \cdot \sigma_\tau^2 & \lambda_2 \cdot \lambda_3 \cdot \sigma_\tau^2 & \lambda_2 \cdot \lambda_4 \cdot \sigma_\tau^2 \\ & & \lambda_3^2 \cdot \sigma_\tau^2 & \lambda_3 \cdot \lambda_4 \cdot \sigma_\tau^2 \\ & & & \lambda_4^2 \cdot \sigma_\tau^2 \end{bmatrix}$$

Since the covariance matrix is symmetric only the entries on and above the main diagonal are shown.

The correlation between the true scores τ and τ_i $(i = 2,3,4)$ is given by:

$$\text{Corr}(\tau, \tau_i) = \frac{\text{Cov}(\tau, \tau_i)}{\sqrt{\text{Var}(\tau) \cdot \text{Var}(\tau_i)}} = \frac{\lambda_i \cdot \sigma_\tau^2}{\sqrt{(\sigma_\tau^2) \cdot (\lambda_i^2 \cdot \sigma_\tau^2)}} = 1.0$$

Similarly, the correlation between the true scores τ_i und τ_j $(i, j = 2,3,4; i \neq j)$ is given by:

$$\text{Corr}(\tau_i, \tau_j) = \frac{\text{Cov}(\tau_i, \tau_j)}{\sqrt{\text{Var}(\tau_i) \cdot \text{Var}(\tau_j)}} = \frac{\lambda_i \cdot \lambda_j \cdot \sigma_\tau^2}{\sqrt{(\lambda_i^2 \cdot \sigma_\tau^2) \cdot (\lambda_j^2 \cdot \sigma_\tau^2)}} = 1.0$$

The assumption of a linear relationship between the true scores of the tests leads to a considerable simplification of the covariance structure of the true scores:

The the covariance matrix of the true scores for the general CTT model comprises $n \cdot (n+1)/2 = 4 \cdot 5/2 = 10$ unknown variances and covariances.

In the congeneric model these 10 unknown parameters are represented by 4 unknown parameters: σ_τ^2, λ_2, λ_3,and λ_4.

The model implied covariance matrix (upper part) of the observed test scores has the following structure:

$$\begin{bmatrix} \sigma_\tau^2+\sigma_{\varepsilon_1}^2 & \lambda_2\cdot\sigma_\tau^2 & \lambda_3\cdot\sigma_\tau^2 & \lambda_4\cdot\sigma_\tau^2 \\ & \lambda_2^2\cdot\sigma_\tau^2+\sigma_{\varepsilon_2}^2 & \lambda_2\cdot\lambda_3\cdot\sigma_\tau^2 & \lambda_2\cdot\lambda_4\cdot\sigma_\tau^2 \\ & & \lambda_3^2\cdot\sigma_\tau^2+\sigma_{\varepsilon_3}^2 & \lambda_3\cdot\lambda_4\cdot\sigma_\tau^2 \\ & & & \lambda_4^2\cdot\sigma_\tau^2+\sigma_{\varepsilon_4}^2 \end{bmatrix}$$

The 10 observed variances and covariances of the test scores are thus explained by means of 8 free parameters:

σ_τ^2, λ_2, λ_3, λ_4, $\sigma_{\varepsilon_1}^2$, $\sigma_{\varepsilon_2}^2$, $\sigma_{\varepsilon_3}^2$, and $\sigma_{\varepsilon_4}^2$.

The first four parameters represent the covariance structure of the true scores, whereas the last four parameters represent the covariance structure of the errors.

The 8 unknown parameters can be determined from the model implied variances and covariances of the observed test scores by solving the model (or covariance) equations for the unknown parameters (cf. Exercise 4-4).

The 10 model equations linking model implied (co-) variances to the observed ones follow directly from the model implied covariance matrix of the observed scores shown above (for an example of covariance equations, cf. Ex. 4-1, on page 47).

Concept 4-4: *Parameters (of a statistical model):*

Parameters are fixed but usually unknown quantities that characterize the population distribution in question.

A statistical model is used for modeling the population distribution. Thus, the *parameters of a statistical model* characterize the population distribution in case of the model being (approximatively) correct.

In CTT models, the parameters characterize the mean and covariance structure of the test scores that are predicted by a model. In this case the model does not represent the complete distribution but only the first and second (central) moments of the distribution (i.e. means and (co-) variances). Consequently, in CTT the parameters characterize the first and second central moments of the population only.

Concept 4-5: *(Sample-) Statistics:*

A *sample statistic* (or simply a *statistic*) is a function of the observed sample. Thus its value is determined by the observed sample values.

In CTT models the relevant statistics are the means and the (co-) variances of the observed test scores.

Notation 4-1: Parameters and statistics

Parameters and sample statistics have to be kept strictly apart. This is reflected by the notation used: Parameters are denoted by Greek letters (e.g. μ, σ_τ^2, λ_i), whereas statistics are symbolized by Latin letters (e.g. $\bar{x}$ s_Y^2).

Let us briefly summarize the main aspects of the congeneric model:
The assumed linear relationship between the true scores, given by Equation 4-7, results in a simplification of the covariance structure of the true scores. Specifically, the correlation between true scores becomes 1.0 (For an illustration, cf. Ex. 4-2). As a consequence the addition of further congeneric tests (i.e. tests whose true scores conform to Equation 4-7) results in a stronger increase of observed variances and covariances than of new parameters. For example, with 3 tests the number of free parameters equals the number of observed variances and covariances (=6). With 4 tests the number of variances and covariances is 10. However, the number of free parameters is only 8.
The addition of a new congeneric test to n existing tests requires 2 additional free parameters: λ_i and $\sigma_{\varepsilon_i}^2$. However, the number of observed variances and covariances increases by $n+1$ (n new covariances and 1 new variance).

Comment 4-1: *The basic principle underlying the congeneric model*

The presentation of the model using Equation 4-7 does not elucidate the main principle underlying the congeneric model. Using structural equations for representing the model (Section 4.3) reveals that the restriction given by Equation 4-7 amounts to the assumption that the tests measure the same true score however with different sensitivity.

4.2.3.2 DIGRESSION: COMPUTATION OF THE MODEL IMPLIED COVARIANCE MATRIX USING MATRIX METHODS

The computation of the covariance matrix of true scores as well as of the model implied covariance matrix of test scores (cf. Ex. 4-2, page 52) can be achieved with little effort using matrices (as well as a program, like Excel, Matlab or R that enables the manipulation of matrices.
For the computation of the relevant matrices of the congeneric model the specification of the following three matrices is required:

$$\boldsymbol{\lambda} = \begin{bmatrix} 1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \lambda_n \end{bmatrix}, \qquad \boldsymbol{\Phi} = \left[\sigma_\tau^2\right], \qquad \boldsymbol{\Psi} = \begin{bmatrix} \sigma_{\varepsilon_1}^2 & 0 & \dots & 0 \\ 0 & \sigma_{\varepsilon_2}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{\varepsilon_n}^2 \end{bmatrix}.$$

$\boldsymbol{\lambda}$ (»lambda«) denotes a $n \times 1$ column vector of the slope parameters of the linear relationship in Equation 4-7. $\boldsymbol{\Phi}$ (»Phi«) is a 1×1 and $\boldsymbol{\Psi}$ (»Psi«) is an $n \times n$ matrix.

Notation 4-2: Notation for matrices and vectors

Matrices and vectors are denoted by bold letters. In case of matrices containing parameters Greek bold letters are used. Vectors are denoted by lowercase letters and matrices by uppercase letters.

The covariance matrix $\boldsymbol{\Theta}$ (»Theta«) of the true scores results from the specified matrices by mean of the matrix multiplication:

$$\boldsymbol{\Theta} = \boldsymbol{\lambda} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\lambda}^{\mathbf{T}}.$$

The symbol »$^{\mathbf{T}}$« represents the transposition of a vector or a matrix, i.e. the exchange of rows and columns. For the actual example the matrix multiplication looks like this:

Konkret sieht die dargestellte Matrizenoperation wie folgt aus:

$$\begin{aligned} \boldsymbol{\Theta} &= \boldsymbol{\lambda} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\lambda}^{\mathbf{T}} \\ &= \begin{bmatrix} 1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} \cdot \left[\sigma_\tau^2\right] \cdot \begin{bmatrix} 1 & \lambda_2 & \cdots & \lambda_n \end{bmatrix} \\ &= \begin{bmatrix} \sigma_\tau^2 & \lambda_2 \cdot \sigma_\tau^2 & \cdots & \lambda_n \cdot \sigma_\tau^2 \\ \lambda_2 \cdot \sigma_\tau^2 & \lambda_2^2 \cdot \sigma_\tau^2 & \cdots & \lambda_2 \cdot \lambda_n \cdot \sigma_\tau^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_n \cdot \sigma_\tau^2 & \lambda_n \cdot \lambda_2 \cdot \sigma_\tau^2 & \cdots & \lambda_n^2 \cdot \sigma_\tau^2 \end{bmatrix} \end{aligned}$$

The model implied covariance matrix $\boldsymbol{\Sigma}$ of the test scores results from this matrix by adding the covariance matrix $\boldsymbol{\Psi}$ of the errors:

$$\boldsymbol{\Sigma} = \boldsymbol{\Theta} + \boldsymbol{\Psi}.$$

Comment 4-2: *On the usage of matrices*

To unexperienced people the usage of matrices has usually a dissuasive effect. However, the employment of matrices (as well as of a proper computer program) actually simplifies the computation (cf. Exercise 4-2).

In addition, the usage of matrices has a mnemonic advantage since complicated formulas can often be expressed in a simple way by using matrices.
Specifically, the computation of the reliability in complex measurement models is greatly simplified by using matrices (cf. Section 4.4.2.2).

Let us now consider the second test model of CTT.

4.2.3.3 THE τ (TAU) EQUIVALENT TEST MODEL

The τ-equivalent test model constitutes a specific case of the congeneric model by fixing the regression coefficients $\lambda_i \left(i = 2,\ldots,n\right)$ in Equation 4-7 to the value 1.0.

Concept 4-6: *The model of τ-equivalent tests:*

Given:

$Y_1, Y_2, \ldots, Y_n$ observed test scores for *n* tests.

$\tau_1, \tau_2, \ldots, \tau_n$ true scores associated with the *n* tests.

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$ measurement errors associated with the *n* tests.

We chose one of the tests, e.g. the first test, as a reference with respect to which the true scores of the other tests are linear functions. The true score of the selected test is denoted by the symbol τ (no subscript).
The *n* tests are *congeneric* if the associated true scores conform to the linear relation:
The *n* tests are *essential τ-equivalent*, if the true scores conform to the following linear relationship:

$$\tau_i = \tau + \alpha_i \quad \left(i = 2,\ldots,n\right) \tag{4-10}$$

The *n* tests are *τ-equivalent* in case of: $\alpha_i = 0 \left(i = 2,\ldots,n\right)$.

The model of (essentially) τ-equivalent tests implies the following covariance matrix of the test scores:

$$\begin{bmatrix} \sigma_\tau^2 + \sigma_{\varepsilon_1}^2 & \sigma_\tau^2 & \cdots & \sigma_\tau^2 \\ & \sigma_\tau^2 + \sigma_{\varepsilon_2}^2 & \cdots & \sigma_\tau^2 \\ & & \ddots & \vdots \\ & & & \sigma_\tau^2 + \sigma_{\varepsilon_n}^2 \end{bmatrix} \tag{4-11}$$

Consequerntly the model predicts identical covariances between each of the *n* tests. The covariance predicted by the model conforms to the variance of the true scores.

Comment:
The distinction between *τ-equivalent* and *essential τ-equivalent* tests concerns the mean structure only and not the covariance structure (cf. Section 4.6.1): In case of τ-equivalent tests the observed means of the test scores are assumed to be identical for the *n* tests. This additional restriction [$\alpha_i = 0$ $(i = 2,\ldots,n)$] is not part of the essential τ-equivalent test model. The distinction is irrelevant if one is interested in the covariance structure only.

Contrary to the congeneric model the restrictions imposed by the τ-equivalent model on the covariance structure of the observed test scores is obvious. The parameters of the essential τ-equivalent model can be estimated with 2 tests only. In this case, there are 3 observed test statistics (the two variances of the 2 test scores and the covariance between the two test scores), and the model comprises 3 free parameters: σ_τ^2, $\sigma_{\varepsilon_1}^2$, and $\sigma_{\varepsilon_2}^2$. The addition of a new τ-equivalent test to *n* existing ones requires the addition of one parameter only ($\sigma_{\varepsilon_i}^2$) whereas the number of observed variances and covariances increases by $n+1$ (*n* new covariances and 1 new variance). Thus, with 3 τ-equivalent tests the model comprises 4 free parameters to explain 6 free data points.

Let us no consider the last and most restricted model of CTT that constitutes a special case of the τ-equivalent model.

4.2.3.4 THE PARALLEL TEST MODEL

The congeneric and τ-equivalent test models comprise restrictions for constraining the covariance structure of the true scores. The parallel test model adds restrictions on the error variances.

Concept 4-7: *The model of parallel tests:*

n test are *parallel* if they are τ-equivalent and, in addition the error variances are assumed to be identical:

$$\sigma_{\varepsilon_1}^2 = \sigma_{\varepsilon_2}^2 = \cdots = \sigma_{\varepsilon_n}^2 \left(= \sigma_\varepsilon^2\right).$$

The model implied covariance matrix of the observed test scores has the following structure:

$$\begin{bmatrix} \sigma_\tau^2 + \sigma_\varepsilon^2 & \sigma_\tau^2 & \cdots & \sigma_\tau^2 \\ & \sigma_\tau^2 + \sigma_\varepsilon^2 & \cdots & \sigma_\tau^2 \\ & & \ddots & \vdots \\ & & & \sigma_\tau^2 + \sigma_\varepsilon^2 \end{bmatrix} \qquad (4\text{-}12)$$

Thus the model predicts not only identical covariances between test scores but also identical variances.

The *n* tests are *strictly parallel* if they are parallel, and, in addition, the means of the test scores are identical.

Tests that are strictly parallel can be assumed to be perfectly equivalent for testing latent abilities. They provide exactly the same information about the latent ability. By consequence, instead of repeated application of the same test, one can use strictly parallel tests for assessing an examinee's ability. The mean of the test scores can be used as an estimate of the true score and their variance as an estimate of the error variance.

Similar to the τ-equivalent model, the parameters of the parallel test model can be estimated using two tests only. However, the model can also be tested with two tests since the model predicts that the variances of the two tests are equal. Note also that the addition of further parallel tests to exiting one does not require additional model parameters: The model explains the covariance structure of the tests using only two parameters: σ_τ^2 and σ_ε^2.

This terminates our presentation of the core of CTT. On the basis of the assumptions that enter the classical test models formulas for the reliabilities of (unweighted) sums of test scores can be derived. These equations are used in typical applications as estimators of the precision of the test (cf. Section 4.4.2).

Let us now consider some critical issues associated with CTT.

4.2.4 Criticism of CTT

The following criticism of CTT concerns two closely related aspects of CTT:

1. The conception of the true score as an expectation.
2. The assumption of *within-subjects* distribution of test scores for the same test.

Let us take a closer look at these two problematic aspects of CTT.

4.2.4.1 The true score concept in CTT

It is important to realize that the true score in CTT does not conform to the construct of a latent ability. Rather, the true score corresponds to the expectation of the test scores with repeated testing of the same person. To illustrate the concept, Lord and Novick (1968) present the following thought experiment:

«Suppose we ask an individual, Mr. Brown, repeatedly whether he is in favor of the United Nations; suppose further that after each question we "wash his brains" and ask him the same question again. Because Mr. Brown is not certain as to how he feels about the United Nations, he will sometimes give a favorable and sometimes an unfa-

vorable answer. Having gone through this procedure many times, we then compute the proportion of times Mr. Brown was in favor of the United Nations…» (Lord & Novick, 1968, page 29).

Accordingly, there exists a true score for each person and each test, and the true score corresponds to the expectation of the test scores of the person with repeated application of the same test. Thus the true score does not exist independently from the test: It is a test dependent quantity.

By contrast, the *latent variable conception* assumes a latent ability construct that exists independently of a test, and each person possesses a specific value on this latent construct (cf. the discussion in Chapter 1).

The true score conception faces a number of difficulties (cf. Borsboom, 2005):

1. The fact that for each test there is a true score leads to an *inflation of true scores*. For example application of a test under different light conditions results in different true scores. Consequently, variations of situational factors generate new true scores.
2. The existence of a true score does not implicate that test measures something useful. For example, a test, consisting in the addition of two numbers as quickly as possible, and to bind the shoe laces as fast as possible, has an assigned true score.
   In this way true score theory precludes a differentiation between constructs and pseudo constructs (cf. Chapter 2.3).
3. The true score conception raises the issue of the relationship between true scores and latent constructs. Obviously, the proponents of true score theory regard the main purpose of mental tests in measuring mental abilities. Since the true score is a quantity that is related to tests a relationship between true scores and mental construct has to be established.

In conclusion, the true score conception raises a number of conceptual problems that can be prevented by a latent variable conception. By consequence, the true score conception seem to be inappropriate in scientific context. The true conception may however be useful in non-scientific contexts when tests are used as a tool of selection. For example, the LSAT (law school admission test) is used as an entrance test to law schools. The empirical validity of the test, as measured, for instance, by the correlation with study success can be used for assessing the value of the test (For a detailed discussion of the concept of empirical validity, cf. Section 4.5.2).

Let us now take a critical look at the within-subjects distribution of test scores that is used for the definition of the true score.

4.2.4.2 ON THE UTILITY OF THE WITHIN-SUBJECTS DISTRIBUTION

In the thought experiment of Lord and Novick, described above, each application of the test produces a sample value from a within-subject distribution. The independence between values from different draws is assured by the brain washing procedure which eliminates memory effects and by the assumption that the other influences, represented by the errors ε_{pi} are distributed independently. The true score τ_{pi} is defined as the expectation of the resulting test scores and the variance of the within-subject distribution corresponds to the variance of the errors.

The utility of a within-subject distribution can be questioned for different reasons (cf. Borsboom, 2005; Holland, 1990):

1. Despite the fact that the thought experiments provides a frequentist interpretation of the within-subjects distribution, a sampling process as described by the thought experiment cannot be performed in reality. In order to get an estimate of the true score and the error variance, associated with the within-subject distribution, one has to resort to parallel measures.
2. The within-subject distribution is required only for the definition of the true score. It is completely irrelevant for the specification of the axioms and the classical test models of CTT.
3. In addition, the within-subjects distribution is not required for the specification of the reliability construct that is based on the *between-subject* distribution of the true scores.

In summary, the within-subject distribution is required only for the definition of the true score and completely irrelevant otherwise. However, as shown above, the true score concept is itself problematic and should be replaced by a latent variable conception that is not affected by these problems and, in addition, renders a within-subject distribution useless. It should also be noted that the concept of true score and the within-subjects distribution are unique to CTT. These constructs are not used in item response models and thus irrelevant for modern psychometrics.

Before we turn to its modern conception let us summarize the main characteristics of CTT:

1. CTT is concerned with the mean and covariance structure of test scores only. No exact distributional assumptions are made. There for the theory has also been termed *weak true score theory* in contrast to *strong true score theory* that makes precise distributional assumptions.
2. The three main aspects of CTT are:
   (a) The decomposition of an observed test score of a person into a true and error score. The true score is defined as the expectation

of a with-subject distribution, and the variance of this distribution corresponds to the variance of the error scores.

(b) The axioms of CTT specify restrictions on the covariance structure of the true and error scores. However, these restrictions are insufficient for the identification of the model parameters (i.e. the variances and covariances of true und error scores).

(c) The classical test models of CTT, the congeneric, τ-equivalent and parallel model, incorporate additional restricts that enable the identification of parameters in case of a sufficient number of tests being employed.

3. The true score concept and the within-subject distribution of test scores are not useful and are better replaced by the latent variable conception.

This presentation misses one important theoretical construct of CTT: the reliability of test of sums of test. A thorough treatment of this construct is presented in Section 4.4. Previously to the discussion of this construct, as well as the construct of validity (cf. Section 4.5), CTT will be presented in the clothes of new psychometrics, as discussed in the classical paper of Jöreskog (1971).

4.3 Representing CTT Models Using Linear Structural Equations

In the following, CTT models are represented using linear structural equation models. The resulting models are also called confirmative factor analytic (CFA) or linear structural relation (LISREL) models (cf. Notation 4-4 on page 63). The CFA models lead to the same model implied covariance and mean structure as the CTT models described above. They, thus, cannot be distinguished empirically from the models in the formulation given above. However, the interpretation of the models differs. Note also that the CFA models are but special cases of the general test model exhibited in Figure 3-1 (on page 24).

Our presentation consists of two parts. First is a discussion of the basics of structural equation models. This is followed by the representation of the models of CTT by means of linear structural equations.

4.3.1 Structural Equation Models

Here is an explication of the concept if structural equations:

Concept 4-8: *Linear and nonlinear structural equations (Mulaik, 2009; Pearl, 2009):*

A *structural equation* is an equation for representing a c*ausal* relationship between a dependent variable *y* and a set of independent variables: $x_1, x_2, \ldots, x_p$. A structural equation has the following general form:

$$y = f\left(x_1, x_2, \ldots, x_p\right) \qquad (4\text{-}13)$$

The independent variables $x_1, x_2, \ldots, x_p$ are regarded as causes, whereas the dependent variable is considered as the effect or outcome.
The symbol *f*() denotes a function whose precise form is not further specified.
In the case *linear* structural equations the equations take on the form of a multiple, linear regression equation:

$$y = \lambda_1 \cdot x_1 + \lambda_2 \cdot x_2 + \cdots + \lambda_p \cdot x_p \qquad (4\text{-}14)$$

The regression coefficients $\lambda_1, \lambda_2, \ldots, \lambda_p$ are also called *structural coefficients* or *loading coefficients*. They are free parameters that are estimated from the data.

Structural equation models comprise a set of structural equations, with a single equation for each of the dependent variables: $y_1, y_2, \ldots, y_n$.

$$\begin{aligned} y_1 &= f_1\left(x_{11}, x_{12}, \ldots, x_{1p_1}\right) \\ y_2 &= f_2\left(x_{21}, x_{22}, \ldots, x_{2p_2}\right) \\ &\vdots \\ y_n &= f_n\left(x_{n1}, x_{n2}, \ldots, x_{np_n}\right) \end{aligned} \qquad (4\text{-}15)$$

Note that the equations for different dependent variables comprise (possibly) different sets of independent variables with (possibly) different numbers of independent variables [as indicated by the indices $p_i \ (i = 1, 2, \ldots, n)$].

In the case of linear structural equations this amounts to a set of linear (regression) equations:

$$\begin{aligned} y_1 &= \lambda_{11} \cdot x_{11} + \lambda_{12} \cdot x_{12} + \cdots + \lambda_{1p_1} \cdot x_{1p_1} \\ y_2 &= \lambda_{21} \cdot x_{21} + \lambda_{22} \cdot x_{22} + \cdots + \lambda_{2p_2} \cdot x_{2p_2} \\ &\vdots \\ y_n &= \lambda_{n1} \cdot x_{n1} + \lambda_{n2} \cdot x_{n2} + \cdots + \lambda_{np_n} \cdot x_{np_n} \end{aligned} \qquad (4\text{-}16)$$

Within a system of structural equations variables can be at the same time dependent and independent variables, i.e. a dependent variable can enter the equation of another dependent variable as an independent one. In this way, dependent variables may be linked together.

Notation 4-3: Order of indices

The order of the indices conforms to the following principle

The first index is refers to the dependent variable and the second index refers to the independent variable.

Example: λ_{34} denotes the structural coefficient of the fourth independent variable within the third equation, i.e., the equation with the third dependent variable y_3 on the left-hand side of the equation.

Notation 4-4: Acronyms and abbreviations

1. In the context of structural equation modeling the following acronyms are used:

   SEM = *S*tructural *E*quation *M*odels / *M*odeling
   LISREL = *Li*near *S*tructural *Rel*ations

   The acronym *LISREL* refers, on the one hand, to the general model developed by Karl Jöreskog, and, on the other hand, to a statistical program for estimating and testing linear structural equation models.

2. For the representation of CTT models only *linear* structural equations are employed. For convenience, we shall use the notation *structural equations* to denote linear structural equations.

As already noted, structural equation models are used for causal models. The models can be best represented by means of causal diagrams.

4.3.1.1 CAUSAL DIAGRAMS AND LINEAR CAUSAL MODELS

Causal diagrams enable a convenient representation of the causal relationships between a set of variables.

Concept 4-9: *Causal diagrams:*

A *causal diagram* is a graph consisting of vertices and edges:

- Die *vertices* represent the variables.
- Die *edges* represent (predominantly) causal relationships between variables.

A causal diagram is used to depict the causal structure of a set of variables,

Figure 4-1 depicts the main components of a causal diagram that represents the causal relationships between different hidden and manifest causal variables, respectively.
The conventions and notations presented in context of the description of the general psychometric model (cf. Figure 3-1 on page 24) apply here as well (Here is a short repetition):

1. The circles and rectangles represent variables denoted by the different letters. The variables of the causal model in Figure 4-1 are: η_1, η_2, ε_1, ε_2, X, Y_1 and Y_2.

The variables can be divided into two groups:

- Unobserved (latent) variables are represented by circles and denoted by Greek letters: η_1, η_2, ε_1, and ε_2.
- Observed (manifest) variables are represented by rectangles and denoted by Latin letters: X, Y_1 and Y_2.

2. The arrows represent linear causal influences. The strength of the effect is not further specified.
3. The double arrows represent correlations and covariances, respectively, between two variables. These correlations are not explained by means of causal influences.

   The absence of a double error symbolizes the absence of a correlation (i.e. the correlation and covariance, respectively between the two variables is assumed to be zero).

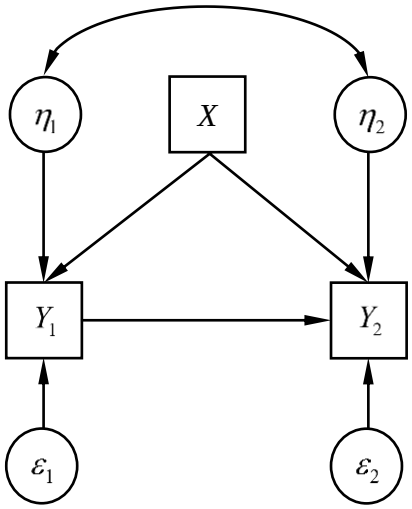


Figure 4-1: *A causal diagram representing a causal structure.*

If the causal diagram is augmented by adding model parameters the diagram represents a *causal model* (and not only a causal structure). Figure 4-2 depicts a diagram of a linear causal model.

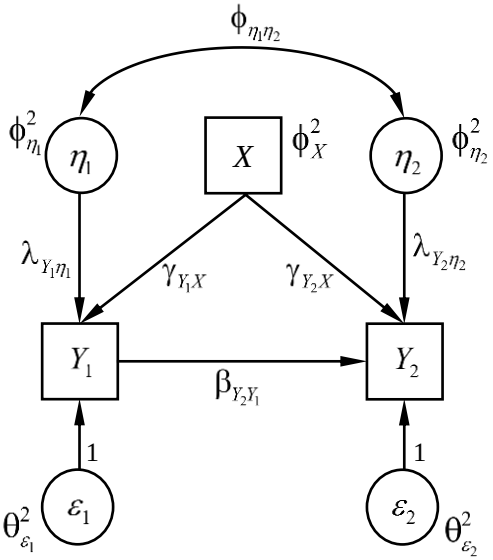


Figure 4-2: *The causal diagram of a linear causal model corresponding to the causal structure represented by the diagram in Figure 4-1.*

The following classification of variables of a causal model is of great importance.

Concept 4-10: *Exogenous vs. endogenous variables*

Exogenous variables are variables whose mean and covariance structure is not explained (or predicted) by the model.
Endogenous variables are variables whose mean and covariance structure is explained (or predicted) by the causal structure.
The distinction between the two types of variables can be identified by inspection of the causal diagram:
Endogenous variables are exactly those variables to which at least one arrow points. In addition, no double arrow can be incident at an endogenous variable.
In the model of Figure 4-2 there are two endogenous variables: Y_1 and Y_2.

Comment 4-3: *Endogenous variables, error variables, and the usage of double arrows*

In general, each (manifest or latent) endogenous variable has an error variable attached to it (cf. ε_1 and ε_2 in Figure 4-2). It represents all possible causes that are not represented by a variable in the model exerting an effect on the endogenous variable.
The variance of the error variable represents the variation of the endogenous variable that is not explained by the variables exerting a causal influence on the endogenous variable.

The absence of incident double arrows (covariance arcs) at endogenous variables results from the fact that the former represent unexplained covariances. Since the covariance structure of the endogenous variables is explained in the model by means of the causal influences of other variables in the model (as well as the covariance structure of these variables) there is no need to attach double arrows to endogenous variables.

In order to represent an unexplained covariance between two endogenous variables a covariance arc is drawn between the error terms attached to the two endogenous variables. This represents a residual covariance between two endogenous variables that cannot be explained by the model.

The parameters of linear causal models can be partitioned into the following groups (see also the descriptions of the parameters of general psychometric model of Figure 3-1 on page 24):

1. *Variances of and covariances between exogenous variables that are not error variables:* In general the Greek letter ϕ (»phi«) denotes variances and covariances of exogenous variables.
   In the model of Figure 4-2 the variances of the three exogenous variables are denoted by ϕ_X^2, $\phi_{\eta_1}^2$ and $\phi_{\eta_2}^2$ whereas the covariance between η_1 und η_2 is symbolized by $\phi_{\eta_1\eta_2}$. The missing covariance arcs between *X* and η_1 as well as *X* and η_2 indicate that these covariances are assumed to be zero: $\phi_{X\eta_1} = \phi_{X\eta_2} = 0$.
2. *Variances of and covariances between exogenous variables that are error variables:* In general the Greek letter θ (»theta«) denotes variances and covariances of error variables.
   In the model of Figure 4-2 the variances of the two error variables are denoted by $\theta_{\varepsilon_1}^2$ and $\theta_{\varepsilon_2}^2$. The missing covariance arc between ε_1 and ε_2 indicate that this covariance is assumed to be zero: $\theta_{\varepsilon_1\varepsilon_2} = 0$.
3. *Structural coefficients of arrows emanating from a manifest exogenous variable and ending in a (manifest or latent) variable:* The Greek letter γ (»gamma«) is used to denote this type of structural coefficients.
   In the model of Figure 4-2 the variable *X* is an observed exogenous variable exerting a causal influence on both Y_1 and Y_2 The associated structural coefficients have been denoted γ_{Y_1X} and γ_{Y_2X}.
4. *Structural coefficients of arrows from a latent exogenous variable to a manifest variable:* The Greek letter λ (»lambda«) is used to denote this type of structural coefficients.

In the model of Figure 4-2 the variables η_1 and η_2 are latent variables exerting an influence on the observed variables Y_1 and Y_2. The associated structural coefficients have been denoted $\lambda_{Y_1\eta_1}$ and $\lambda_{Y_2\eta_2}$.

5. *Structural coefficients of arrows between two endogenous (latent or manifest) variables:* The Greek letter β (»beta«) is used to denote this type of structural coefficients.
   In the model of Figure 4-2 the variables Y_1 and Y_2 are (manifest) endogenous variables with Y_1 exerting a causal influence on Y_2. The respective structural coefficient is represented by the symbol $\beta_{Y_2Y_1}$.
6. *Structural coefficients of arrows pointing from error terms to manifest variables:* The structural coefficients are fixed to the value 1.0.

Notation 4-5: Structural coefficients
The structural coefficients in linear structural equation models are regression weights, i.e. slope parameters of the linear regression equation.
The structural coefficients of arrows from a latent variable to an observed one (denoted by λ) are usually called *loading coefficient* or simply *loadings*.

Comment 4-4: *Missing mean parameters*
Contrary to the general psychometric model in Figure 3-1 the model of Figure 4-2 does not contain mean parameters. This is due to the fact that in the following the models are used for modeling the covariance structure only, ignoring the mean structure for the moment. In Section 4.6 where the modeling of the mean structure is discussed the mean parameters, required for modeling the mean structure are added to the model.

We have now discussed all components of a linear causal model. We next turn to the problem of recovering the linear structural equation from the diagrams of the causal model.

4.3.1.2 Determining the linear structural equations from causal diagrams of causal models

There is a simple method to get the system of linear structural equation from the diagram of a causal model.

Method 4-1: *Recovering the system of linear structural equations from the diagram of a causal model*
To determine the system of linear structural equations perform the following steps:

1. For each endogenous variable a linear equation is added. Thus, the structural equation model comprises as many equations as there are endogenous variables in the model.

2. The endogenous variable makes up the left-hand side of the equation.
3. Any variable with an arrow pointing to the endogenous target variable enters the right-hand side of the equation.
   Each of these variables is multiplied by the structural coefficient associated with the arrow from the variable to the endogenous variable, and the sum of these products makes up the right-hand side of the equation.
   Note: We thus get a linear regression equation with the endogenous variable as the dependent variable, and the variables with an arrow pointing to the endogenous variable being an independent variable. The structural coefficients make up the regression weights.
4. The structural coefficients with a value of 1.0 are suppressed in the equation. Thus only the variable enters the equation.
   The latter convention concerns error variables with an arrow pointing to the endogenous variable whose coefficient is usually 1.0 (see above).

Ex. 4-3: Determining the system of linear structural equation from the causal diagram:

Given: The causal model depicted in Figure 4-2 (page 65).

The model comprises 2 endogenous variables: Y_1 and Y_2. The accompanying system of linear structural equations looks like this:

$$Y_1 = \gamma_{Y_1 X} \cdot X + \lambda_{Y_1 \eta_1} \cdot \eta_1 + \varepsilon_1$$

$$Y_2 = \gamma_{Y_2 X} \cdot X + \lambda_{Y_2 \eta_2} \cdot \eta_2 + \beta_{Y_2 Y_1} \cdot Y_1 + \varepsilon_2$$

Notice that linear structural equation model contains information only about the causal (or structural) relations but not about covariance structure of the exogenous variables. Consequently, the causal diagram of the model contains more information than the system of linear structural equations.

4.3.1.3 PREDICTING AN OBSERVED COVARIANCE STRUCTURE USING LINEAR CAUSAL MODELS

The system of linear structural equations represents the causal model of the scientist, that is, it implements her assumptions of how the observed data have been generated. In order to test the model empirically, predictions are derived from the model and compared with the data. The predictions are concerned with the covariance and mean structure of the endogenous variables (the variances, covariances and means). For the moment the mean structure will be ignored.

Let us now consider how predictions are derived from the linear structural equations and the covariance structure of the exogenous variables. To this end we assume that we have observed *n* endogenous variables $Y_1, Y_2, \ldots, Y_n$ (e.g. the results of *n* test items). The covariance matrix of these *n* variables is a $n \times n$ matrix with the variances of the variables on the main diagonal and the covariances above and below the main diagonal:

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & \cdots & Y_n \end{array} \\ \begin{bmatrix} \sigma^2_{Y_1} & \sigma_{Y_1Y_2} & \cdots & \sigma_{Y_1Y_n} \\ \sigma_{Y_2Y_1} & \sigma^2_{Y_2} & \cdots & \sigma_{Y_2Y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{Y_nY_1} & \sigma_{Y_nY_2} & \cdots & \sigma^2_{Y_n} \end{bmatrix} \end{array}$$

Since the covariance matrix is symmetric, i.e. $\sigma_{Y_iY_j} = \sigma_{Y_jY_i}$, the entries of the rows are identical to the entries of the columns. Thus, the first line of the matrix is identical the second one: $\sigma_{Y_1Y_2} = \sigma_{Y_2Y_1}$, $\sigma_{Y_1Y_2} = \sigma_{Y_2Y_1}$, …, $\sigma_{Y_1Y_n} = \sigma_{Y_nY_1}$.

As noted above (cf. Section 4.1), the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Y}}$ contains $n \cdot (n+1)/2$ unique variances and covariances. For each of these $n \cdot (n+1)/2$ unique variances and covariances a so called *covariance equation* has to be set up on the basis of the linear causal model.

Concept 4-11: *Covariance equations:*

A *covariance equation* is an equation with the to be modeled (or predicted) variance or covariance on the left-hand side.

The right-hand side of the covariance equation is made up by an expression that is a function of the model parameters of the causal model, i.e. the variances and covariances of the exogenous variables and the structural coefficients.

Consequently, a covariance equation represents the variances and covariances of the observed quantities in terms of the parameters of the causal model.

The generation of the covariance equations proceeds in two steps:

1. Reduction of the linear structural equations: Endogenous variables that appear on the right-hand side of the structural equations are replaced successively by the right-hand sides of the accompanying structural equation until there are no more endogenous variables on the right-hand sides of the linear equations (for details on the method of reducing linear structural equations (cf. *Basics of covariance algebra*, Method 2-3).

2. The variances and covariances of the observed endogenous variables are computed using the covariance algebra (cf. *Basics of covariance algebra*).

With respect to our causal model in Figure 4-2 (page 65) the two steps look like this:

Ex. 4-4: Specification of the covariance equations:

Given: The system of linear structural equations for the causal model of Figure 4-2 (page 65):

$$Y_1 = \gamma_{Y_1X} \cdot X + \lambda_{Y_1\eta_1} \cdot \eta_1 + \varepsilon_1$$
$$Y_2 = \gamma_{Y_2X} \cdot X + \lambda_{Y_2\eta_2} \cdot \eta_2 + \beta_{Y_2Y_1} \cdot Y_1 + \varepsilon_2$$

1. Reduction of the equation for Y_2 by replacing the endogenous variable Y_1 by its equation and collection coefficients for each exogenous variable:

   $$Y_2 = \left(\gamma_{Y_2X} + \beta_{Y_2Y_1} \cdot \gamma_{Y_1X}\right) \cdot X + \beta_{Y_2Y_1} \cdot \lambda_{Y_1\eta_1} \cdot \eta_1 + \lambda_{Y_2\eta_2} \cdot \eta_2 + \beta_{Y_2Y_1} \cdot \varepsilon_1 + \varepsilon_2$$

   The other equation need not be reduced since there are only exogenous variables on the right-hand side.
2. Computation of variances and covariances of the endogenous variables.

   The model comprises two endogenous variables, By consequence two observed variances: $\mathrm{Var}(Y_1)$ and $\mathrm{Var}(Y_2)$, as well as one covariance $\mathrm{Cov}(Y_1, Y_2)$ are to be explained.

   Using the simplified rules of covariance algebra (cf. *Basics of covariance algebra*, Method 2-4) results in:

$$\mathrm{Var}(Y_1) = \gamma_{Y_1X}^2 \cdot \mathrm{Var}(X) + \lambda_{Y_1\eta_1}^2 \cdot \mathrm{Var}(\eta_1) + \mathrm{Var}(\varepsilon_1)$$

$$\begin{aligned}\mathrm{Var}(Y_2) = {} & \left(\gamma_{Y_2X} + \beta_{Y_2Y_1} \cdot \gamma_{Y_1X}\right)^2 \cdot \mathrm{Var}(X) + \left(\beta_{Y_2Y_1} \cdot \lambda_{Y_1\eta_1}\right)^2 \cdot \mathrm{Var}(\eta_1) + \\ & \lambda_{Y_2\eta_2}^2 \cdot \mathrm{Var}(\eta_2) + \beta_{Y_2Y_1}^2 \cdot \mathrm{Var}(\varepsilon_1) + \mathrm{Var}(\varepsilon_2) + \\ & 2 \cdot \lambda_{Y_1\eta_1} \cdot \lambda_{Y_2\eta_2} \cdot \beta_{Y_2Y_1} \cdot \mathrm{Cov}(\eta_1, \eta_2)\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(Y_1, Y_2) = {} & \gamma_{Y_1X} \cdot \left(\gamma_{Y_2X} + \beta_{Y_2Y_1} \cdot \gamma_{Y_1X}\right) \cdot \mathrm{Var}(X) + \beta_{Y_2Y_1} \cdot \lambda_{Y_1\eta_1}^2 \cdot \mathrm{Var}(\eta_1) + \\ & \beta_{Y_2Y_1} \cdot \mathrm{Var}(\varepsilon_1) + \lambda_{Y_1\eta_1} \cdot \lambda_{Y_2\eta_2} \cdot \mathrm{Cov}(\eta_1, \eta_2)\end{aligned}$$

Notation 4-6: Operator- vs. parameter notation:
For the representation of the variances and covariances the operator notation has been employed (cf. *Basics of covariance algebra*). Using the parameter notation the equations look like this:

$$\sigma_{Y_1}^2 = \gamma_{Y_1X}^2 \cdot \phi_X^2 + \lambda_{Y_1\eta_1}^2 \cdot \phi_{\eta_1}^2 + \theta_{\varepsilon_1}^2$$

$$\sigma^2_{Y_2} = \left(\gamma_{Y_2X} + \beta_{Y_2Y_1} \cdot \gamma_{Y_1X}\right)^2 \cdot \phi^2_X + \left(\beta_{Y_2Y_1} \cdot \lambda_{Y_1\eta_1}\right)^2 \cdot \phi^2_{\eta_1} +$$
$$\lambda^2_{Y_2\eta_2} \cdot \phi^2_{\eta_2} + \beta^2_{Y_2Y_1} \cdot \theta^2_{\varepsilon_1} + \theta^2_{\varepsilon_2} + 2 \cdot \lambda_{Y_1\eta_1} \cdot \lambda_{Y_2\eta_2} \cdot \beta_{Y_2Y_1} \cdot \phi_{\eta_1\eta_2}$$
$$\sigma_{Y_1Y_2} = \gamma_{Y_1X} \cdot \left(\gamma_{Y_2X} + \beta_{Y_2Y_1} \cdot \gamma_{Y_1X}\right) \cdot \phi^2_X + \beta_{Y_2Y_1} \cdot \lambda^2_{Y_1\eta_1} \cdot \phi^2_{\eta_1} +$$
$$\beta_{Y_2Y_1} \cdot \theta^2_{\varepsilon_1} + \lambda_{Y_1\eta_1} \cdot \lambda_{Y_2\eta_2} \cdot \phi_{\eta_1\eta_2}$$

In our example the parameters cannot be uniquely estimated from the observed data since there are many more free parameters than free data points. The model is thus not identified.

Comment 4-5: *Modeling variances and covariances by means of structural equation models and models of classical test theory.*

Please note that our description of the modeling of variances and covariances by means of linear structural equations reflects the description of modeling of variances and covariances using the assumptions of the classical test models (cf. Ex. 4-1 [page 47] and Ex. 4-2 [page 52]).

This suggests that structural equation models can be used for modeling the classical test models.

Following to this exposition of linear structural equation models we are now in the position to express the test models of CTT as structural equation models.

4.3.2 Representing the Test Models of CTT as Linear Structural Equation (LISREL) Models

It will now be shown that the test models of CTT can be expressed as linear structural equation models that result in the same predictions with respect to the covariance and mean structure (the mean structure will be ignored for the moment). It should be noted that linear structural equations and classical test model are based on different assumptions (cf. Section 4.3.3). However, the divergent assumptions do not result in different predictions with respect to the covariance and mean structure of the test items.

4.3.2.1 Representation of the General Test Model of CTT

Let us consider once again the basic equation of CTT:

$$Y_i = \tau_i + \varepsilon_i$$

Obviously, the equation can be conceived of as a simple linear structural equation if the latent variables τ_i and ε_i are regarded as variables exerting a causal influence on the observed test score. Figure 4-3 depicts the associated causal diagram.

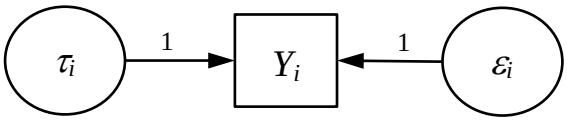


Figure 4-3: *A causal diagram of the structural equation representing the basic equation of CTT.*

The fact that the basic equation of CTT can be conceived of as a linear structural equation constitutes the foundation for the representation of the classical test models by means of linear structural equation (LISREL) models.

Figure 4-4 exhibits the causal diagram of the general test model that incorporates the axioms of CTT.

Notation 4-7: LISREL Notation:

The symbols for denoting latent variables and parameters has been changed to the notation used with linear structural equation models, specifically:

1. The symbol τ, denoting true score variables, has been replaced by the symbol η symbolizing latent constructs.
2. The symbols σ_τ^2 and $\sigma_{\tau_i\tau_j}$, representing the variance and covariance, respectively, of the true scores have been replaced by the symbols ϕ_i^2 and ϕ_{ij}.
3. The symbol σ_ε^2, denoting the variance of the error variables have been replaced by θ_i^2.

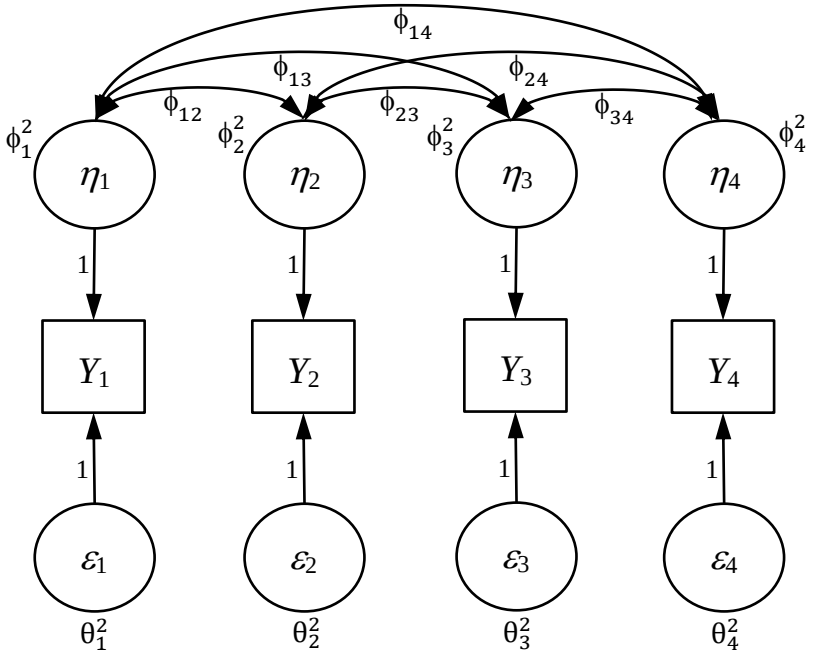


Figure 4-4: *Causal diagram of the LISREL model representing the general classical test model with 4 test items implementing the axioms of CTT.*

The causal diagram enables the determination of the linear structural equations. (cf. Method 4-1 on page 67). Application of covariance algebra (cf. *Basics of covariance algebra*) reveals that the model implied covariance matrix is exactly the same as the one resulting from using the axioms of CTT for imposing restrictions on (cf. Ex. 4-1 on page 47) and applying covariance algebra. It turns (Exercise 4-6) out that the resulting structure of the implied covariance matrix is identical to that of the general classical test model (cf. Ex. 4-1 on page 47).

The causal diagram of the LISREL model implements the axioms of CTT in the following way:

- $\mathrm{Kov}(\tau_i, \varepsilon_i) = 0$ is represented by the fact that the covariance arcs between circles representing error variables and those representing latent construct of the *same observed variable* are missing.
- $\mathrm{Kov}(\tau_i, \varepsilon_j) = 0$ is implemented by missing covariance arcs between circles representing error variables and circles representing latent constructs for *different observed variables*.
- $\mathrm{Kov}(\varepsilon_i, \varepsilon_j) = 0$ is implemented by missing covariance arcs between circles representing error variables for different observed variables.

Consequently, the axioms of CTT are implemented within the causal model by means of restrictions on specific model parameters: The re-

levant covariances are restricted to be zero. In the causal diagrams this is represented by missing covariance arcs.
This way of representation enhances the comprehensibility of the significance of the axioms. This will become even more prominent in case of representing the three classical test models by means of causal diagrams.

4.3.2.2 REPRESENTATION OF THE CLASSICAL TEST MODELS

Let us now reconstruct the three classical test models by means of linear structural equation (LISREL) models.

4.3.2.2.1 The model of congeneric tests

Figure 4-5 depicts the LISREL model of the model of four congeneric tests. The figure reveals that the central aspect of the congeneric model consists in the fact that the four tests, Y_1, Y_2, Y_3 and Y_4, are measures of the same latent construct η.

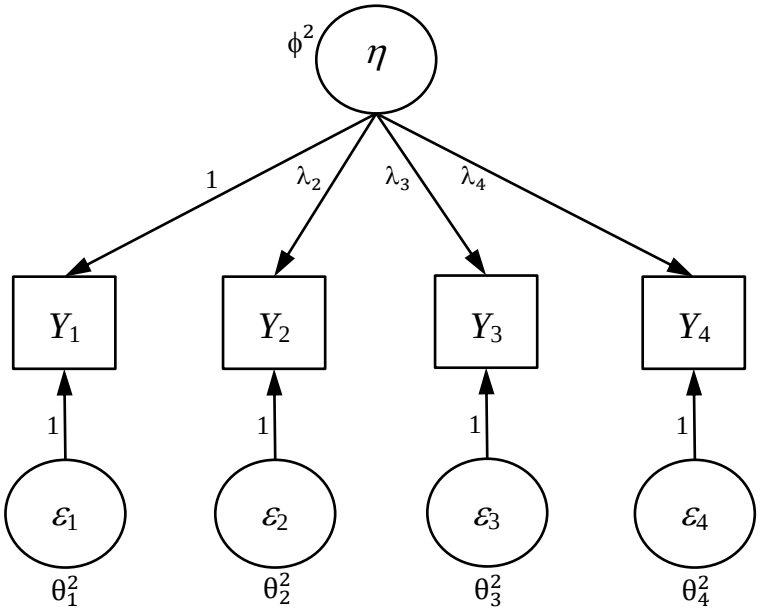


Figure 4-5: *Causal diagram of the LISREL model of four congeneric tests.*

The figure also reveals that the four measures may measure the latent construct in the same way and my thus not be equally suitable for measuring the latent construct. This is evidenced, one the hand, by the (possibly) different loading coefficients representing the strength of the causal influence that is exerted by the latent construct on the measures.
On the other hand, the error variances may be different. The quality of a measure increases with the size of the loading coefficient and with a decrease of the error variance since in this case the measurement is predominantly influenced by the measured construct rather than by other sources of influence (represented by the error variable). Thus

measures with a great loading coefficient and small error variance are to be preferred.
The loading coefficient of the first test has been fixed to 1.0. This fixes the scale of the latent variable.

Principle 4-1: *Fixing the scale and location of latent constructs*

Latent variables have no scale (unit of measurement) and no location. By consequence the scale and the location of the latent construct has to be fixed.
Fixing the scale of latent constructs can be done in two different ways:

1. By setting the variance of the latent variables to a fixed value (usually 1.0), or
2. By setting the loading coefficient of an observed variable to a fixed value (usually 1.0).

The location of latent variables can also be fixed in two different ways:

1. By setting the mean of a latent construct to a specific value (in general 0).
2. By setting the intercept parameter of one observed variables associated with a latent construct to a specific value (in general 0).

Comment: The fixing of the location is required only in case of modeling mean structure of tests.

Similar as for the general test model of CTT it can be shown, using covariance algebra (Exercise 4-6), that the model implied covariance matrix is identical to that of the congeneric test model of (Ex. 4-2 on page 52). Thus both types of models lead to the same empirical predictions.
The model is identified if at least 3 tests are available for measuring the latent construct. In case of 4 or more test items the model can be tested statistically.

4.3.2.2.2 The model of essential τ-equivalent tests

Figure 4-6 depicts the causal diagram of the LISREL model representing the τ-equivalent test model. It results from the congeneric model by setting each of the loading coefficients to 1.0.

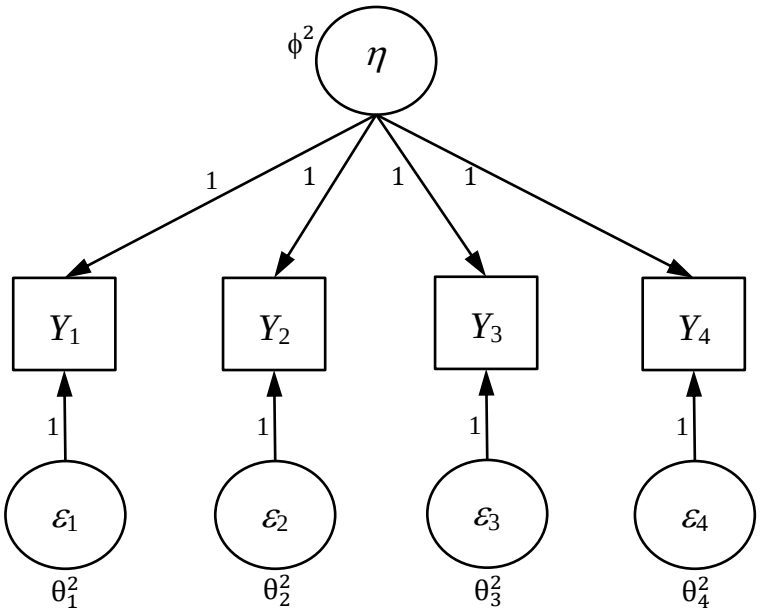


Figure 4-6: *Causal diagram of the LISREL model of four τ-equivalent tests.*

It thus implements additional restrictions on the parameters. These restrictions implement the assumption of an equal causal influence of the latent construct on the measured variables.

As for the congeneric model, in the case of τ-equivalent test models, the LISREL model results in the same implied covariance matrix as the τ-equivalent test model (Exercise 4-6).

Due to the additional restriction the model parameters are identified in case of three test items. In this case the model can also be tested which results from the fact that in case of 3 test items the model predicts the equality of the covariances between observed test items: $\sigma_{12} = \sigma_{13} = \sigma_{23}$, where σ_{12} denotes the covariance between the first and second test, σ_{13} represents the covariance between Items 1 and 3, and σ_{23} symbolizes the covariance between Test 2 und 3. The prediction of equal covariances can thus be tested by comparing the model predicted covariances with the observed ones.

4.3.2.2.3 The parallel test model

Figure 4-7 exhibits the causal diagram of the LISREL model implementing the parallel test model.

The diagram differs from that in Figure 4-6 only by the fact that the parameters representing the error variances are the same and are, thus, represented by the same symbol: $\theta_1^2 = \theta_2^2 = \theta_3^2 = \theta_4^2 = \theta^2$.

Again, the model implied covariance matrix conforms to that of the parallel model. The model requires two test for identification of the

two parameter ϕ^2 and θ^2. With two test items the model can also be tested empirically since it predicts equal variances of the 2 test items.

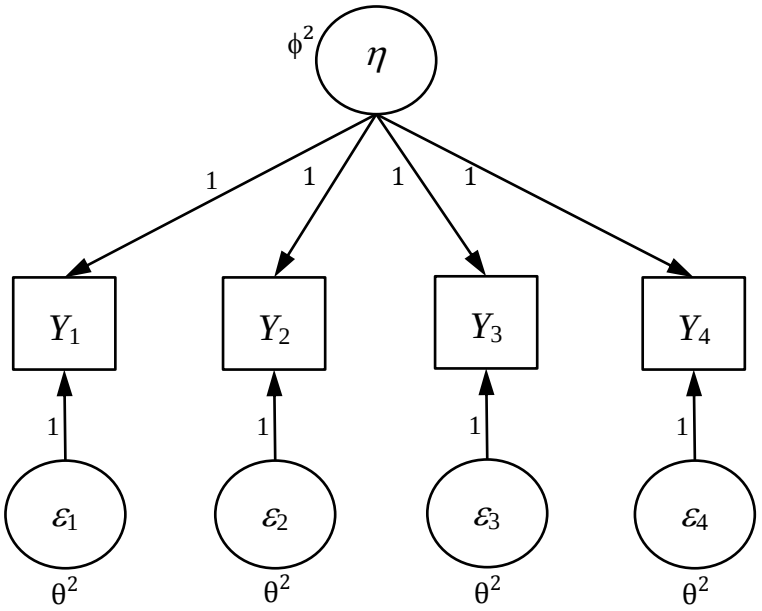


Figure 4-7: *Causal diagram of the LISREL model of four parallel tests.*

The structural equation model depicted in Figure 4-7 reveals why parallel tests are perfectly equivalent measurement instruments: The causal influences of the latent construct and of the error are exactly the same for each test item.

To illustrate the different models and how they may be used to draw conclusion about the structure of tests, let us consider an example of Jöreskog (1971) that analyzes data from Lord using structural equation models.

Ex. 4-5: Analyzing the structure of tests using structural equation models (Jöreskog, 1971):

Given: Four different vocabulary tests:

X_1, X_2 represent two tests, consisting of 15 items each, that have been applied without time pressure.

Y_1, Y_2 represent two tests consisting, of 75 items each, that have been applied under time pressure.

The number of examinees was $N = 649$. Tab. 4-1 contains the covariance matrix of the 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. 4-1: *Covariance matrix of four tests (From Jöreskog, 1971).*

Jöreskog investigated the following 4 hypotheses:

H_1: X_1 and X_2, as well as Y_1 und Y_2 are parallel. The two pairs are not congeneric, however.

H_2: X_1 and X_2, as well as Y_1 und Y_2 are parallel. The four test are congeneric.

H_3: X_1 and X_2, as well as Y_1 und Y_2 are congeneric. The four measures together are not congeneric, however.

H_4: The four 4 tests are congeneric but not parallel.

Explanation of the logic behind the hypotheses:

The 4 hypotheses refer to 2 aspects of the tests:

1. The first aspect concerns the question of whether the two speeded tests measure the same construct as the two tests without time constraints. The amounts to the question of whether the four tests are congeneric.
   The hypotheses H_2 and H_4 claim that the 4 tests are congeneric, thus assuming that both types of tests (with vs. without time constraints) measure the same construct.
   The hypotheses H_1 and H_3 do not assume that the two types of tests measure the same construct.
2. The second concerns the question of whether the two subtests X_1, and X_2 as well as Y_1, and Y_2 are parallel forms The hypotheses H_1 und H_2 assume that this is the case, whereas the hypotheses H_3 and H_4 assert that this not the case.

Tab. 4-2: *Assertions of the 4 hypotheses with respect to two aspects of the four tests: Identical construct and parallel subtests*

	Subtests parallel	
Identical construct (congeneric)	Yes	No
Yes	H_2	H_4
No	H_1	H_3

Tab. 4-2 depicts a cross classification of the two aspects and the claims of the four hypotheses. For example, hypothesis H_2 makes the strongest assertion that the four tests are congeneric and the two subtests are parallel forms.

An elegant method to test the 4 hypotheses consists in the generation of model that conforms to the most general (i.e. the most unrestricted hypothesis). The other models are then based on this model by specifying relevant restrictions conforming to the hypothesis.

The most general model is given by hypothesis H_3. The associated SEM is shown in Figure 4-8.

The model accompanying hypothesis H_1 results from this model by specifying the following restrictions:

$$\lambda^X_1 = \lambda^X_2 = \lambda^X \qquad \lambda^Y_1 = \lambda^Y_2 = \lambda^Y \qquad \text{and} \qquad \theta^{X\,2}_1 = \theta^{X\,2}_2 = \theta^{X\,2} \qquad \theta^{Y2}_1 = \theta^{Y2}_2 = \theta^{Y\,2} \;.$$

The resulting model is identified by fixing the scale of the two latent variables. This is accomplished by fixing the variance parameters: $\phi^2_{\eta_X} = 1$, and $\phi^2_{\eta_Y} = 1$.

The model representing H_2 results from this model by adding the restriction $\phi_{\eta_X \eta_Y} = 1$. This corresponds to a perfect correlation between the latent constructs η_X and η_Y : $\rho_{\eta_X \eta_Y} = 1$.

Comment:

The model with variance and covariance parameters for the latent constructs all equal to 1.0 corresponds to the congeneric model.

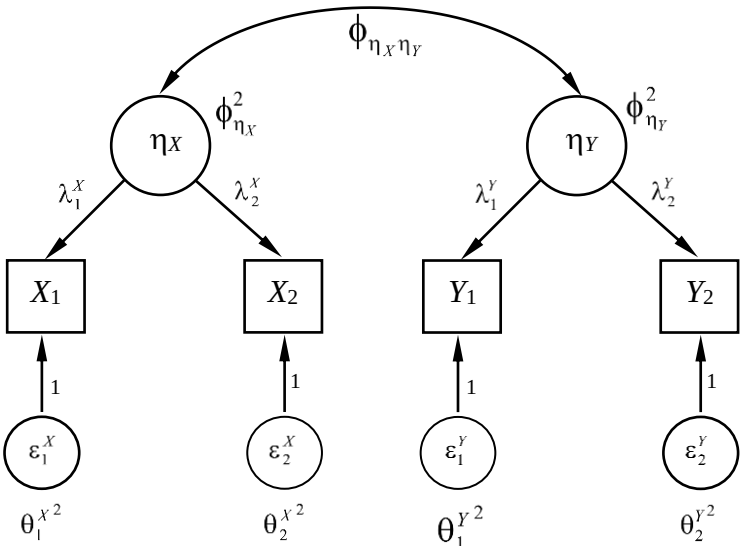


Figure 4-8: *SEM model for testing hypothesis H_3.*

Finally, the model for testing hypothesis H_4 results from model H_3 by adding the constraint $\phi_{\eta_X \eta_Y} = 1$ to the latter.

Note:
The equivalence $\phi_{\eta_X \eta_Y} = 1 \Leftrightarrow \rho_{\eta_X \eta_Y} = 1$ results due to the fact that the variance parameters have been fixed to $\phi^2_{\eta_X} = 1$ and $\phi^2_{\eta_Y} = 1$.

The previous exposition has demonstrated that linear structural equation models result in the same predications as classical test models in their original formulation. Both types of models are thus empirically not discriminable. There exist however differences with respect to their interpretation. These will be discussed in the following section.

4.3.3 On the Difference between the Classical and the SEM Approach

The structural equation modeling (SEM) approach differs from the classical approach in three respects:

1. The SEM approach is based on a latent variable conception and not on the true score conception of classical test theory. As shown in Section 4.2.4.1, the true score approach comprises a number of critical aspects that are not shared by the latent variable approach.
2. The SEM approach puts forward causal assumptions that are implemented by means of the model equations.
3. The SEM approach makes distributional assumptions whereas assumptions of CTT are concerned with covariances and correlations only. Lord and Novick (1968) thus termed CTT as *weak true score theory* which they contrasted to *strong true score theory* that incorporates distributional assumptions (see Concept 4-1 on page 45).

In the following, the causal and distributional assumptions are considered more closely.

4.3.3.1 TEST AND MEASUREMENT MODELS AS CAUSAL MODELS

As mentioned above (Section 4.3.1), structural equation models serve for the modeling of causal relationships. Consequently, linear structural equation models conceptualize the test models of CTT as causal models. Here is a definition of test models from a causal perspective.

Concept 4-12: *Test model / Measurement model:*

A test or measurement model is a causal model representing the measurement or test situation. It thus models how the observed test scores evolved due to the causal influences of various causal factors.

The model consists of the following components:

(i) The set of observed measures (indicators, test items etc.).
(ii) The set of relevant causal variables and their relationships exerting a direct causal influence on the observed measures.
(iii) A detailed specification of the causal relationship between measures and the causal variables exerting an influence on the former.

Concerning the test models of CTT there are two different causal factors exerting an influence on the test items:

1. The latent constructs η that are measured by the test item. A test item is can be a reliable and valid measure of the target construct it intends to measure only if the latent construct exerts a causal influence on the test item. By consequence, a higher score on the latent construct should result in a higher score on the measured variable. Note that there is no causal influence in the reverse direction, i.e., an increased test score does not lead to a higher score of the latent ability. It can only indicate that the ability might be high but never causally influence the latent construct.
2. The error term ε represents the residual causal influences resulting variation of the measurements that is considered as measurement error. The models assume that the causal factors represented by ε are uncorrelated with the target constructs measured by the test item.

Linear structural equations enable the modeling of additional causal factors that as relevant influence factors. Specifically, multitrait-multimethod (MTMM) models (cf. Figure 4-36) are used to model the effect of different methods used to measure latent traits, and latent trait state (LTS) models enable the modeling of situational influences by measuring latent traits on different occasions. These extensions of the classical models will not be treated here.

4.3.3.1.1 Criticism of causal conceptualizing measurement models

The conceptualization of test models as causal models has been criticized by Zumbo (2007). He argues that the concept of causality is itself too problematic to be used as a foundational concept for defining measurement models and measurement concepts like validity. There are two main counter-arguments against this type of criticism:

(1) Modern science is full of problematic concepts that are nevertheless useful and commonly employed. Examples are the concepts of probability and the construct of a natural law. Up to now no one has developed a generally accepted conceptualization of these constructs.
(2) Recent developments on causal networks and causal reasoning have resulted in a deep understanding of the concept of causality. (cf., e.g., Pearl, 2009). Specifically, Schurz and Gebharter (2016) argue that causality is a theoretical construct that results in empirical testable predictions. Schurz and Gebharter specify conditions, in terms of axioms, that enable the empirical testing of causal structures (see also Spirtes, Glymour, & Scheines, 1993).

Due to these arguments, the criticism of Zumbo (2007) does not seem to be well-founded. Note however, that the predictions of structural models are not based on the causal interpretation of the SEM models. It has been noted above in the context of the discussion of the cognitive function of theoretical constructs (cf. Section 2.1.3) that within a formalized theory theoretical constructs are but variables (or placeholders) whose semantic content is irrelevant for the theoretical predictions of the theory.

Similarly, the predictions of structural equation models are solely based on the form of the structural equations, and the covariance structure of the exogenous variables. The predictions concerning the covariance structure of the observed test scores are computed from these two elements of the SEM model by means of covariance algebra. The assumption of causality is, thus, not required for deriving the model predictions.

The causal interpretation of measurement and test models, respectively, serves a similar cognitive function as theoretical constructs in general: They enable a better understanding and a coherent interpretation of the test models. The assumption that latent traits and latent states, as well as other aspects of the test situation, exert a causal influence on the observed test scores seems to be quite a natural and uncontroversial assumption.

4.3.3.2 DISTRIBUTIONAL ASSUMPTIONS

As mentioned above (Section 4.2.1), CTT is concerned with means, variances, and covariances (or correlations). There are no assumptions concerning the distributions of test scores, true scores, and errors. This renders CTT very general. However, as an unfortunate consequence, the empirical adequacy of the test models cannot be tested statistically.

By contrast, linear structural equation model make use of distributional assumptions. Specifically, it is assumed that the latent abilities and the error variables conform to a (joint) multivariate normal distribution. This enables one to perform statistical tests of the models (cf. Exercise 4-8).

It should be stressed however that the assumption of multivariate normality is not an inherent feature of linear structural equation models. It is but a convenient assumption that enables the efficient estimation of the models.
This ends our discussion of the modeling of classical test models and performing test item analysis, respectively, using linear structural equation models. We now turn to the discussion of reliability that constitutes probability one of the most important concepts of classical test theory.

4.4 Reliability: Concept and Estimation

The reliability of a test is one of the most important concepts of CTT. In this section we investigate this concept. Following to an exposition of the concept, we investigate how the concept is used in traditional approaches (Section 4.4.1). This is followed by a treatment of the reliability of the unweighted sum of test scores of individual items (Section 4.4.2). Finally, the concept of maximal reliability and the optimal weighting of test scores will be discussed (Section 4.4.3).
The concept of reliability is based on the concept of the true score variance of a test.

Concept 4-13: *True score variance*

The *true score variance* of a test corresponds to that variation of the test that is due to variation of the latent constructs.

Comments:

1. The true score variance has to be distinguished strictly from the variance of the latent constructs.
2. The *true score variance* concerns those variations of the measurement that are due to variations of the latent constructs.

The concept of reliability can now be defined using the concept of the true score variance.

Concept 4-14: *Reliability*

The *reliability of a test* is the proportion of the true score variance within the complete variance of the test scores. Consequently, the reliability corresponds to the quotient of the true score variance and the variance of the test scores.

The concepts of *true score variance* and *reliability* are theoretical constructs since they cannot be observed but have to be estimated on the basis of an underlying psychometric model. This sort of measurement is called *model dependent measurement*. By consequence the reliability and its estimation is based on the correctness of the assumptions that are incorporated into the psychometric model. In case of the

model approximating the population poorly, the estimated reliabilities are usually biased.
The following example exhibits the relevance of the concept of reliability for the measurement of latent constructs.

Ex. 4-6: Significance of the concept of reliability:

Given: The test scores of 2 examinees on an intelligence test: $Y_1 = 102$ and $Y_2 = 105$.

The test scores indicate that Examinee 2 might have a slightly higher value in the latent intelligence construct than Examinee 1. However, in the light of existing measurement errors, one might question whether this conclusion is really justified.

The variance of the two test scores is $(Y_1 - Y_2)^2 / 2$. It is, thus, a function of the difference of the test scores.

The validity of a conclusion from the observed test scores to the values of the underlying latent constructs depends on how the test scores were created. Let us consider to different extreme scenarios:

(3) There was no measurement error present. By consequence, the variation of test scores (i.e. the difference between test scores) has been determined completed by the variation of the latent construct (i.e. the difference between the two values of the latent construct). The observed variance of test scores corresponds to the true score variance, and the reliability of the test is thus 1.0.

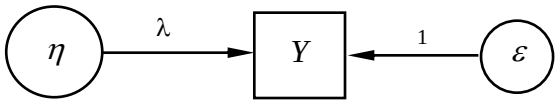
(4) The observed difference was completely caused by measurement error. The contribution of the variation of the true scores to the variation of the test scores is thus zero. Consequently, the true score variance, as well as the reliability of the test, is zero.

Obviously, a conclusion from the observed difference of test scores on the difference of the underlying constructs is perfectly justified for the first scenario and completely unjustified for the second one. This illustrates the importance of the reliability with respect to inference of differences between the latent constructs on the basis of differences between observed test scores.

The next example demonstrates the computation of the reliability of a single test item, assuming a simple psychometric model.

Ex. 4-7: Specification of the reliability of a test item:

Given: A simple measurement model (Figure 4-9):



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

Figure 4-9: *Causal diagram of a simple linear psychometric model.*

The variance of the observed score *Y* can be partitioned into two variance components. This is easily demonstrated using covariance algebra:

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Cov}(Y, Y) \\ &= \mathrm{Cov}(\lambda \cdot \eta + \varepsilon, \lambda \cdot \eta + \varepsilon) \\ &= \lambda^2 \cdot \mathrm{Cov}(\eta, \eta) + \mathrm{Cov}(\varepsilon, \varepsilon) \\ &= \lambda^2 \cdot \mathrm{Var}(\eta) + \mathrm{Var}(\varepsilon)\end{aligned}$$

The first summand on the right-hand side represents the true score variance and the second summand the error variance. Dividing both sides of the equation by $\mathrm{Var}(Y)$ results in:

$$1 = \frac{\lambda^2 \cdot \mathrm{Var}(\eta)}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}(\varepsilon)}{\mathrm{Var}(Y)}$$

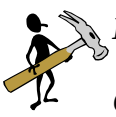
The first term on the right-hand side represents the proportion of the true score variance within the total observed variance of the test scores. The reliability of the test is thus:

$$\mathrm{Rel}(Y) = \frac{\lambda^2 \cdot \mathrm{Var}(\eta)}{\mathrm{Var}(Y)}. \qquad (4\text{-}17)$$

It can be demonstrated that, in the present case, the reliability of *Y* corresponds to the squared correlation between *Y* and the latent construct η (cf. Exercise 4-10):

$$\mathrm{Rel}(Y) = \frac{\left[\mathrm{Cov}(\eta, Y)\right]^2}{\mathrm{Var}(\eta) \cdot \mathrm{Var}(Y)} = R^2_{Y.\eta} = R^2_{Y\eta}. \qquad (4\text{-}18)$$

The decomposition of the variance of a test score into the true score and error variance is unique as long as errors and latent constructs are uncorrelated. In this case the reliability of a test score can be computed using the estimated parameters resulting from a structural analysis of the test items.



Method 4-2: *Decomposition of the variance of an observed test score into true score and error variance:*

Given: The linear structural equation of the observed test score *Y*: $Y = \lambda_1 \cdot \eta_1 + \lambda_2 \cdot \eta_2 + \cdots + \lambda_p \cdot \eta_p + \varepsilon$

Assumption: $\mathrm{Cov}(\eta_i, \varepsilon) = 0$, for all $i = 1, \ldots, p$.

ϕ_1^2

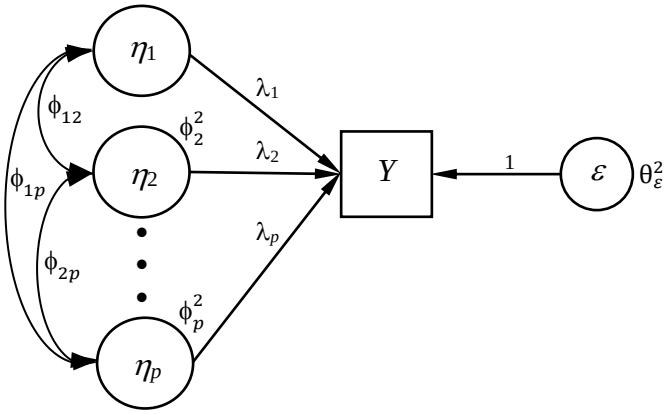


Figure 4-10: *Measurement model for illustrating the decomposition of the observed variance into true score and error variance.*

The desired decomposition follows from the measurement model by means of covariance algebra using the assumption of the model. For the given model in Figure 4-10 the following decomposition is obtained:

$$\begin{aligned}\mathrm{Var}(Y) = {} & \lambda_1^2 \cdot \mathrm{Var}(\eta_1) + \lambda_2^2 \cdot \mathrm{Var}(\eta_2) + \cdots + \lambda_n^2 \cdot \mathrm{Var}(\eta_p) + \\ & + 2 \cdot \lambda_1 \cdot \lambda_2 \cdot \mathrm{Cov}(\eta_1, \eta_2) + 2 \cdot \lambda_1 \cdot \lambda_3 \cdot \mathrm{Cov}(\eta_1, \eta_3) + \cdots \\ & + 2 \cdot \lambda_{p-1} \cdot \lambda_p \cdot \mathrm{Cov}(\eta_{p-1}, \eta_p) + \mathrm{Var}(\varepsilon)\end{aligned}$$

The term $\mathrm{Var}(\varepsilon)$ represents the error variance. The sum of the other terms constitute the true score variance.

Since $\mathrm{Cov}(\eta_i, \varepsilon) = 0$, for all *i*, there are no covariance terms including error variables and variables denoting latent constructs. The decomposition is thus unique.

In case of $\mathrm{Cov}(\eta_i, \varepsilon) \neq 0$, for some of the *i*, the decomposition is not unique since it is unclear whether to assign these terms to the true score or to the error variance. In this case one could compute a lower bound of the true score variance and the reliability, respectively, by assigning each of these covariance terms to the error variance.

Parameter representation of the decomposition:

Using the model parameters, shown in Figure 4-10, the decomposition of the variance σ_Y^2 of *Y* can be represented by the model parameters:

$$\begin{aligned}\sigma_Y^2 = {} & \lambda_1^2 \cdot \phi_1^2 + \lambda_2^2 \cdot \phi_2^2 + \cdots + \lambda_p^2 \cdot \phi_p^2 + 2 \cdot \lambda_1 \cdot \lambda_2 \cdot \phi_{12} + \\ & 2 \cdot \lambda_1 \cdot \lambda_3 \cdot \phi_{13} + \cdots + 2 \cdot \lambda_{p-1} \cdot \lambda_p \cdot \phi_{p-1,p} + \theta_\varepsilon^2\end{aligned}$$

Computation of the true score variance and reliability using matrices:

A convenient way to compute the true score variance consists in the usage of matrices since the expression representing the true score variance:

$$\lambda_1^2 \cdot \phi_1^2 + \lambda_2^2 \cdot \phi_2^2 + \cdots + \lambda_p^2 \cdot \phi_p^2 +$$
$$2 \cdot \lambda_1 \cdot \lambda_2 \cdot \phi_{12} + 2 \cdot \lambda_1 \cdot \lambda_3 \cdot \phi_{13} + \cdots + 2 \cdot \lambda_{p-1} \cdot \lambda_p \cdot \phi_{p-1,p}$$

Can be computed by means of the matrix product:

$$\boldsymbol{\lambda}^{\mathbf{T}} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\lambda},$$

where:

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_p \end{bmatrix} \text{ and } \boldsymbol{\Phi} = \begin{bmatrix} \phi_1^2 & \phi_{12} & \cdots & \phi_{1p} \\ \phi_{21} & \phi_2^2 & \cdots & \phi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{p1} & \phi_{p2} & \cdots & \phi_p^2 \end{bmatrix}.$$

$\boldsymbol{\lambda}$ denotes the $p \times 1$ column vector of loading coefficients and $\boldsymbol{\Phi}$ the $p \times p$ covariance matrix of the latent constructs. The symbol $^{\mathbf{T}}$ represents the operation of transposing a matrix or a vector. In the present case the column vector $\boldsymbol{\lambda}$ is transformed to a row vector.

The variance of *Y* is thus given by:

$$\sigma_Y^2 = \boldsymbol{\lambda}^{\mathbf{T}} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\lambda} + \theta_\varepsilon^2. \qquad (4\text{-}19)$$

Consequently, the reliability ρ_{YY} of the test *Y* is given by:

$$\rho_{YY} = \frac{\boldsymbol{\lambda}^{\mathbf{T}} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\lambda}}{\boldsymbol{\lambda}^{\mathbf{T}} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\lambda} + \theta_\varepsilon^2}, \qquad (4\text{-}20)$$

Comment:

The reliability can also be computed by means of the equation:

$$\rho_{YY} = \frac{\sigma_Y^2 - \theta_\varepsilon^2}{\sigma_Y^2} \qquad (4\text{-}21)$$

This follows immediately from Equation (4-19).

The following example demonstrates the computation of the reliability of a test using the matrix method.

Ex. 4-8: Computation of the reliability of a test:

Given: The measurement model of Figure 4-11:

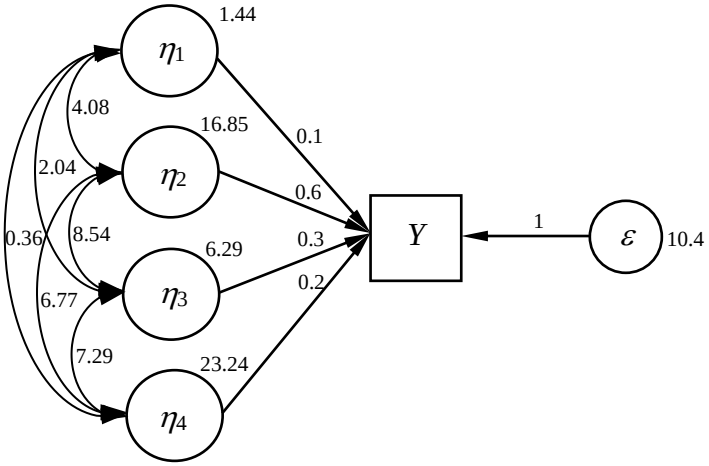


Figure 4-11: *Measurement model to demonstrate the computation of the reliability of a test Y.*

The vector of the loading coefficients $\boldsymbol{\lambda}$, the covariance matrix of the latent constructs $\boldsymbol{\Phi}$, and the error variance θ_ε^2:

$$\boldsymbol{\lambda} = \begin{bmatrix} 0.1 \\ 0.6 \\ 0.3 \\ 0.2 \end{bmatrix}, \quad \boldsymbol{\Phi} = \begin{bmatrix} 1.44 & 4.08 & 2.04 & 0.36 \\ 4.08 & 16.85 & 8.54 & 6.77 \\ 2.04 & 8.54 & 6.29 & 7.29 \\ 0.36 & 6.77 & 7.29 & 23.24 \end{bmatrix}, \text{ and } \theta_\varepsilon^2 = 10.4 .$$

Application of Equation 4-19 results in the variance of *Y* predicted by the model:

$$\begin{aligned} \sigma_Y^2 &= \boldsymbol{\lambda}^{\mathbf{T}} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\lambda} + \theta_\varepsilon^2 \\ &= \begin{bmatrix} 0.1 & 0.6 & 0.3 & 0.2 \end{bmatrix} \cdot \begin{bmatrix} 1.44 & 4.08 & 2.04 & 0.36 \\ 4.08 & 16.85 & 8.54 & 6.77 \\ 2.04 & 8.54 & 6.29 & 7.29 \\ 0.36 & 6.77 & 7.29 & 23.24 \end{bmatrix} \cdot \begin{bmatrix} 0.1 \\ 0.6 \\ 0.3 \\ 0.2 \end{bmatrix} + 10.4 \\ &= 13.777 + 10.4 \\ &= \underline{\underline{24.177}} \end{aligned}$$

Thus the true score variance is 13.777. The reliability of *Y* is given by:

$$\rho_{YY} = \frac{\boldsymbol{\lambda}^{\mathbf{T}} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\lambda}}{\boldsymbol{\lambda}^{\mathbf{T}} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\lambda} + \theta_\varepsilon^2} = \frac{13.777}{24.177} = \underline{\underline{0.570}} .$$

Comment on the usage of the model predicted variance of Y:

The reliability was computed using the variance of *Y* predicted by the model and not by means of the observed variance of *Y*, i.e., the variance computed from the sample. If the model is correct the model predicted variance is a more efficient estimator of the variance of *Y* and should thus be used.

If, by contrast, the model is a bad approximation of the population the estimated variances (and reliabilities) are, in general, biased and thus useless.

Let us terminate this presentation of the concept of *reliability* by asserting the following four observations.

1. The concept of r*eliability* refers to the test and not to the theoretical construct to be measured.
2. The concept of reliability is a population-based quantity, with *varying values of the latent construct*. Otherwise, the true score variance, and thus the reliability is zero. In this case, differences between observed test scores cannot be explained by means of differences of the underlying latent constructs.
3. The reliability of a test cannot be observed directly but can only be measured in a *model dependent* way. This process assumes that the assumptions incorporated in the model are (approximately) correct.
4. The variance of the error is given by the following equation (cf. Exercise 4-11): $\mathrm{Var}(\varepsilon) = \mathrm{Var}(Y) \cdot \left[1 - \mathrm{Rel}(Y)\right]$. The square root of this quantity is called the *standard error of measurement*.

4.4.1 Traditional Approaches to Measuring the Reliability of a Test

There exist three closely related approaches for determining the reliability of a test that do not require analysis of the structure of a test.

Method 4-3: *Traditional methods for measuring the reliability of a test:*

1. *Test-Retest Method:*
   The same test is applied at different time points to the same subjects. The correlation between the test scores from the two applications is a measure of the reliability of the test.
2. *Alternative Test Versions:*
   The test is available in two versions. The two versions are applied to different participants. The correlation between the test scores from the two versions is a measure of the reliability of the test (versions).

3. *Test Halves:*
   The whole test is split into two halves (e.g. even and uneven test items). The two test halves are applied to different groups. The reliability of the test conforms to the corrected correlation coefficient using the Spearman-Brown formula:

$$\rho_{YY'} = \frac{2 \cdot r}{1 + r},$$

   Where $\rho_{YY'}$ denotes the reliability of the test and *r* symbolizes the correlation of the test scores from different halves (For details on the Spearman-Brown formula, see Concept 4-15 on page 92).
   The Spearman-Brown formula is used since the whole test is double the length of the two halves.

The three methods differ from the one described by Method 4-2 (page 85) in two fundamental ways:

1. The reliabilities are computed using the sample correlations and not model predicted quantities.
2. Apparently, no model assumptions are required.

Consequently, the determination of the reliability of a test using one of the three methods seems to contradict the claim that the reliability of a test can be measured in a model dependent way only. However, the impression of a model free measurement of reliability by means of the traditional methods is incorrect since these methods are based on the assumption *that the repeated measures, test versions, and test halves, respectively, are parallel*. The correlation coefficient and the Spearman-Brown formula are unbiased estimators of the reliability only in case of parallel tests (Exercise 4-13).
Usually the assumption of parallel measures is not tested. Remember, that the parallel model predicts that the variances of the observed test scores are equal (cf. Section 4.2.3.4).
The following examples demonstrate various possibilities of deviations from the parallel test model:

Ex. 4-9: Possible reasons for deviations from the parallel test model:

Given: Three test models for modeling the structure of two tests (Figure 4-12):

Model (a):
This model represents a situation with test Y_1 exerting a direct influence on test Y_2. This may be the case with repeated application of the same test due to memory effects: The examinee remembers her previous response.

Model (b):
This model represents a situation where the covariance between test Y_1 and Y_2 cannot be explained completely by the measured construct since there is a second latent construct ξ exerting an influence on both measures. This may be the case if, for example, a specific answer to a test item has a high social desirability.

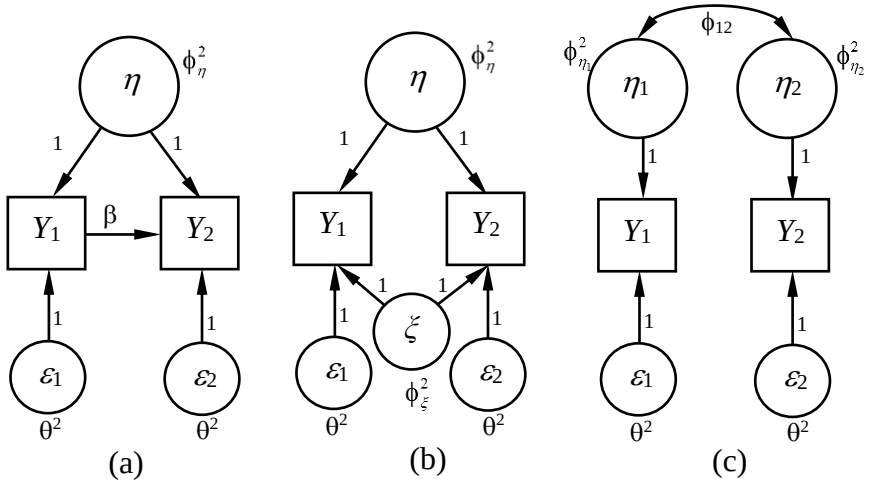


Figure 4-12: *Measurement model demonstrating possible deviations from parallel tests: (a) the first measurement exerts an influence on the second measurement, (b) the presence of a second latent construct ξ exerting an influence on both measures, and (c) the measurement concerns a transient state instead of a stable trait.*

Model (c):
This model represents a situation where transient states, like emotional states, are measured at different time points (cf. Exercise 4-7). Consequently, different latent constructs are measured at different time points.

In case of one of the models representing the measurement process correctly, the correlation coefficient is not an adequate estimate of the reliability of the two tests.
Up to now we have discussed methods that apply to single tests or test items. In the following section the reliability of the sum scores of test items will be discussed.

4.4.2 The Reliability of Sum Scores

A common practice consists in computing the sum Y of the scores of the single test items $Y_1, Y_2, \ldots, Y_n$: $Y = Y_1 + Y_2 + \cdots + Y_n$. This raises the question of the reliability $\mathrm{Rel}(Y)$ of the sum score.

In the following, we first discuss the most commonly used procedures and coefficients, respectively: *Spearman-Brown formula, coefficient* α (also called *Cronbach's* α), and *Guttman's* λ_2. Subsequently, the computation of the reliability based on the analysis of the covariance structure of the test items is presented.

4.4.2.1 TRADITIONAL MEASURES OF THE RELIABILITY OF SUM SCORES

The traditional measures of the reliability of the sum of test scores can be computed using the (co-) variances and correlations, respectively, of observed test scores. Consequently, their usage is quite common. However, one should keep in mind that these estimates of reliability are also based on model assumptions that limit their usage. It is thus important to understand these limitations and the resulting biases of the estimates in case of violations of the model assumptions.

Concept 4-15: *Spearman-Brown formula for increased test length: Reliability of the sum of m parallel tests*

Given:

m parallel tests: $Y_1, Y_2, \ldots, Y_n$. The reliability of a single item is ρ (since items are parallel the reliability is identical for all items).

The reliability $\mathrm{Rel}(Y)$ of the sum $Y = Y_1 + Y_2 + \cdots + Y_n$ of the *n* items, is given by:

$$\mathrm{Rel}(Y) = \frac{n \cdot \rho}{1 + (n-1) \cdot \rho}. \qquad (4\text{-}22)$$

Equation (4-21) is known as the *Spearman-Brown* formula.

Comment:

The Spearman-Brown formula assumes that the test items are parallel since only in this case the items measuring the latent construct have the same reliability (assuming the classical case with loading coefficients being all of equal size).

In case of violations of the assumption of parallel items the Spearman-Brown coefficient either overestimates or underestimates the reliability of the sum of the items. Which case of bias obtains depends on the covariance structure. The same principle applies in case of coefficient α (cf. Principle 4-2 on page 105).

The most commonly employed coefficient of reliability is coefficient α that is also called Cronbach's α.

Concept 4-16: *Coefficient α (Cronbach's α): The reliability of the sum of m τ-equivalent tests*

Given:

n τ-equivalent tests: $Y_1, Y_2, \ldots, Y_n$.

The reliability of the sum $Y = Y_1 + Y_2 + \cdots + Y_n$ is given by:

$$\alpha = \frac{n}{n-1} \cdot \frac{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathrm{Cov}\left(Y_i, Y_j\right)}{\mathrm{Var}\left(Y\right)} \tag{4-23}$$

or

$$\alpha = \frac{n}{n-1} \cdot \left(1 - \frac{\sum_{i=1}^{n} \mathrm{Var}\left(Y_i\right)}{\mathrm{Var}\left(Y\right)} \right) \tag{4-24}$$

Comment:

If the *n* tests are parallel coefficient α is identical to the Spearman-Brown coefficient (Exercise 4-18).

To obtain a better understanding of the equations underlying coefficient α consider the schematic representation of the covariance matrix of the test scores for the different tests (Figure 4-13). The matrix has been partitioned into three regions: The main diagonal contains the variances that are denoted by the letter *V*. The off-diagonal parts of the matrix, denoted by *C*, contain the covariances.

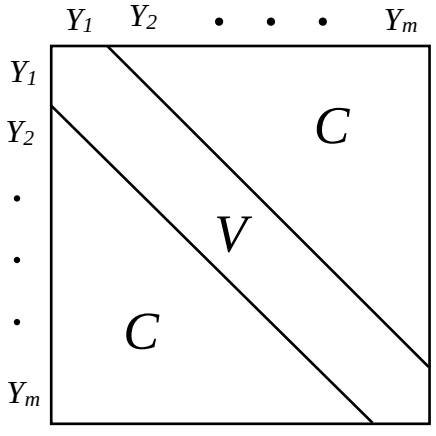


Figure 4-13: *Schematic representation of the covariance matrix of the variables $Y_1, Y_2, \ldots, Y_n$: C = covariances, V = variances.*

The nominator of coefficient α consists of the sum of the two regions denoted by *C*, and denominator contains the sum of all entries of the matrix. This quotient is multiplied by the factor $n/(n-1)$.

The equation of coefficient α can thus be written in schematic form:

$$\alpha = \frac{n}{n-1} \cdot \frac{C+C}{C+C+V}. \tag{4-25}$$

The letters C and *V* represent the sum of the entries in the respective regions of the covariance matrix.

The identity of the two different equation of coefficient α (cf. Equations 4-22 and 4-23) can be easily demonstrated: The variance $\mathrm{Var}(Y)$ consists of the sum of all entries of the covariance matrix (cf. *Basics of Covariance Algebra*). By subtraction of the sum of the variances (i.e. the sum of the entries in the main diagonal denoted by *V* in Figure 4-13) from $\mathrm{Var}(Y)$ one gets the sum of all covariances in the matrix. This leads directly to the expression of Equation 4-23, as the following derivation exhibits:

$$\begin{aligned}
\alpha &= \frac{n}{n-1} \cdot \frac{C+C}{C+C+V} \\
&= \frac{n}{n-1} \cdot \frac{C+C+V-V}{C+C+V} \\
&= \frac{n}{n-1} \cdot \left(\frac{C+C+V}{C+C+V} - \frac{V}{C+C+V} \right) \\
&= \frac{n}{n-1} \cdot \left(1 - \frac{V}{C+C+V} \right)
\end{aligned} \tag{4-26}$$

Comment 4-6: *Cronbach's α and Stigler's law of eponymy*

The naming of Cronbach's α (instead of coefficient α) conforms to Stephen Stigler's »law of eponymy« according to which scientific discoveries are usually not named after their discoverers.

In the case of coefficient α it was Louis Guttman (1916-1987), who developed this coefficient, and not Lee Cronbach (1916-2001) whose name it got.

Guttman (1945) proved that coefficient α provides as lower bound on the reliability of the sum of test scores, assuming uncorrelated errors (a concise prove of this fact can be found in Lord and Novick, 1968, page 88-89).

The name *coefficient* α is due to Cronbach (1951). He proved that α from $2 \cdot n$ tests is identical to the mean of the α values, where the mean is computed over the α values from all possible splits of the $2 \cdot n$ item into two halves with n items each. All in all there are $1/2 \cdot (2 \cdot n)! / (n!)^2$ different such splits (for a simple proof of this fact, cf. Lord and Novick, 1968).
Kuder and Richardson (1937) developed the formula of coefficient α for the special case of binary test items, i.e., items with two possible outcomes only.

Comment:
As one might expect, Stigler's law of eponymy has been proposed first by the sociologist Robert K. Merton (1910-2003), and not by Stephen Stigler (1941-) according to whom it is named.

A further coefficient is Guttman's λ_2. This coefficient provides, in case of uncorrelated errors, also a lower bound on the reliability that is at least as good as the one given by coefficient α:

Concept 4-17: *Guttman's λ_2: A lower bound of the reliability of the sum of m tests in case of uncorrelated errors (Guttman, 1945).*

Given:

n tests: $Y_1, Y_2, \ldots, Y_n$ with uncorrelated errors.

The reliability of the sum $Y = Y_1 + Y_2 + \cdots + Y_n$ is always greater or equal to Guttman's λ_2:

$$\lambda_2 = \frac{\sqrt{\frac{n}{n-1} \cdot \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left[\mathrm{Cov}(Y_i, Y_j) \right]^2} + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathrm{Cov}(Y_i, Y_j)}{\mathrm{Var}(Y)}. \quad (4\text{-}27)$$

Equivalently,

$$\lambda_2 = 1 - \frac{\sum_{i=1}^{n} \mathrm{Var}(Y_i)}{\mathrm{Var}(Y)} + \frac{\sqrt{\frac{n}{n-1} \cdot \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left[\mathrm{Cov}(Y_i, Y_j) \right]^2}}{\mathrm{Var}(Y)} \quad (4\text{-}28)$$

Comments:

1. If the *n* tests are τ-equivalent coefficient α and Guttman's λ_2 are identical: $\lambda_2 = \alpha$ (cf. Exercise 4-17 and Ex. 4-10).

2. Despite the fact that Guttman's λ_2 can be computed using popular statistical software, like SPSS, Guttman's λ_2 is used less often than the more common coefficient α.

   This is somewhat surprising in the light of the fact that, in case of uncorrelated errors, Guttman's λ_2 provides a better lower bound than coefficient α.

3. Coefficient α und Guttman's λ_2 are based on the following inequalities:

   The inequality underlying coefficient α:

$$\sum_{i=1}^{n} \mathrm{Var}(\eta_i) \geq \frac{1}{n-1} \cdot \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathrm{Cov}(\eta_i, \eta_j) . \qquad (4\text{-}29)$$

   The inequality underlying Guttman's λ_2:

$$\sum_{i=1}^{m} \mathrm{Var}(\eta_i) \geq \sqrt{\frac{n}{n-1} \cdot \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left[\mathrm{Cov}(\eta_i, \eta_j) \right]^2} . \qquad (4\text{-}30)$$

(for further details, cf. Lord and Novick, 1968)

Ex. 4-10: Computation of coefficient α and Guttman's λ_2 for congeneric and τ-equivalent tests:

Given: The covariance matrix of 5 congeneric tests (Tab. 4-3):

Tab. 4-3: *The covariance matrix of five congeneric tests.*

	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. 4-4: *The covariance matrix of five τ-equivalent tests.*

	Y_1	Y_2	Y_3	Y_4	Y_5
Y_1	3.62	1.43	1.43	1.43	1.43
Y_2	1.43	2.50	1.43	1.43	1.43
Y_3	1.43	1.43	2.02	1.43	1.43
Y_4	1.43	1.43	1.43	6.84	1.43
Y_5	1.43	1.43	1.43	1.43	3.24

Coefficient α is given by:

$$\alpha = \frac{n}{n-1} \cdot \frac{C+C}{C+C+V}$$
$$= \frac{5}{4} \cdot \frac{2 \cdot (1.8+0.36+4.32+1.08+0.2+2.4+0.6+0.48+0.12+1.44)}{2 \cdot (1.8+0.36+4.32+1.08+0.2+2.4+0.6+0.48+0.12+1.44) + 3.62+2.5+2.02+6.84+3.24}$$
$$= \underline{\underline{0.730}}$$

Guttman's λ_2 is given by:

$$\lambda_2 = 1 - \frac{\sum_{i=1}^{n} \mathrm{Var}(Y_i)}{\mathrm{Var}(Y)} + \frac{\sqrt{\frac{n}{n-1} \cdot \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \left[\mathrm{Cov}(Y_i, Y_j) \right]^2}}{\mathrm{Var}(Y)}$$
$$= 1 - \frac{(3.62+2.5+2.02+6.84+3.24)}{2 \cdot (1.8+0.36+4.32+1.08+0.2+2.4+0.6+0.48+0.12+1.44) + 3.62+2.5+2.02+6.84+3.24} + \frac{\sqrt{\frac{5}{4} \cdot 2 \cdot (1.8^2+0.36^2+4.32^2+1.08^2+0.2^2+2.4^2+0.6^2+0.48^2+0.12^2+1.44^2)}}{2 \cdot (1.8+0.36+4.32+1.08+0.2+2.4+0.6+0.48+0.12+1.44) + 3.62+2.5+2.02+6.84+3.24}$$
$$= \underline{\underline{0.787}}$$

Given: The covariance matrix of five τ-equivalent tests (Tab. 4-4):

Computation of α and λ_2 results in the same value:

$$\alpha = \frac{n}{n-1} \cdot \frac{C+C}{C+C+V}$$
$$= \frac{5}{4} \cdot \frac{1.43 \cdot 5 \cdot 4}{1.43 \cdot 5 \cdot 4 + 3.62+2.5+2.02+6.84+3.24}$$
$$= \frac{1.43 \cdot 25}{1.43 \cdot 20 + 3.62+2.5+2.02+6.84+3.24}$$
$$= \underline{\underline{0.764}}$$

$$\lambda_2 = \frac{\sqrt{\frac{n}{n-1}\cdot\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\left[\mathrm{Cov}\left(Y_i,Y_j\right)\right]^2}+\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\mathrm{Cov}\left(Y_i,Y_j\right)}{\mathrm{Var}\left(Y\right)}$$

$$= \frac{\sqrt{\frac{5}{4}\cdot 1.43^2\cdot 20}+1.43\cdot 20}{1.43\cdot 20+3.62+2.5+2.02+6.84+3.24}$$

$$= \frac{\sqrt{1.43^2\cdot 25}+1.43\cdot 20}{1.43\cdot 20+3.62+2.5+2.02+6.84+3.24}$$

$$= \frac{1.43\cdot 25}{1.43\cdot 20+3.62+2.5+2.02+6.84+3.24}$$

$$= \underline{\underline{0.764}}$$

We next tackle the problem of estimating the reliability of the sum of tests by using the methods of modern psychometrics.

4.4.2.2 COMPUTATION OF THE RELIABILITY OF WEIGHTED SUM SCORES IN THE CONTEXT OF COVARIANCE STRUCTURE ANALYSIS

The analysis of the covariance structure of the test items enables one to compute an unbiased estimate of the reliability of the weighted sum of the test items provided that the structural equation model represents the structure of the test items correctly.

Method 4-4: *Computation of the reliability of a weighted sum of test scores using the estimated parameters of the confirmatory factor analytic model*

Given:

The general linear psychometric model of first order with *p* latent constructs $\eta_1,\eta_2,\ldots,\eta_p$ and *n* test items $Y_1,Y_2,\ldots,Y_n$. Figure 4-14 depicts the causal diagram of the linear structural equation model (or confirmatory factor analytic model).

Wanted:

An estimator of the reliability of the weighted sum of the *n* test scores:

$$Y = w_1\cdot Y_1 + w_2\cdot Y_2 + \cdots + w_n\cdot Y_n$$

The analysis of the covariance structure provides us with the following two matrices:

1. The model implied ($n \times n$) covariance matrix of the observed scores:

$$\hat{\boldsymbol{\Sigma}} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & \cdots & Y_n \end{array} \\ \begin{bmatrix} \hat{\sigma}^2_{Y_1} & \hat{\sigma}_{Y_1Y_2} & \cdots & \hat{\sigma}_{Y_1Y_n} \\ \hat{\sigma}_{Y_2Y_1} & \hat{\sigma}^2_{Y_2} & \cdots & \hat{\sigma}_{Y_2Y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{Y_nY_1} & \hat{\sigma}_{Y_nY_2} & \cdots & \hat{\sigma}^2_{Y_n} \end{bmatrix} \end{array}$$

2. The ($n \times n$) estimated covariance matrix of errors:

$$\hat{\boldsymbol{\Theta}} = \begin{array}{c} \\ \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{array} \begin{array}{c} \begin{array}{cccc} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{array} \\ \begin{bmatrix} \hat{\theta}^2_1 & \hat{\theta}_{12} & \cdots & \hat{\theta}_{1n} \\ \hat{\theta}_{21} & \hat{\theta}^2_2 & \cdots & \hat{\theta}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\theta}_{n1} & \hat{\theta}_{n2} & \cdots & \hat{\theta}^2_n \end{bmatrix} \end{array}$$

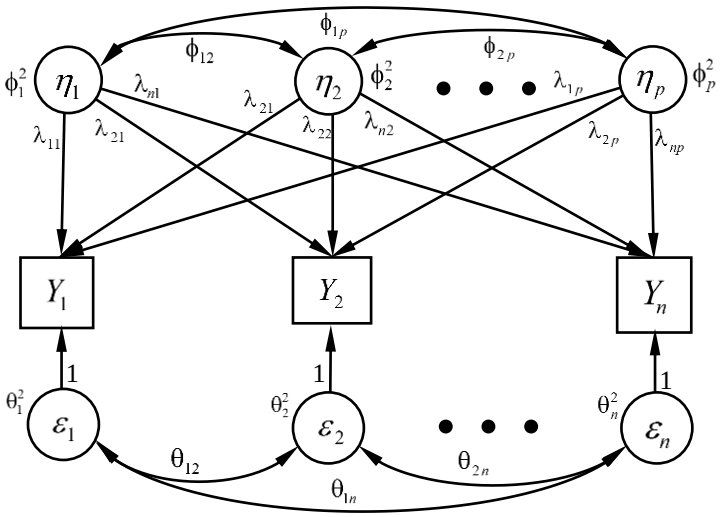


Figure 4-14: *Confirmatory factor analytic model of first order.*

Notation 4-8: Hat symbol

The hat symbol ^ indicates that the covariance matrices (and all its entries) are model based estimates that result from the structural analysis of the sample data.

The computation of the reliability comprises the following three steps:

1. Compute the true score variance (cf. Concept 4-13 on page 83) of the weighted sum scores by means of the following matrix product:

$$\hat{\mathrm{Var}}_{\boldsymbol{\eta},\mathbf{w}}(Y) = \mathbf{w}^{\mathbf{T}} \cdot \left(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Theta}}\right) \cdot \mathbf{w}. \qquad (4\text{-}31)$$

The symbols have the following meaning:

$\hat{\mathrm{Var}}_{\boldsymbol{\eta},\mathbf{w}}(Y)$ represents the estimated true score variance of the weighted sum of the observed scores.

$\mathbf{w}^{\mathbf{T}} = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}$ denotes a row vector with the weights of the weighted sum: $Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + \cdots + w_n \cdot Y_n$.

The symbol « $^{\mathbf{T}}$ » represents the operation of transposing a matrix or a vector: exchanging the rows and columns of the matrix and vector, respectively.

2. Compute the model predicted variance of the weighted sum of the observed scores: $Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + \cdots + w_n \cdot Y_n$.

$$\hat{\mathrm{Var}}_{\mathbf{w}}(Y) = \mathbf{w}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}} \cdot \mathbf{w}. \qquad (4\text{-}32)$$

3. The (estimated) reliability $\hat{\mathrm{Rel}}_{\mathbf{w}}(Y)$ of the weighted sum is given by dividing the (estimated) true score variance by model predicted variance of the weighted sum scores:

$$\hat{\mathrm{Rel}}_{\mathbf{w}}(Y) = \frac{\hat{\mathrm{Var}}_{\boldsymbol{\eta},\mathbf{w}}(Y)}{\hat{\mathrm{Var}}_{\mathbf{w}}(Y)}. \qquad (4\text{-}33)$$

Comment: The true score variance $\hat{\mathrm{Var}}_{\boldsymbol{\eta},\mathbf{w}}(Y)$ can be computed in an alternative way using the covariance matrix of the latent constructs and the matrix of loading coefficients.

The analysis of the covariance structure provides us with the following two matrices:

1. The estimated ($p \times p$) covariance matrix of the latent constructs:

$$\hat{\boldsymbol{\Phi}} = \begin{array}{c} \\ \eta_1 \\ \eta_2 \\ \vdots \\ \eta_p \end{array} \begin{array}{c} \begin{array}{cccc} \eta_1 & \eta_2 & \cdots & \eta_p \end{array} \\ \begin{bmatrix} \hat{\phi}_1^2 & \hat{\phi}_{12} & \cdots & \hat{\phi}_{1p} \\ \hat{\phi}_{21} & \hat{\phi}_2^2 & \cdots & \hat{\phi}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\phi}_{p1} & \hat{\phi}_{p2} & \cdots & \hat{\phi}_p^2 \end{bmatrix} \end{array}, \text{ and}$$

2. Die ($n \times p$) matrix of loading coefficients:

$$\hat{\mathbf{\Lambda}} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{array} \begin{array}{c} \begin{array}{cccc} \eta_1 & \eta_2 & \cdots & \eta_p \end{array} \\ \begin{bmatrix} \hat{\lambda}_{11} & \hat{\lambda}_{12} & \cdots & \hat{\lambda}_{1p} \\ \hat{\lambda}_{21} & \hat{\lambda}_{22} & \cdots & \hat{\lambda}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\lambda}_{n1} & \hat{\lambda}_{n2} & \cdots & \hat{\lambda}_{np} \end{bmatrix} \end{array}.$$

The estimated true score variance $\hat{\text{Var}}_{\boldsymbol{\eta},\mathbf{w}}(Y)$ of the weighted sum $Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + \cdots + w_n \cdot Y_n$ can be computed by means of the matrix product:

$$\hat{\text{Var}}_{\boldsymbol{\eta},\mathbf{w}}(Y) = \mathbf{w}^{\mathbf{T}} \cdot \hat{\mathbf{\Lambda}} \cdot \hat{\mathbf{\Phi}} \cdot \hat{\mathbf{\Lambda}}^{\mathbf{T}} \cdot \mathbf{w} \,. \qquad (4\text{-}34)$$

The following examples illustrate the method.

Ex. 4-11: Computation of the reliability of the sum of test scores of five tests for a general factor analytic model:

Given: The model shown in Figure 4-15 for modeling the covariance structure of 5 tests Y_1-Y_5.

Wanted: The reliability of the sum $Y = Y_1 + Y_2 + Y_3 + Y_4 + Y_5$ of the five tests.

The single matrices resulting from the model look like this (cf. Figure 4-15):

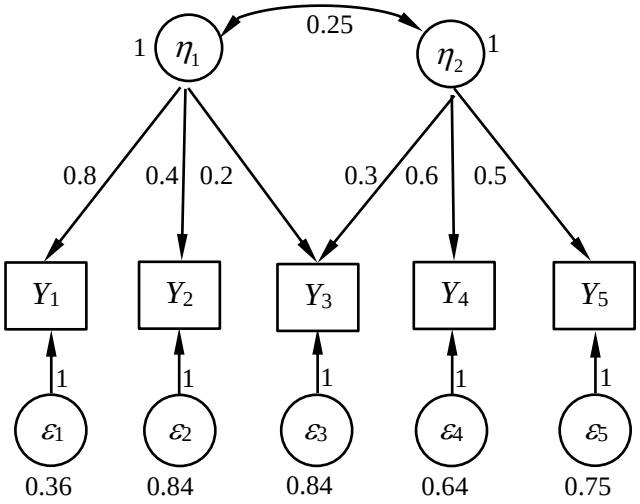


Figure 4-15: *Structural equation model of five tests.*

1. The model implied covariance matrix of observed scores:

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

2. The estimated covariance matrix of the errors:

$$\hat{\boldsymbol{\Theta}} = \begin{array}{c} \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{array} \begin{array}{c} \begin{array}{ccccc} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 & \varepsilon_5 \end{array} \\ \begin{bmatrix} 0.36 & 0 & 0 & 0 & 0 \\ 0 & 0.84 & 0 & 0 & 0 \\ 0 & 0 & 0.84 & 0 & 0 \\ 0 & 0 & 0 & 0.64 & 0 \\ 0 & 0 & 0 & 0 & 0.75 \end{bmatrix} \end{array}$$

3. The estimated covariance matrix of the latent constructs:

$$\hat{\boldsymbol{\Phi}} = \begin{array}{c} \\ \eta_1 \\ \eta_2 \end{array} \begin{array}{c} \begin{array}{cc} \eta_1 & \eta_2 \end{array} \\ \begin{bmatrix} 1.00 & 0.25 \\ 0.25 & 1.00 \end{bmatrix} \end{array}$$

4. The estimated matrix of loading coefficients:

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

The weight vector looks like this:

$$\mathbf{w} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \ \mathbf{w^T} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The true score variance of the sum $Y = Y_1 + Y_2 + Y_3 + Y_4 + Y_5$ is given by:

$$\hat{\mathrm{Var}}_{\boldsymbol{\eta},\mathbf{w}}(Y) = \mathbf{w}^{\mathbf{T}} \cdot \left[\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Theta}}\right] \cdot \mathbf{w}$$

$$= \begin{bmatrix}1 & 1 & 1 & 1 & 1\end{bmatrix} \cdot \left(\begin{bmatrix} 1.000 & 0.320 & 0.220 & 0.120 & 0.100 \\ 0.320 & 1.000 & 0.110 & 0.060 & 0.050 \\ 0.220 & 0.110 & 1.000 & 0.210 & 0.175 \\ 0.120 & 0.060 & 0.210 & 1.000 & 0.300 \\ 0.100 & 0.050 & 0.175 & 0.300 & 1.000 \end{bmatrix} - \begin{bmatrix} 0.36 & 0 & 0 & 0 & 0 \\ 0 & 0.84 & 0 & 0 & 0 \\ 0 & 0 & 0.84 & 0 & 0 \\ 0 & 0 & 0 & 0.64 & 0 \\ 0 & 0 & 0 & 0 & 0.75 \end{bmatrix} \right) \cdot \begin{bmatrix}1 \\ 1 \\ 1 \\ 1 \\ 1\end{bmatrix}$$

$$= \underline{\underline{4.9}}$$

Using the alternative method for computing the true score variance leads to the same result:

$$\hat{\mathrm{Var}}_{\boldsymbol{\eta},\mathbf{w}}(Y) = \mathbf{w}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Lambda}} \cdot \hat{\boldsymbol{\Phi}} \cdot \hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \mathbf{w}$$

$$= \begin{bmatrix}1 & 1 & 1 & 1 & 1\end{bmatrix} \cdot \begin{bmatrix} 0.8 & 0.0 \\ 0.4 & 0.0 \\ 0.2 & 0.3 \\ 0.0 & 0.6 \\ 0.0 & 0.5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0.8 & 0.4 & 0.2 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.3 & 0.6 & 0.5 \end{bmatrix} \cdot \begin{bmatrix}1 \\ 1 \\ 1 \\ 1 \\ 1\end{bmatrix}$$

$$= \underline{\underline{4.9}}$$

The model predicted variance is:

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

$$= \begin{bmatrix}1 & 1 & 1 & 1 & 1\end{bmatrix} \cdot \begin{bmatrix} 1.000 & 0.320 & 0.220 & 0.120 & 0.100 \\ 0.320 & 1.000 & 0.110 & 0.060 & 0.050 \\ 0.220 & 0.110 & 1.000 & 0.210 & 0.175 \\ 0.120 & 0.060 & 0.210 & 1.000 & 0.300 \\ 0.100 & 0.050 & 0.175 & 0.300 & 1.000 \end{bmatrix} \cdot \begin{bmatrix}1 \\ 1 \\ 1 \\ 1 \\ 1\end{bmatrix}$$

$$= \underline{\underline{8.33}}$$

The estimated reliability of Y is therefore:

$$\hat{\mathrm{Rel}}(Y) = \frac{\hat{\mathrm{Var}}_{\boldsymbol{\eta},\mathbf{w}}(Y)}{\hat{\mathrm{Var}}_{\mathbf{w}}(Y)} = \frac{4.9}{8.33} = \underline{\underline{0.588}}$$

Comment:

By comparison, $\alpha = 0.500$ and $\lambda_2 = 0.513$. Thus both traditional coefficients underestimate the true reliability.

The following example illustrates the effect of a differential weighting of test items on the reliability of the weighted sum.

Ex. 4-12: Reliability of the weighted sum of test scores in the general factor analytic model (continuation of Ex. 4-11):

Given:

- The model of Figure 4-15 on page 101;
- The weight vector for weighting the five tests:
  $\mathbf{w}^{\mathbf{T}} = \begin{bmatrix}0.5 & 0.2 & 0.1 & 0.15 & 0.05\end{bmatrix}$.

Wanted: The reliability of the weighted sum of the 5 tests.

$$Y = 0.5 \cdot Y_1 + 0.2 \cdot Y_2 + 0.1 \cdot Y_3 + 0.15 \cdot Y_4 + 0.05 \cdot Y_5$$

The estimated true score variance of *Y* is given by:

$$\hat{\mathrm{Var}}_{\boldsymbol{\eta},\mathbf{w}}(Y) = \mathbf{w}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Lambda}} \cdot \hat{\boldsymbol{\Phi}} \cdot \hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \mathbf{w}$$

$$= \begin{bmatrix} 0.5 & 0.2 & .01 & 0.15 & 0.05 \end{bmatrix} \cdot \begin{bmatrix} 0.8 & 0.0 \\ 0.4 & 0.0 \\ 0.2 & 0.3 \\ 0.0 & 0.6 \\ 0.0 & 0.5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0.8 & 0.4 & 0.2 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.3 & 0.6 & 0.5 \end{bmatrix} \cdot \begin{bmatrix} 0.50 \\ 0.20 \\ 0.10 \\ 0.15 \\ 0.05 \end{bmatrix}$$

$$= \underline{\underline{0.307}}$$

The estimated variance of *Y* is given by:

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

$$= \begin{bmatrix} 0.5 & 0.2 & 0.1 & 0.15 & 0.05 \end{bmatrix} \cdot \begin{bmatrix} 1.000 & 0.320 & 0.220 & 0.120 & 0.100 \\ 0.320 & 1.000 & 0.110 & 0.060 & 0.050 \\ 0.220 & 0.110 & 1.000 & 0.210 & 0.175 \\ 0.120 & 0.060 & 0.210 & 1.000 & 0.300 \\ 0.100 & 0.050 & 0.175 & 0.300 & 1.000 \end{bmatrix} \cdot \begin{bmatrix} 0.50 \\ 0.20 \\ 0.10 \\ 0.15 \\ 0.05 \end{bmatrix}$$

$$= \underline{\underline{0.456}}$$

Consequently, the estimated reliability of *Y* is given by:

$$\hat{\mathrm{Rel}}(Y) = \frac{\hat{\mathrm{Var}}_{\boldsymbol{\eta},\mathbf{w}}(Y)}{\hat{\mathrm{Var}}_{\mathbf{w}}(Y)} = \frac{0.307}{0.456} = \underline{\underline{0.675}}$$

Comment:
Obviously, the differential weighting of tests results in a significant higher reliability (0.675), compared to the unweighted sum (0.588). The determination of the optimal weights that maximize the reliability will be treated in Section 4.4.3.

As noted above, coefficient α is one of the most frequently used coefficients of reliability. In the following, some problems associated with the use of coefficient α and Guttman's λ_2 are discussed.

4.4.2.3 CRITICAL ISSUES CONCERNING COEFFICIENT α AND GUTTMAN'S λ_2

Problems associated with the use of coefficient α and of Guttman's λ_2 are concerned with their interpretation.

4.4.2.3.1 Over- and underestimation of the reliability by coefficient α and Guttman's λ_2

The previous examples illustrated that coefficient α and Guttman's λ_2 can underestimate the true reliability of the sum of tests (cf. Ex. 4-11).

However, both coefficient α and Guttman's λ_2 can also overestimate the true reliability.

Principle 4-2: *Overestimation and underestimation of the reliability by coefficient α and Guttman's λ_2*

Coefficient α and Guttman's λ_2 can overestimate or underestimate the true reliability of the sum of the test items. Which case of bias obtains depends on the covariance structure:

1. The true reliability is underestimated in the following situations:
   (i) The tests are congeneric.
   (ii) The tests are loading on more than a single construct and the errors are uncorrelated.
2. In case of correlated errors both coefficients can lead to an overestimation of the true reliability.
   The overestimation is due to the fact that with correlated errors the covariance between the observed scores cannot be attributed entirely to the latent variables as assumed by α and λ_2.

Ex. 4-13: Overestimation of the reliability by α and λ_2:

Given: The covariance matrix of 5 tests (Tab. 4-5):

The covariance matrix in Tab. 4-5 corresponds to the implied covariance matrix of observed scores of the model in Figure 4-16 (Notice the covariance arcs between the error variables).

Coefficient α and Guttman's λ_2 computed from the covariance matrix in Tab. 4-5 have the following values: $\alpha = .697$ and $\lambda_2 = .705$.

Tab. 4-5: *Covariance matrix of test scores for five test items.*

	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

The true reliability of the sum Y is however $\mathrm{Rel}(Y) = .510$. Thus the two coefficients overestimate the correct reliability considerably.

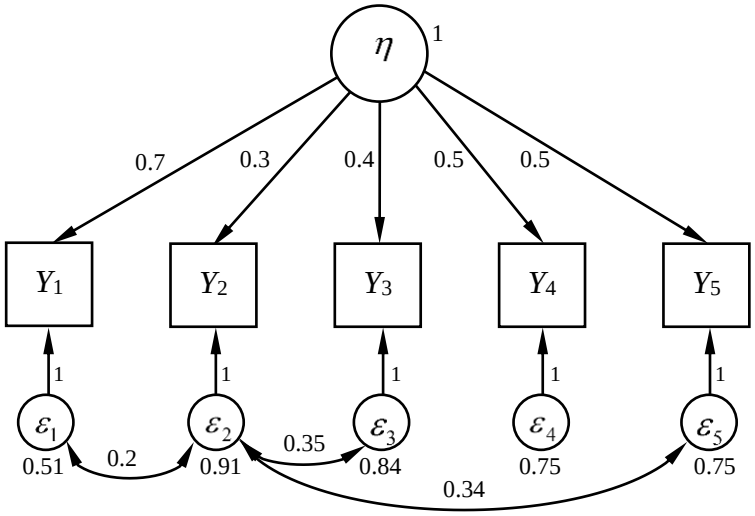


Figure 4-16: *Model underlying the data of Tab. 4-5.*

Comment 4-7*: Over- and underestimation of the true reliability by means of coefficient α and Guttman's λ_2 in practical applications*

I suspect that, in general, published values of coefficient α *overestimate* the true reliability of the sum of the test scores. This supposition is based on the following consideration:

Sums of test scores for which coefficient α is computed comprise usually at least 10-15 test items. It is quite improbable that the covariance of these items can be explained by the existence of a single common latent construct exerting an influence on the tests. Rather, there remains a residual covariance between test scores that is not explained by the latent construct (as shown in Figure 4-16).

In conclusion, the question of whether coefficient α and Guttman's λ_2 overestimate or underestimate the true reliability depends on the covariance structure of the test scores. Thus an analysis of the covariance structure is required. This, again, exhibits the problems of traditional psychometric methods, i.e., of employing coefficients based on observed measures without further consideration of the assumptions underlying these coefficients.

4.4.2.3.2 A possible erroneous interpretation of coefficient α and Guttman's λ_2 as measures of homogeneity

Coefficient α as well as Guttman's λ_2 must not be interpreted as coefficients of homogeneity, that is, the assumption that the tests are based on a single underlying construct. The following example illustrates that test score based on many underlying constructs are compatible with high valued of coefficient α and Guttman's λ_2.

Ex. 4-14: Coefficient α, Guttman's λ_2, and the homogeneity of tests (Green, Lissitz & Mulaik, 1977).

Given: The model shown in Figure 4-17.

The model assumes that the test scores of the 10 test items are influenced by 5 uncorrelated latent traits with each trait affecting 4 tests. In addition, each test is affected by two latent variables only (e.g. variable Y_4 is affected by the latent constructs η_1 and η_5 only). Consequently, the 10 test items have no single underlying trait and are thus not homogenous.

Tab. 4-6 exhibits the model implied covariance matrix. Coefficient α and Guttman's λ_2, based on this covariance matrix, are relatively high: $\alpha = .829$ and $\lambda_2 = .848$.

Tab. 4-6: *The implied covariance matrix of the model of Figure 4-17.*

	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

Coefficient α and Guttman's λ_2 are thus not sensible measures of the homogeneity of the test items. The latter can be assessed by fitting the congeneric model to the test items. If the fit of the model is reasonable, homogeneity of the items may be assumed.

This example demonstrates again the problem of drawing conclusion using coefficients that are based on the observed test scores only, without analyzing the structure of the tests.

The present example exhibits, however, another problem of the two reliability coefficients: $\alpha = .829$, as well as $\lambda_2 = .848$, are lower than the reliability of the single test items which is .98. The true reliability of the sum of the 10 test items is however .995 and thus higher than the reliability of a single test item. Thus, α and λ_2 not only underestimate the reliability of the sum of the test items but also the reliability of the single items.

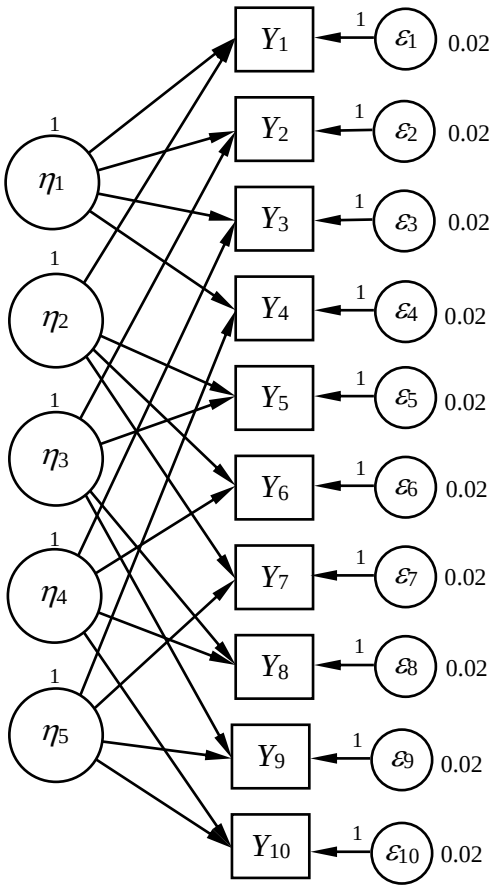


Figure 4-17: *A model demonstrating the inappropriateness of coefficient α and Guttman's λ_2 as measures of homogeneity. The latent variables are uncorrelated and each observed variable is influenced by two latent variables. The loading coefficients are all equal to $\lambda = 0.7$.*

4.4.2.4 PROBLEMS ASSOCIATED WITH THE RELIABILITY OF UNWEIGHTED SUMS OF TEST ITEMS

Coefficient α and Guttman's λ_2 both presume the unweighted sum of test items. In the following, it will be demonstrated that the usage of unweighted sums may be associated with a number of problems.

Comment 4-8*: Weights for the computation of weighted sum scores (lectures from elementary statistics):*

Principles of elementary statistics dictate that in pooling scores from different populations the latter should be weighted, where the *precisions* (i.e. the inverse variances) should be used as weights.

In case of summing single sample values from different populations the values have to be weighted by the precisions $1/\sigma_i^2$. The symbol σ_i^2 denotes the variance of population *i*, from which the score was drawn.
In case of pooling sample means the inverse squared standard errors of the means n_i/σ_i^2 are used. The symbol n_i denotes the sample size of the sample from which the mean was computed.
In case of the population variances being unknown sample estimates of the variances are used.

As shown in Ex. 4-12 (page 103) an unweighted sum of test items can result in a considerably lower reliability than the weighted sum. In addition, unweighted sums can violate requirements concerning the monotony of the reliability of sums of items.

Principle 4-3: *Requirements concerning the monotony of the reliability of sums:*

1. The addition of reliable test items to an existing set of items should result in an increase of the reliability of the sum scores.
2. Replacing, within a set of test items, an item with one of higher reliability should lead to an increased reliability of the sum of the test items.
3. Reducing the correlation between two constructs of two tests should result in a decrease of the reliability of the sum of the tests.

Li, Rosenthal, and Rubin (1996) demonstrated the violation of each of these requirements in case of using unweighted sums of test items.

Comment 4-9*:* *Violation of Principle 4-3 in case of using coefficient α and Guttman's λ_2:*

It was shown above in Ex. 4-14 (page 107) that the usage of coefficient α and Guttman's λ_2 resulted in an estimate of the reliability of the sum of the 10 test items that was lower than the reliability of a single item.
However, this result may be due to the fact that the two coefficients underestimate the true reliability and not to the fact that they are based on simple sums. However, the following examples, illustrate that using simple sums of tests items may be problematic per se.

Ex. 4-15: The reliability of the sum of reliable test items does not increase with test length.

Given: The τ-equivalent model of Figure 4-18.

The set of test items comprises two groups of items:

- Items Y_1 and Y_2 are of high reliability: $\mathrm{Rel}(Y_i) = .6$.
- The residual items $Z_1,\ldots,Z_5$ are of low reliability only: $\mathrm{Rel}(Y_i) = .1$.

Using the Spearman-Brown formula we get the reliability of the sum $Y = Y_1 + Y_2$ of the two items (note that Y_1 and Y_2 are parallel):

$$\mathrm{Rel}(Y) = \frac{n \cdot \rho}{1 + (n-1) \cdot \rho} = \frac{2 \cdot .6}{1 + (2-1) \cdot .6} = .75 .$$

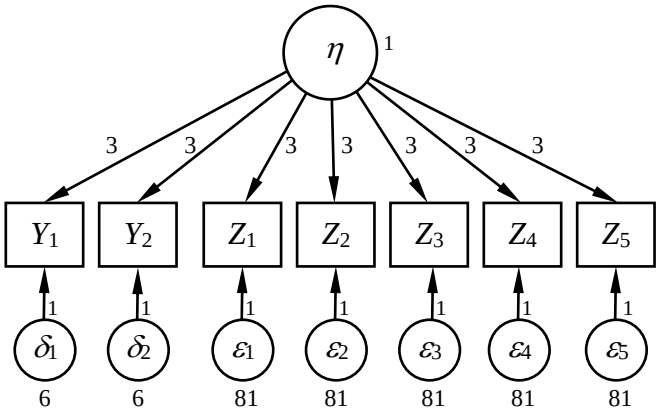


Figure 4-18: *A model demonstrating that adding reliable test items to a set of already existing ones can result in a decrease of the reliability of the sum of the test items.*

Using coefficient α for computing the sum of the 7 test items $Z = Y_1 + Y_2 + Z_1 + Z_2 + Z_3 + Z_4 + Z_5$ results in a lower reliability (Note that in the present case coefficient α is an unbiased estimate of the reliability of the sum since the 7 test items are τ-equivalent): $\alpha = \mathrm{Rel}(Z) = .514$.

The addition of 5 reliable items $Z_1, Z_2, \ldots, Z_5$ to the items Y_1 and Y_2 results in a decrease of reliability of the sum. This illustrates the violation of the first requirement of Principle 4-3 concerning the monotony of the reliability of sums.

The following example demonstrates the violation of the second requirement of Principle 4-3.

Ex. 4-16: The reliability of the sum of items does not increase if an item is replaced by one with a higher reliability.

Given: The two models shown in Figure 4-19.

The model on the left-hand side comprises two test items with reliabilities .6 and .1. For the reliability of the sum we get: $\alpha = \text{Rel}(Y+Z) = .29$.

In the model on the right-hand side the item with reliability .6 has been replaced by an item of higher reliability (.65), resulting in a lower reliability of the sum: $\text{Rel}(Y+Z) = .23$, $\alpha = .22$.

The reduction of the reliability of the sum for the model on the right-hand side is due to the fact that the new item exhibits a lower variance. By consequence, the item gets a lower weight with respect to the other items thus resulting in a decrease of the reliability of the sum.

Note the reliability of the sum of the two items is considerably lower than the reliability of item *Y*. This, again, demonstrates a violation of the first requirement of Principle 4-3.

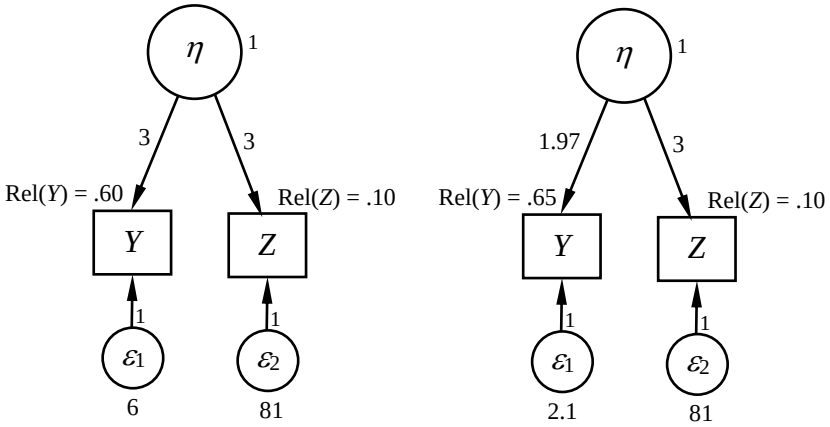


Figure 4-19: *Two models demonstrating that the replacement of an item by a more reliable one can result in a decrease of the reliability of the sum.*

The next example demonstrates the violation of the third monotony requirement of Principle 4-3 (as well as of the first one).

Ex. 4-17: The reliability of the sum of items need not increase with the correlation between the latent variables.

Given: The two models of Figure 4-20.

The model on the left-hand side corresponds to that of Figure 4-19. Thus the reliability of the sum is: $\alpha = \text{Rel}(Y+Z) = .29$.

In the model on the right-hand side the correlation between the latent constructs was reduced to .95. Importantly, the variance of the test item with reliability .60 is higher than the variance of corresponding item in the left figure. The reliability of the sum of the two items for the model on the right-hand side is: $\text{Rel}(Y+Z) = .34$, (Coefficient $\alpha = .32$).

Increasing the variance of the item with the higher reliability increases its contribution to the reliability of the sum. As a consequence, the reduction of the reliability of the sum due to the decrease of the correlation between the latent constructs is more than compensated.

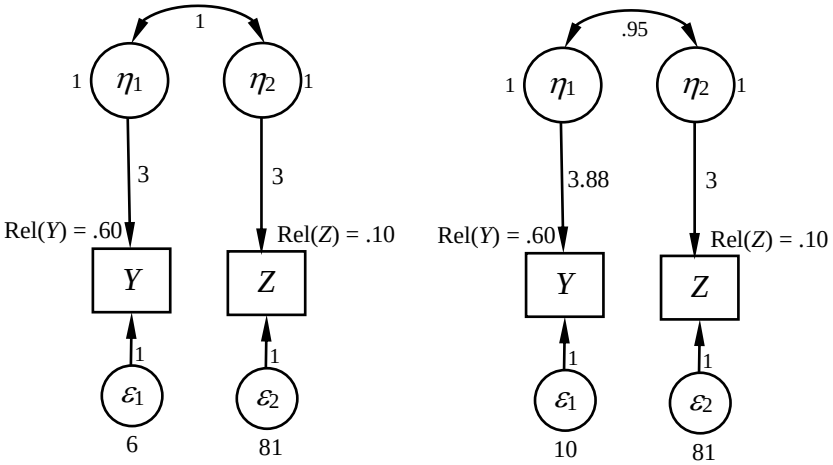


Figure 4-20: *Two models demonstrating that a reduction of the correlation between two latent constructs, keeping the reliabilities of the single test items constant, can result in an increase of the reliability of the sum.*

The three previous examples illustrate strikingly the problems associated with usage of unweighted sum scores. This raises two questions.

Questions:

1. *Is there a measure of reliability that does not violate the three monotony properties of Principle 4-3?*
2. *How have the single test items to be weighted to get the optimal (maximal) reliability?*

It turns out that an optimal weighting of test items exists resulting in the maximal reliability that conforms to the monotony property of Principle 4-3. The computation of the maximal reliability is based on an analysis of the covariance structure of the test items.

4.4.3 Maximal Reliability and the Optimal Weighting of Tests

The maximal reliability of the weighted sum of test scores is obtained by finding those weights that maximize the reliability. Assume that an analysis of the covariance structure has been performed that resulted in estimates of covariance matrix of the latent constructs $\hat{\mathbf{\Phi}}$, the matrix of the estimated loading coefficients $\hat{\mathbf{\Lambda}}$, and the covariance matrix of the errors $\hat{\mathbf{\Theta}}$. One can estimate the reliability of the weighted sum using weight vector $\mathbf{w}$ by means of the matrix formula described above in Method 4-4 (on page 98):

$$\hat{\mathrm{Rel}}_{\mathbf{w}}(Y) = \frac{\hat{\mathrm{Var}}_{\boldsymbol{\eta}.\mathbf{w}}(Y)}{\hat{\mathrm{Var}}_{\mathbf{w}}(Y)} = \frac{\mathbf{w}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Lambda}} \cdot \hat{\boldsymbol{\Phi}} \cdot \hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \mathbf{w}}{\mathbf{w}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}} \cdot \mathbf{w}}, \qquad (4\text{-}35)$$

where $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Lambda}} \cdot \hat{\boldsymbol{\Phi}} \cdot \hat{\boldsymbol{\Lambda}}^{\mathbf{T}} + \hat{\boldsymbol{\Theta}}$ is the model implied covariance matrix. In order to find the maximal reliability one has to look for that weight vector that maximizes the expression in Equation 4-13. Note however that a weight vector $\mathbf{w}_0$ that maximizes the reliability is not unique since each multiple of $\mathbf{w}_0$ (say $c \cdot \mathbf{w}_0$, with c being an arbitrary constant) will also be a weight vector the maximizes Equation 4-34. This is the case because the constant c appears in the nominator and denominator of Equation 4-34 and thus cancels.

In this situation, one usually chooses the weight vector of length 1: $\|\mathbf{w}_0\| = 1$, where $\|\mathbf{w}_0\| = \sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}$ is the square root of the sum of the squared components of $\mathbf{w}_0$.

The maximal reliability as well as the weight vector that maximizes the reliability can be found by means of matrix methods that will be described below (cf. Method 4-6 on page 115). In case of the congeneric, τ-equivalent or parallel model the maximal reliability as well as the optimal weights can be represented by means of simple algebraic expressions. We thus treat this special case before tackling the more complex general case.

Method 4-5: *Computation of the maximal reliability of the optimally weighted sum of congeneric, τ-equivalent, and parallel test items and the associated optimal weights.*

Given: The congeneric, τ-equivalent or parallel test model of *n* test items: $Y_1, Y_2, \ldots, Y_n$ (cf. Concept 4-3, Concept 4-6, and Concept 4-7).

1. The *maximal reliability* of the optimally weighted sum of the test items is given by:

$$\mathrm{Rel}_{\max}(Y) = \frac{\frac{\nu_1^2}{1-\nu_1^2} + \frac{\nu_2^2}{1-\nu_2^2} + \cdots + \frac{\nu_n^2}{1-\nu_n^2}}{1 + \frac{\nu_1^2}{1-\nu_1^2} + \frac{\nu_2^2}{1-\nu_2^2} + \cdots + \frac{\nu_n^2}{1-\nu_n^2}} \qquad (4\text{-}36)$$

The symbols $\nu_i \ (i = 1, 2, \ldots, n)$ denote the *standardized* loading coefficients.

The standardized weights can be computed from the unstandardized ones as follows:

$$\nu_i = \frac{\phi_j}{\sigma_{Y_i}} \cdot \lambda_i \qquad (i = 1, 2, \ldots, n). \qquad (4\text{-}37)$$

σ_{Y_i} denotes the standard error of test Y_i, and ϕ_j denotes standard error of the latent construct on which Y_i is loading with loading coefficients λ_i.

2. The *optimal* weights $w_1, w_2, \ldots, w_n$ for computing the weighted sum $Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + \cdots + w_n \cdot Y_n$ of maximal reliability are given by:

$$w_i = \frac{\lambda_i}{\theta_i^2} \qquad (i = 1, 2, \ldots, n). \qquad (4\text{-}38)$$

λ_i denotes the *unstandardized* loading coefficient of test Y_i and θ_i^2 represents the associated error variance.

Ex. 4-18: Estimation of the maximal reliability of the weighted sum of congeneric tests

Given: The covariance matrix of 5 congeneric tests (Tab. 4-3 on page 96).

Wanted: The maximal reliability of the optimally weighted sum of the tests with the associated optimal weights.

Tab. 4-7 depicts the relevant quantities involved in the computation of the maximal reliability and of the optimal weights.

Column 2 contains the unstandardized loading coefficients.

Column 3 contains the standardized loading coefficients.

Column 4 contains the terms that enter the computation of the maximal reliability (cf. Equation 4-35).

Column 5 contains the error variances required for computing the optimal weights.

Column 6 contains the non-normalized optimal weights.

Column 7 contains the normalized optimal weights.

Tab. 4-7: *Data for computing the maximal reliability and the optimal weights.*

Test	λ_i	ν_i	$\frac{\nu_i^2}{1-\nu_i^2}$	θ_i^2	w_i	w_i^{normal}
Y_1	1.8	0.946	8.526	0.38	4.737	0.897
Y_2	1.0	0.632	0.667	1.50	0.667	0.126
Y_3	0.2	0.141	0.020	1.98	0.101	0.019
Y_4	2.4	0.918	5.333	1.08	2.222	0.421
Y_5	0.6	0.333	0.125	2.88	0.208	0.039

The maximal reliability is given by:

$$\begin{aligned}\mathrm{Rel}_{\max}(Y) &= \frac{\frac{\nu_1^2}{1-\lambda_1^2}+\frac{\nu_2^2}{1-\lambda_2^2}+\cdots+\frac{\nu_n^2}{1-\nu_n^2}}{1+\frac{\nu_1^2}{1-\nu_1^2}+\frac{\nu_2^2}{1-\nu_2^2}+\cdots+\frac{\nu_n^2}{1-\nu_n^2}} \\ &= \frac{8.526+0.667+0.020+5.333+0.125}{1+(8.526+0.667+0.020+5.333+0.125)} \\ &= \underline{\underline{.936}}\end{aligned}$$

The optimal (unstandardized) weight for test Y_1 is given by:

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

The other weights can be computed in a similar way.

The maximal reliability (.936) is distinctly higher that the reliability of the simple sum of the test items (.821) as well as of the reliabilities due to coefficient α (.730) and Guttman's λ_2 (.787).

In case of test items loading on multiple latent constructs, the simple computational formulas of Method 4-5 for computing the maximal reliability and the optimal weights are no longer valid. In this case, the maximal reliability and the optimal weights can either be determined by means of optimization or using matrix methods.

Comment 4-10:

The presentation of the procedures in Method 4-6 assumes that the reader has some specialized knowledge about matrices. This material is not required for understanding the stuff in subsequent chapters and may thus be skipped without loss of continuity.

Method 4-6: *Determining the maximal reliability and the optimal weights of the weighted sum of test scores in the general linear latent trait model.*

Given: The general linear latent trait model of Figure 4-14 (on page 99).

Wanted: The maximal reliability of the optimally weighted sum of the tests with the associate optimal weights.

Method I: Direct maximization of the reliability:

The method consists in maximizing the reliability given by the equation:

$$\hat{\mathrm{Rel}}_{\mathbf{w}}(Y) = \frac{\mathbf{w}^{\mathbf{T}}\cdot\hat{\boldsymbol{\Lambda}}\cdot\hat{\boldsymbol{\Phi}}\cdot\hat{\boldsymbol{\Lambda}}^{\mathbf{T}}\cdot\mathbf{w}}{\mathbf{w}^{\mathbf{T}}\cdot\hat{\boldsymbol{\Sigma}}\cdot\mathbf{w}}$$

with respect to the weights of the weight vector **w**.

Comment:
This method is inferior to the two methods presented below. However, it enables the determination of the maximal reliability and the associated weights with programs, like Excel, that do not dispose of the matrix functions required for the other two methods (Excel has an optimizer, called the *Solver*, that can be used for performing the optimization).

Method II (Greene & Carmines, 1980):

1. The maximal reliability conforms to the greatest eigenvalue of the following matrix:

$$\hat{\mathbf{\Sigma}}^{-1/2} \cdot \hat{\mathbf{\Lambda}} \cdot \hat{\mathbf{\Phi}} \cdot \hat{\mathbf{\Lambda}}^{\mathbf{T}} \cdot \hat{\mathbf{\Sigma}}^{-1/2}. \qquad (4\text{-}39)$$

The symbols have the following meaning:

$\hat{\mathbf{\Phi}}$ denotes the estimated covariance matrix of the latent constructs,

$\hat{\mathbf{\Lambda}}$ denotes the estimated matrix of loading coefficients,

$\hat{\mathbf{\Sigma}}$ denotes the model implied covariance matrix.

Comment:
The exact form of these matrices is shown in Method 4-4 on page 98f.

$\hat{\mathbf{\Sigma}}^{-1/2}$ denotes a matrix that may be conceptualized as the inverse square root of the matrix $\hat{\mathbf{\Sigma}}$ in the following sense:

$$\hat{\mathbf{\Sigma}}^{-1/2} \cdot \hat{\mathbf{\Sigma}} \cdot \hat{\mathbf{\Sigma}}^{-1/2} = \mathbf{I},$$

where **I** is the identity matrix, i.e. a diagonal matrix with only 1 in the main diagonal (all other entries are zero).

The matrix $\hat{\mathbf{\Sigma}}^{-1/2}$ can be obtained by means of a singular value decomposition (SVD) of the covariance matrix $\hat{\mathbf{\Sigma}}$.

The function `svd` of the program R performs the SVD of a matrix. Similar functions exist for other software packages.

The SVD provides the matrix factorization:

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

The symbols have the following meaning:

$\mathbf{V}$ denotes the matrix of the orthogonal eigenvectors of $\hat{\mathbf{\Sigma}}$.

$\mathbf{\Delta}$ denotes a diagonal matrix with the singular values (i.e. the eigenvalues of $\hat{\mathbf{\Sigma}}$) $\delta_i^2 \quad (i = 1, 2, \ldots, n)$ on the main diagonal in decreasing order. In case of $\hat{\mathbf{\Sigma}}$ being a proper covariance matrix the singular values are all greater than zero:

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

Using the matrices resulting from the SVD the matrix $\hat{\mathbf{\Sigma}}^{-1/2}$ is computed by the equation:

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

where,

$$\mathbf{\Delta}^{-1/2} = \begin{bmatrix} \frac{1}{\delta_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\delta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\delta_n} \end{bmatrix}.$$

Thus, one generates a diagonal matrix with the inverse of the square roots of the singular values on the main diagonal (note that in case of a proper covariance matrix the singular values are greater than zero and the inverses of their square roots are thus finite).

2. The optimal weights associated with the maximal reliability are given by:

$$\mathbf{w}_0 = \hat{\mathbf{\Sigma}}^{-1/2} \cdot \mathbf{u}_0. \qquad (4\text{-}40)$$

The vector $\mathbf{u}_0$ is the eigenvector associated with the greatest eigenvalue.

Background information on the method:
The method is based on the following theorem (see, for example, Schott, 2005):
Let $\mathbf{\Sigma}$ denote a ($n \times n$) covariance matrix. The *Rayleigh quotient* is defined as:

$$\rho = \frac{\mathbf{w}^{\mathbf{T}} \cdot \mathbf{\Sigma} \cdot \mathbf{w}}{\mathbf{w}^{\mathbf{T}} \cdot \mathbf{w}},$$

where **w** denotes an arbitrary ($n \times 1$) vector $\left(\neq \mathbf{0}\right)$.

The theorem states:

1. The maximal value of ρ corresponds to the maximal eigenvalue of $\boldsymbol{\Sigma}$.
2. The weight vector $\mathbf{w}_0$ that maximizes ρ is the eigenvector that is associated with the maximal eigenvalue.

In the actual case the following expression has to be maximized with respect to **w**:

$$\hat{\mathrm{Rel}}_{\mathrm{w}}(Y) = \frac{\mathbf{w}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Lambda}} \cdot \hat{\boldsymbol{\Phi}} \cdot \hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \mathbf{w}}{\mathbf{w}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}} \cdot \mathbf{w}}$$

Unfortunately the right-hand side of the equation is not a Rayleigh quotient (because of the denominator).
We thus define **w** in terms of the vector **u** as follows:

$$\mathbf{w} = \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \mathbf{u} \qquad (4\text{-}41)$$

Consequently, inserting the right-hand side of 4-37 into the equation of the reliability we get:

$$\begin{aligned} \hat{\mathrm{Rel}}_{\mathrm{w}}(Y) &= \frac{\mathbf{w}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Lambda}} \cdot \hat{\boldsymbol{\Phi}} \cdot \hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \mathbf{w}}{\mathbf{w}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}} \cdot \mathbf{w}} \\ &= \frac{\mathbf{u}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \hat{\boldsymbol{\Lambda}} \cdot \hat{\boldsymbol{\Phi}} \cdot \hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \mathbf{u}}{\mathbf{u}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \hat{\boldsymbol{\Sigma}} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \mathbf{u}} \\ &= \frac{\mathbf{u}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \hat{\boldsymbol{\Lambda}} \cdot \hat{\boldsymbol{\Phi}} \cdot \hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \mathbf{u}}{\mathbf{u}^{\mathbf{T}} \cdot \mathbf{u}} \end{aligned} \qquad (4\text{-}42)$$

The transition to the last line is based on the identity (already presented above): $\hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \hat{\boldsymbol{\Sigma}} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2} = \mathbf{I}$.
The expression in the last line is a Rayleigh quotient. Thus, the maximal eigenvalue of the matrix $\hat{\boldsymbol{\Sigma}}^{-1/2} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\Lambda}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}}^{-1/2}$ corresponds to the maximal reliability (according to the theorem presented above).
Let $\mathbf{u}_0$ denote the eigenvector associated with the maximal eigenvalue, then we get the optimal weight vector $\mathbf{w}_0$ by means of the transformation (cf. Equation 4-40):

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

Method III (Li, 1997):
1. Determine the greatest eigenvalue $\rho_{\max}$ of the matrix:

$$\hat{\boldsymbol{\Theta}}^{-1/2} \cdot \hat{\boldsymbol{\Lambda}} \cdot \hat{\boldsymbol{\Phi}} \cdot \hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Theta}}^{-1/2} . \qquad (4\text{-}43)$$

$\hat{\boldsymbol{\Theta}}$ denotes the estimated covariance matrix of the errors (the other symbols have the same meaning as previously).
2. The (estimated) maximal reliability is given by:

$$\hat{\text{Rel}}_{\max} = \frac{1}{1+1/\rho_{\max}}. \qquad (4\text{-}44)$$

3. The optimal weights are given by:

$$\mathbf{w}_0 = \hat{\boldsymbol{\Theta}}^{-1/2} \cdot \mathbf{u}_0, \qquad (4\text{-}45)$$

where $\mathbf{u}_0$ is the eigenvector associated with the maximal eigenvalue $\rho_{\max}$.

Comment:
At first sight Method III seems more involved than Method II since it requires the additional step (4-43) for computing the maximal reliability.
However the method is computationally less involved in case of the covariance matrix of errors $\hat{\boldsymbol{\Theta}}$ being a diagonal matrix, i.e. there are no correlated errors. In this case, $\hat{\boldsymbol{\Theta}}^{-1/2}$ is simply a diagonal matrix with the inverse standard errors of the error variables on the main diagonal. This spares the complicated matrix factorization (for computing $\hat{\boldsymbol{\Sigma}}^{-1/2}$) required with Method II.

Ex. 4-19: Maximal reliability in the general test model:

Given: The model of Figure 4-15 (on page 101).

The maximal reliability of the weighted sum of the 5 tests is .687 (by comparison, the reliability of the unweighted sum is .588).
The optimal weights (normalized) are:
w_1 = 0.934, w_2 = 0.200, w_3 = 0.174, w_4 = 0.194, and w_5 = 0.138.
Each of the three methods, described above, leads to the same result. (Exercise 4-26).

It will be shown next that the maximal reliability meets the monotony requirements of Principle 4-3 (page 109).

Ex. 4-20: The maximal reliability meets the monotony requirements of Principle 4-3:

1. Concerning Ex. 4-15 (page 109), the maximal reliability of the sum $Z = Y_1 + Y_2 + Z_1 + Z_2 + Z_3 + Z_4 + Z_5$ is given by:

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

This value is higher than the (maximal) reliability .75 of the sum of the first two items alone.

Comment:
Since the first two items are parallel they receive the same weight. Consequently, the maximal reliability equals the reliability of the (unweighted) sum of the items.

2. Concerning Ex. 4-16 (page 110), the maximal reliability of the model on the left-hand side of Figure 4-19 (page 111) [the model with the item of lower reliability] is given by:

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

For the model on the right-hand side of Figure 4-19 [the model with the item of higher reliability] the maximal reliability is:

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

Thus the combination of test items with a test item of higher reliability results in a higher maximal reliability.

3. Concerning Ex. 4-17 (page 111), the model with perfectly correlated latent constructs has a higher maximal reliability ($\text{Rel}_{\max} = .617$) than the model with an imperfect correlation of .95 between the latent constructs: $\text{Rel}_{\max} = .616$.

In conclusion, the maximal reliability with optimally weighted sum of test items is, in every respect, superior to the reliability of the unweighted sum of the items. However, the computation of the maximal reliability presupposes an analysis of the structure of the test items.

This ends or detailed treatment of the concept of reliability. We, now, turn to the second important concept for assessing the quality of a test.

4.5 Validity: Concept and Estimation

> *The problem of validity is that of whether a test really measures what it purports to measure,* […]
> Kelley (1927, page 14)

4.5.1 Introduction

The concept of *validity* is the second important construct of CTT for assessing the quality of tests. Despite the intuitively appealing characterization given by Kelley (1927) [cf. the citation above], there has been (and still remains) a great deal of confusion and misunderstanding that surrounds the concept of validity. This was in part due to a paper of Cronbach and Meehl (1955) that can be seen as a milestone in the attempt to clarify the concept of validity. However, their concept

of *construct validity* was not well understood by the scientific community (Kane, 2001). Moreover, the concept of construct validity has also become under criticism by proponents of an empiristic view of science that questioned the utility of theoretical constructs in psychology (cf. Bechtoldt, 1959).

An additional source of confusion is the proliferation of different modifiers associated with the concept of validity, to name just a few: *construct, incremental, predictive, convergent, discriminant, criterion-related, concurrent, criterion, factorial, construct-related, structural, content,* and *consequential*. (cf. Newton & Shaw, 2012: Table 1 on page 306). Because of this inflation of validity constructs Newton and Shaw (2012) propose to replace the whole notion of validity by another one: *quality*.

The present approach is based on the distinction between *empirical* and *theoretical validity* that was proposed by Lord and Novick (1968, Chapter 12). According to this distinction *incremental, predictive, criterion, criterion-related,* and *concurrent validity* belong to the category of empirical validity, whereas, *convergent, discriminant, factorial, construct-related, structural,* and *content validity* may be conceived of as members of the class of theoretical validity (The concept of *consequential validity* is not regarded as a useful validity concept for CTT, at all).

The differentiation between empirical and theoretical validity is based on the observation that the former, contrary to the latter, does not require the consideration of theoretical constructs. Thus, the distinction reflects the one already observed in the discussion of reliability: Estimates of validity based on observed scores versus theory (or model) dependent estimates.

The distinction between empirical and theoretical validity also entails that the statement of Kelly (1927), cited above, refers to different aspects in different contexts: In case of empirical validity it refers to the relationship between a test score and an empirical criterion (measure). By contrast, in case of theoretical validity the statement refers to the relationship between a test score and a theoretical construct the test intends to measure.

Let us now take a closer look at the two basic conceptions of validity starting with the concept of empirical validity.

4.5.2 Empirical Validity: The Classical Conception of Validity

Empirical validity can be seen as the classical conception of validity. This is evidenced by the fact that Lord and Novick's (1968) classic book on test theory devotes nearly the whole chapter on validity (Chapter 12 of their book) to empirical validity with only a short discussion of theoretical validity at the end of the chapter.

Empirical validity is concerned with the relationship between observed test scores and a criterion without taking into account the latent constructs and relationships. Consequently, the computation of validity coefficients does not require model assumptions and theoretical considerations. Rather, validity coefficients are based on correlation and regression methods, respectively, involving observed scores. In fact, the different types of empirical validity coefficients can be based on regression methodology, with a criterion or standard of comparison being regressed on observed test scores. The different labels and modifiers are but different names of essentially the same thing, specifically:

- *Criterion* and *criterion-related validity*, respectively, is concerned with the correlation between a test score Y and a criterion C. The associated correlation coefficient $\mathrm{Corr}(Y,C)$ has been termed *validity coefficient*. Equivalently, the standardized regression coefficient that results from regressing the criterion C on the test score Y can be computed.
- *Predictive validity* refers to how well specific tests scores are able to predict future performance (= the criterion). Thus, in this case, the criterion will be measured in the future. The multiple correlation coefficient between the criterion and the different test used as predictors (R or its square R^2) is used as the measure of validity.
- By contrast, in case of *concurrent validity*, test scores and criterion are measured at the same time. Again, R or R^2 can be used as a coefficient of concurrent validity.
- Finally, *incremental validity* refers to the increase of predicted variance in the criterion by adding a specific test to a set of predictors. The difference between R^2 with the test included and R^2 without the test $\left(R^2_{\text{with test}} - R^2_{\text{without test}}\right)$ can be used as a measure of incremental validity.

Measures of empirical validity are useful in situations where the objective does not consists in the measurement of latent constructs. Typical instances are admission tests for screening candidates applying for specific academic studies, like law or medicine. The main goal of this type of tests consists in selecting the candidates with the highest probability of success during the academic study.
If a test is intended to measure latent constructs empirical measures are problematic as the following example demonstrates.

Ex. 4-21: Validity coefficients and the measurement of latent constructs:

Given: The model of Figure 4-21.

Variable η denotes the target construct to be measured (e.g. η = social intelligence) and ξ represents another construct exerting an influence on the test scores *Y* as well as on the criterion *C*. (e. g. ξ = verbal abilities).

The correlation between *Y* and *C* is: $\mathrm{Corr}(Y,C) = 2/3$ (which can be shown by means of covariance algebra).

However, only half of the observed covariance between *Y* and *C* is due to the fact that both *C* and *Y* are measures of the latent target construct η (the residual covariance is due to the influence of ξ).

Even in case of the target construct η exerting an influence on neither *Y* nor *C* the observed correlation between them was ½.

On the other hand, controlling for ξ (if this were possible) reduces the correlation between *Y* and *C* to ½.

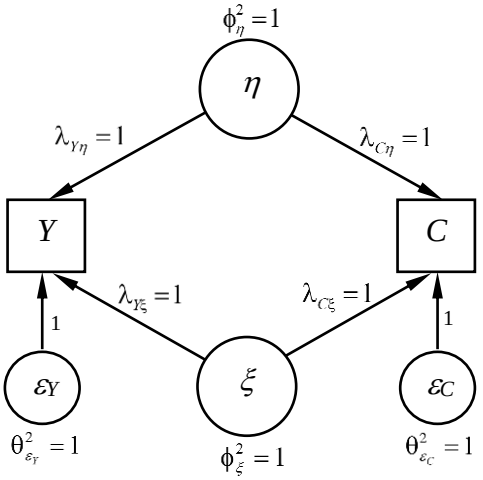


Figure 4-21: *Measurement model including the test Y and the criterion C. η denotes the target construct to be measured by Y and C. ξ represents another latent construct that exerts an influence on both measures.*

This example demonstrates two problems of the empirical validity estimates in the context of the measurement of latent constructs.

1. The empirical validity coefficient does not provide direct information about whether *Y* and *C* are gauging the same target construct η.
2. The empirical validity coefficient does not give direct information about the relationship between the target construct η and the measure *Y* (In the present case this correlation is $1/\sqrt{3}$).

In conclusion, empirical validity coefficients do not provide direct information about the structure of latent constructs and their relationships to the observed measures. This limits their utility for drawing conclusion with respect to latent variable models. This requires the consideration of the concept of theoretical validity.

4.5.3 Theoretical Validity and Latent Variable Models

The definition of Kelley (1927) given at the beginning of this section (that a test is valid if it measures what it purports to measure) makes, in case of the measurement of theoretical construct, an assertion concerning the relationship between the theoretical construct and the measure.

The idea of Kelley has been made more concrete by Borsboom, Mellenbergh, and Van Heerden (2004).

Concept 4-18: *Validity of a Test (Indicator, Measurement)*

(Bollen, 1989; Borsboom, et al., 2004):

A test is *valid*, if the systematic variation of the test scores is due to variations of the target construct the test intends to measure.

Thus, the *validity of a test* corresponds to the *direct structural* (*causal*) relationship between the latent construct and the test).

Comments:

1. The true score variance is a measure of the systematic variation of the test scores (mentioned above).
2. The validity of a test is one aspect of the more general concept of *construct validity* discussed below.

The following example illustrates the basic idea.

Ex. 4-22: Validity of a test

Given:

1. The target construct η: Emotional intelligence;
2. A test Y for measuring η;
3. Another construct ξ: Social competence.

Figure 4-22 depicts two possible measurement models representing the relationships between the variables.

In model (a), on the left-hand side, the test is affected by the target construct only. Consequently, the complete systematic variation in Y is due to variation of the target construct η.

By contrast, for model (b), on the right-hand side, the test scores are affected by both constructs. This limits the validity of the test as a measure of η.

Note that the reliability of Y may be higher in model (b) than in (a). This is due to the fact that the reliability is concerned with the total systematic variation. No distinction is made between the various sources that contribute to the total systematic variation (or true score variance).

The definition of a valid test (of Concept 4-18) refers to the relationship between the latent target construct and the test that is used to measure this construct.

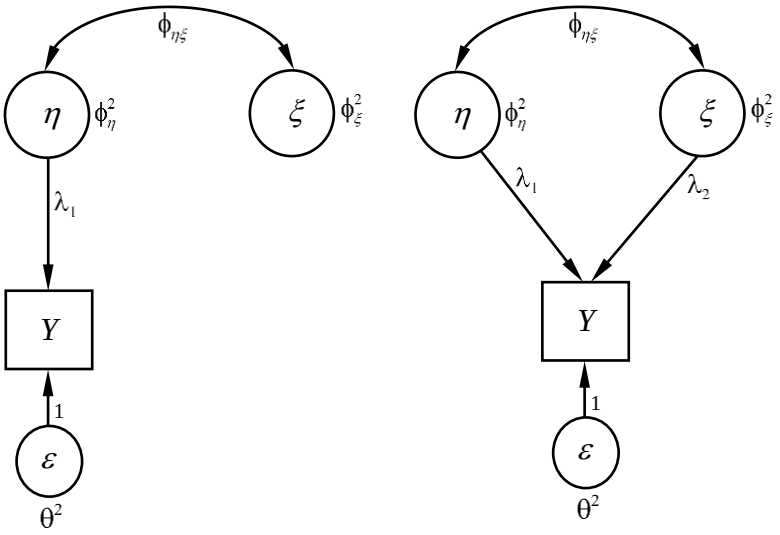


Figure 4-22: *Two measurement models: According to model (a) Y is a valid test of the target construct η; Concerning model (b), Y is not or only partially valid.*

The concept of *construct validity* of Cronbach und Meehl (1955) extends the notion of the validity of a test. According to this concepttion not only the relationship between the latent construct and the measure should be taken into account but all relevant relationships between the entities of a measurement model. Specifically, the following relationships have to be specified correctly in case of construct validity being present:

1. The relations between latent constructs;
2. The relations between latent constructs and measures;
3. The relations between the measures.

The concept of construct validity amounts to the correct specification of a measurement model.

Concept 4-19: *Construct Validity*
(Cronbach and Meehl, 1955):
Construct validity is present, if the measurement model that represents the measurement situation is (approximately) correct (cf. Concept 4-12 on page 81). In this case, *conclusions* about the test, drawn on the basis of the measurement model, are valid. In addition, estimates of various quantities (e.g. estimates of reliability and other relevant parameters) are unbiased.

Comment:
Note that this also excludes the presence of systematic biases due to influences that are not represented in the measurement model since in this case the model would not be correct.

The following example exhibits the close relationship between violations of construct validity and erroneous measurement models.

Ex. 4-23: Construct validity and erroneous measurement models

Given: A model representing the personality construct of *neuroticism* (Figure 4-23).

According to the model the construct of neuroticism is made up of various facets (cf., John & Soto, 2007):
The facets *anxiousness* and *depression* require no further explanation. The facet *hostility* refers, on the one hand, to the own behavior, and, on the other hand, to the interpretation of the behavior of other people (i.e. a tendency to interpret other peoples' behavior as hostile). The facet *impulsivity* refers to a propensity to act precipitously and in a reckless way, for example in stressful situations.

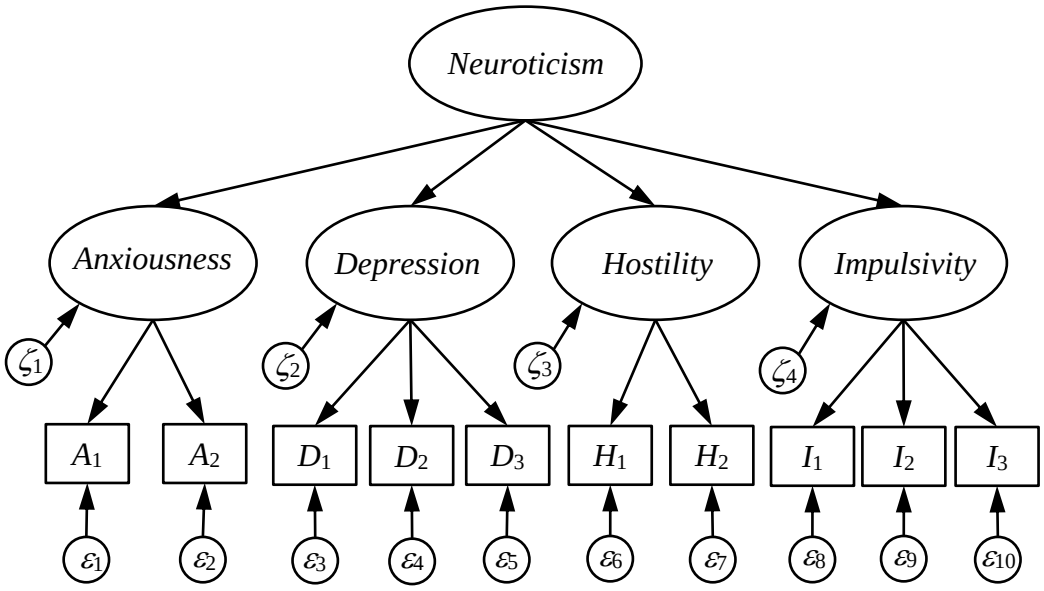


Figure 4-23: *Facets of the personality construct of neuroticism.*

The lowest level represents different tests, denoted by the letters *A*, *D*, *H*, and *I*, for measuring the single facets.

Assume that a researcher uses a single test only for measuring each facet since she assumes that there exists only a single construct thus ignoring the sub-constructs. She assumes the model shown in Figure 4-24. Obviously this simpler model does not reflect correctly the factorial structure of the constructs. This represents a case of missing construct validity.

The concept of construct validity has been criticized recently on being unrealistic, i.e., it does not reflect the reality of scientific psychology (Borsboom et al. 2004; Kane, 2001). Specifically, the concept stresses too much the relationship between theoretical constructs at the cost of the relationship between constructs and measures. However, scientific psychology does not dispose of a dense nomological network of constructs. In addition, the dictum of Kelley (1927) refers to the relationship between latent constructs and measurements and not between constructs.

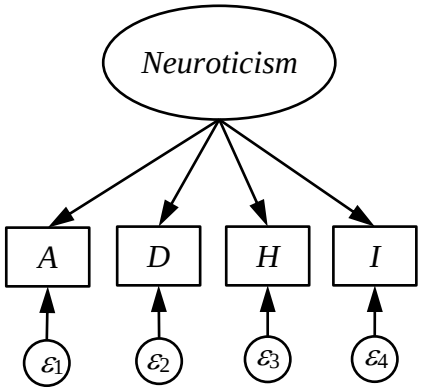


Figure 4-24: *Alternative model of the construct of neuroticism.*

This criticism is not fully justified since, as already noted above, the nomological network of Cronbach and Meehl (1955) also comprises the relationship between constructs and measures (cf. Cronbach & Meehl, 1955, page 290). Moreover, the assessment of the validity of a test has to take into account also constructs other than the target constructs (cf. Ex. 4-22, page 124) as well as their relationships to the target construct (cf. Section 4.5.4.2).

Let us conclude this section about theoretical validity by taking a short look on the other concepts related to theoretical validity, mentioned above: *factorial, construct-related, structural, content validity, convergent,* and *discriminant.* Obviously, the first three terms refer to different aspects of the measurement model and, consequently, to different aspects of construct validity.

Content validity refers to whether a test covers all facets of a theoretical construct. This amounts to whether the measurement model comprising the construct and test (as well as other possible influences) provides an adequate representation of the situation. Thus, content validity can be conceived of as an aspect of construct validity.

Convergent and *discriminant validity* are concerned with whether the observed correlations between measures correspond to the correlations predicted by the theory (or model). Campbell & Fiske (1959) used observed correlation in order to assess these types of validities. Bollen (1989) exhibited the shortcomings of this approach and showed that a proper treatment has to be based on latent variable models. Thus, these two types of validity can also be regarded as special cases of construct validity.

In summary, the concept of construct validity refers to the correctness of the measurement model and, by consequence, the inferences based on this model. It comprises all the different sorts of theoretical validity. The validity of a test (cf. Concept 4-18, page 124) is but one aspect of construct validity that concerns the relationship between the latent construct and its measures.

4.5.4 Model Based Measures of the Validity of a Test

The previous discussion exhibits that in case of measuring theoretical constructs the empirical measures of validity can be misleading (cf. Ex. 4-21 on page 122). The following example provides a further illustration.

Ex. 4-24: The validity of a test:

Given:

1. The target construct η_T to be measured: emotional intelligence;
2. An observed measure Y_T that is an indicator of η_T;
3. A further construct η that is closely related to the target construct η_T.
4. The measure Y is an indicator of the latent construct η (but not of the target construct η_T): social competence.
5. An additional latent construct ξ: verbal ability.

The measurement model in Figure 4-25 represents the measurement situation.

Obviously, the test Y is not a valid test of the target construct η_T, since there is not direct causal influence of this latent construct on Y, and, consequently, the construct η_T does not explain any systematic variance in Y.

Nevertheless, a high correlation between the criterion Y_T and Y might be observed, due to a high correlation between the two constructs η_T and η, as well as the fact that both measures are influenced by a third latent factor ξ.

Ex. 4-24 demonstrates again the shortcomings of empirical measures of validity. Moreover, it also provides an interpretation of Kelley's (1927) definition of validity within the framework of causal models:

A test measures the latent construct it purports to measure if the construct exerts a direct causal influence on the measure.

Note that if the construct η_T is not a direct cause of a test Y then a set of variables can be found (e.g. construct η in Ex. 4-24) such that η_T does not explain any systematic variance in Y as soon as these variables have been taken into account. We, now present two measures of the validity of a test that are based on these considerations.

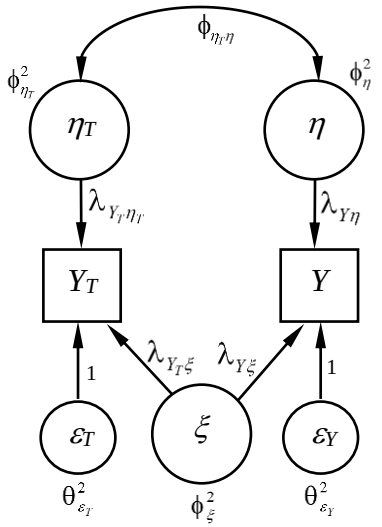


Figure 4-25: *Model of two measurements for measuring latent constructs: η_T = emotional intelligence, η = social competence, and ξ = verbal abilities.*

4.5.4.1 THE LOADING COEFFICIENT AS A MEASURE OF VALIDITY

The definition of the validity of a test Y as a measure of a target construct η (Concept 4-18, page 124), as well as the previous considerations the validity of Y is best gauged by the loading coefficient $\lambda_{Y\eta}$ representing the direct causal influence of the latent construct on the measure.

The unstandardized loading coefficients $\lambda_{Y\eta}$ cannot be used since it depends on the scales of the latent and observed variable. Thus the standardized coefficient $\lambda^s_{Y\eta}$ has to be employed as a measure of the validity of a test.

Comment 4-11:

Remember the relationship between the standardized $\lambda^s_{Y\eta}$ and the unstandardized loading coefficient $\lambda_{Y\eta}$:

$$\lambda^s_{Y\eta} = \lambda_{Y\eta} \cdot \frac{\sigma_\eta}{\sigma_Y},$$

where σ_η and σ_Y denote the standard deviations of the latent construct η and the observed score *Y*, respectively.

If the target construct is the only latent construct exerting an influence on the measure *Y* one gets the following relationship between the reliability of *Y* and the standardized loading coefficient $\lambda^s_{Y\eta}$:

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

This relationship is in accordance with the principle that the maximal validity corresponds to the square root of the reliability of the measure (see, for example, Angoff, 1988).

4.5.4.2 UNIQUE TRUE SCORE VARIANCE AND RELIABILITY

Bollen (1989) presents a second measure of validity that he calls the *unique validity variance*. In the following we use, instead, the name *unique reliability* since this appears to be a better name.

Concept 4-20: *Unique reliability:*

The *unique reliability* $\mathrm{Rel}_\eta(Y)$ consists in that portion of the variance of the measure *Y* that can be *unambiguously* attributed to the latent target construct η that the test *Y* intends to measure. The unique reliability is given by:

$$\mathrm{Rel}_\eta(Y) = \frac{\lambda^2_{Y\eta} \cdot \mathrm{Var}_{\mathrm{unique}}(\eta)}{\mathrm{Var}(Y)} \qquad (4\text{-}46)$$

The symbols have the following meaning:

$\lambda^2_{Y\eta}$	denotes the squared loading coefficient of the path $\eta \rightarrow Y$.
$\mathrm{Var}(Y)$	symbolizes the variance of the test *Y*.
$\mathrm{Var}_{\mathrm{unique}}(\eta)$	denotes the variance of the construct η, that cannot be explained by the other constructs.

Comments:

1. The equation of the unique reliability conforms to the equation of the reliability of a test in case of η being the only latent construct exerting an influence on the measure *Y*. Hence, in this case the unique reliability is simply the reliability of the test.
2. $\mathrm{Var}_{\mathrm{unique}}(\eta)$ corresponds to the residual variance after partialling out the total variance of η that variance which η shares with the other constructs exerting an influence of *Y* (cf. Method 4-7).

Note that since the target construct η may be correlated with other latent constructs, a part of the variation of η can be explained by the varaiotion of the other constructs. This variation in η explained by the other construct should not enter into the computation of the unique reliability.

3. $\lambda_{Y\eta}^2 \cdot \mathrm{Var}_{\mathrm{unique}}(\eta)$ can be denoted as the *unique true score variance:* The variation of the measure *Y* that can be attributed uniquely to the variation of the target construct η.

The computation of the unique reliability, requires the calculation of the unique true score variance $\mathrm{Var}_{\mathrm{unique}}(\eta)$. Method 4-7 explains how this can be done by means of matrices.

Method 4-7: Partialling out of the target construct the variance that can be explained by variations of the other constructs.

Given:

- The factor analytic model (of first order) with *p* latent variables $\eta, \xi_1, \xi_2, \ldots, \xi_{p-1}$ that exert and influence on test *Y*. η symbolizes the target construct the test intends to measure (cf. Figure 4-26). The symbols $\xi_i \quad (i = 1, 2, \ldots, p-1)$ denote the other constructs.
- The variances of and the covariances between the latent constructs $\eta, \xi_1, \xi_2, \ldots, \xi_{p-1}$ have been estimated from the data on the basis of a test model (a linear structural equation model).

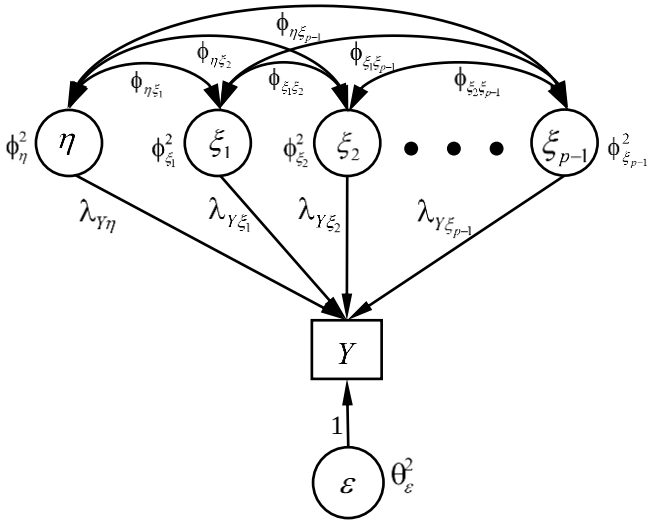


Figure 4-26: *A test Y, measuring the latent constructs $\xi_1, \xi_2, \ldots, \xi_{p-1}$, additionally to the target construct η.*

Wanted:

The residual variance $\mathrm{Var}_{\mathrm{unique}}(\eta)$ after removing the variance of η that can be explained by the constructs: $\xi_1, \xi_2, \ldots, \xi_{p-1}$.

Die residual variance can be computed using linear regression with the target construct η as the dependent and the other constructs $\xi_1, \xi_2, \ldots, \xi_{p-1}$ as the independent variables:

$$\eta = \lambda_{\eta\xi_1} \cdot \xi_1 + \lambda_{\eta\xi_2} \cdot \xi_2 + \cdots + \lambda_{\eta\xi_{p-1}} \cdot \xi_{p-1} + \zeta . \qquad (4\text{-}47)$$

The symbols have the following meaning:

$\lambda_{\eta\xi_1}, \lambda_{\eta\xi_2}, \ldots, \lambda_{\eta\xi_{p-1}}$ denote the regression coefficients, and the symbol ζ (zeta) denotes the residual term. The variance ψ_ζ^2 of ζ corresponds to the unique variance $\mathrm{Var}_{\mathrm{unique}}(\eta)$ that cannot be explained by the independent variables $\xi_1, \xi_2, \ldots, \xi_{p-1}$.

The residual variance can be explained by means of a simple matrix equation:

$$\mathrm{Var}_{\mathrm{unique}}(\eta) = \mathrm{Var}(\eta) - \left(\boldsymbol{\phi}_{\eta\xi}^{\mathbf{T}} \cdot \boldsymbol{\Phi}_{\xi}^{-1} \cdot \boldsymbol{\phi}_{\eta\xi}\right), \qquad (4\text{-}48)$$

where:

$\boldsymbol{\phi}_{\eta\xi}$ denotes a $(p-1)\times 1$ columns vector containing the covariances between the target construct η and the other latent constructs $\xi_1, \xi_2, \ldots, \xi_{p-1}$ exerting an influence on *Y*.

$\boldsymbol{\Phi}_{\xi}^{-1}$ represents the inverse covariance matrix [of dimension $(p-1)\times(p-1)$] between the latent constructs $\xi_1, \xi_2, \ldots, \xi_{p-1}$.

$\mathrm{Var}(\eta)$ symbolizes the variance of the target construct η.

Comment:

The expression $\boldsymbol{\phi}_{\eta\xi}^{\mathbf{T}} \cdot \boldsymbol{\Phi}_{\xi}^{-1} \cdot \boldsymbol{\phi}_{\eta\xi}$ represents that part of the variance of η that can be explained by the variables $\xi_1, \xi_2, \ldots, \xi_{p-1}$.

The following example illustrates the computation of the unique true score variances and reliabilities, respectively.

Ex. 4-25: Unique reliabilities

Given: The model of Figure 4-27.

Wanted: The unique reliability of *Y* with respect to η_1, η_2, and η_3.

1. Die residual variances, after regressing the target construct on the other constructs are:
   $\mathrm{Var}_{\mathrm{unique}}(\eta_1) = 1.893$,
   $\mathrm{Var}_{\mathrm{unique}}(\eta_2) = 1.343$, and
   $\mathrm{Var}_{\mathrm{unique}}(\eta_3) = 2.833$
2. The unique reliabilities for the single constructs are:

$$\mathrm{Rel}_{\eta_1}(Y) = \frac{\lambda^2_{Y\eta_1} \cdot \mathrm{Var}_{\mathrm{unique}}(\eta_1)}{\mathrm{Var}(Y)} = \frac{1.5^2 \cdot 1.893}{7.645} = 0.557,$$

$$\mathrm{Rel}_{\eta_2}(Y) = \frac{\lambda^2_{Y\eta_2} \cdot \mathrm{Var}_{\mathrm{unique}}(\eta_2)}{\mathrm{Var}(Y)} = \frac{0.7^2 \cdot 1.343}{7.645} = 0.086, \text{ and}$$

$$\mathrm{Rel}_{\eta_3}(Y) = \frac{\lambda^2_{Y\eta_3} \cdot \mathrm{Var}_{\mathrm{unique}}(\eta_3)}{\mathrm{Var}(Y)} = \frac{0.3^2 \cdot 2.833}{7.645} = 0.033.$$

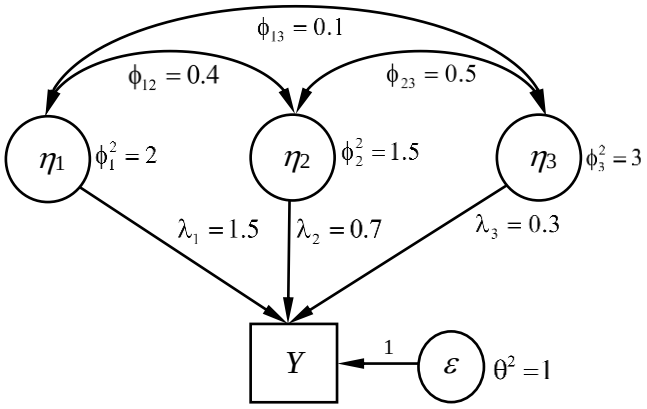


Figure 4-27: *Structural equation model for illustrating the computation of unique reliabilities.*

The detailed computation of $\mathrm{Var}_{\mathrm{unique}}(\eta_1)$ looks like this:

The vector of covariances between η_1 and the two other constructs is given by:

$$\boldsymbol{\phi}_{\eta_1,[\eta_2\eta_3]} = \begin{bmatrix} 0.4 \\ 0.1 \end{bmatrix}, \text{ and } \boldsymbol{\phi}^{\mathbf{T}}_{\eta_1,[\eta_2\eta_3]} = \begin{bmatrix} 0.4 & 0.1 \end{bmatrix}, \text{ respectively.}$$

The covariance matrix of the other two constructs looks like this:

$$\boldsymbol{\Phi}_{[\eta_2 \ \ \eta_3]} = \begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 3.0 \end{bmatrix}.$$

Consequently,

$$\begin{aligned}\mathrm{Var}_{\mathrm{unique}}(\eta_1) &= \mathrm{Var}(\eta_1) - \left(\boldsymbol{\phi}^{\mathbf{T}}_{\eta_1,[\eta_2\eta_3]} \cdot \boldsymbol{\Phi}^{-1}_{[\eta_2\ \ \eta_3]} \cdot \boldsymbol{\phi}_{\eta,[\eta_2\eta_3]}\right) \\ &= 2.0 - \begin{bmatrix}0.4 & 0.1\end{bmatrix} \cdot \begin{bmatrix}1.5 & 0.5 \\ 0.5 & 3.0\end{bmatrix}^{-1} \cdot \begin{bmatrix}0.4 \\ 0.1\end{bmatrix} \\ &= 2.0 - 0.107 \\ &= \underline{\underline{1.893}}\end{aligned}$$

The unique variances of the other constructs may be computed in a similar way.

The unique reliability depends on the correlation between the target construct η that the test *Y* intends to measure and the other latent constructs on which the test is loading: the higher this correlation the lower the unique variance of η.

By consequence, unique reliabilities do not sum to that reliability of the test *Y*. For instance, the sum of unique reliabilities in Ex. 4-25 is 0.677, whereas the reliability is: Rel(*Y*) = 0.869. The sum of the unique reliabilities corresponds to the reliability of the test only in case of the latent constructs being uncorrelated.

4.5.5 The Validity–Reliability-Paradox

In some textbooks (e.g. Schmid, 1992; Rost, 2004) one can find the following statement that has been termed the *validity-reliability paradox*:

> *Increasing the reliability of a test can result in a reduction of the validity of the test.*

An increase of the reliability of the sum of test items can result in a violation of the content validity of the test if the items added differ only slightly in their content. As a result the test items do not capture all aspects (facets) of the construct which amounts to a violation of construct validity.

However, this limitation of content validity is not meant by the validity-reliability paradox. Rather, the latter refers to the fact that *the validity coefficient (i.e. the correlation between the test scores and a given criterion C) can be reduced by increasing the reliability of the single measures.*

Assume, for the sake of concreteness, that there are *n* parallel tests $Y_1, Y_2, \ldots, Y_n$, as well as a criterion *C*. Assuming that the correlation $\mathrm{Corr}(C, Y_i)$ is the same for each test item Y_i it can be shown (Exercise 4-28) that the correlation $\mathrm{Corr}(C, Y)$ between the criterion *C* and the sum $Y = Y_1 + Y_2 + \cdots + Y_n$ of the test items is given by:

$$\mathrm{Corr}(C,Y)=\frac{\sqrt{n}\cdot\mathrm{Corr}(C,Y_i)}{\sqrt{1+(n-1)\cdot\mathrm{Rel}(Y_i)}}. \qquad (4\text{-}49)$$

The symbols have the following meaning:

$\mathrm{Corr}(C,Y_i)$ denotes the correlation between the criterion C and the test item Y_i. It is assumed that this correlation is the same for all test items.

$\mathrm{Rel}(Y_i)$ denotes the reliability of test item Y_i that corresponds to the correlation $\mathrm{Corr}(Y_i,Y_j)$ between the single test items. Note that since the test items are parallel $\mathrm{Corr}(Y_i,Y_j)$ is identical for all pairs of test items, and, by consequence, $\mathrm{Rel}(Y_i)$ is the same for all test items.

Increasing $\mathrm{Rel}(Y_i)$ results in a greater denominator and thus in a decrease of the validity coefficient $\mathrm{Corr}(C,Y)$.

This way of reasoning has a major drawback: *It ignores the latent variable structure.* If the latter is taken into account, the alleged paradox disappears. From the perspective of latent variable models the situation allows for two basic types of models:

1. The criterion *C,* on the one hand, and the measures $Y_1,Y_2,\ldots,Y_n$, on the other hand, load on different constructs: In this case there exists no paradox at all since $Y_1,Y_2,\ldots,Y_n$ are, obviously, not valid measures of the construct that is measured by the criterion *C*.
2. The criterion *C* and the measures $Y_1,Y_2,\ldots,Y_n$ are measuring the same underlying construct. In this case, an increase of $\mathrm{Rel}(Y_i)$ can never lead to a decrease of $\mathrm{Corr}(C,Y)$.

Let us consider the two cases in greater detail. Figure 4-28 exhibits a latent variable model representing the first case. Note that the *n* parallel tests $Y_1,Y_2,\ldots,Y_n$ are not loading on the target construct η that is measured by the criterion *C*. Rather they are loading on a different construct ξ. Consequently, $Y_1,Y_2,\ldots,Y_n$ cannot be regarded as valid tests of the target construct η.

By consequence, the empirical validity coefficient $\mathrm{Corr}(C,Y)$ and its size, respectively, is not a useful measure of the validity of the tests Y_i $(i=1,2,\ldots,n)$.

The reliability $\mathrm{Rel}(Y)$ can be made greater by increasing the variance ϕ_ξ^2 of the construct ξ (i.e. the population becomes more heterogeneous

with respect to the latent construct ξ, everything else staying the same). In this case $\mathrm{Rel}(Y)$ increases since (Exercise 4-29):

$$\mathrm{Rel}(Y) = \frac{n \cdot \phi_\xi^2}{n \cdot \phi_\xi^2 + \theta_\varepsilon^2} = \frac{1}{1 + \theta_\varepsilon^2 / \left(n \cdot \phi_\xi^2\right)}. \qquad (4\text{-}50)$$

Hence, increasing the variance ϕ_ξ^2 results in a decreased denominator, and, by consequence, in a higher reliability of the sum.

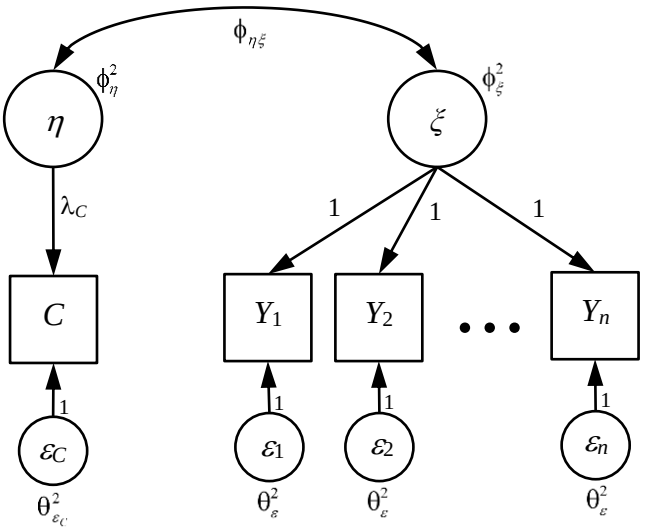


Figure 4-28: *Structural equation model for illustrating the validity-reliability paradox.*

The correlation $\mathrm{Corr}(C,Y)$ is given by (Exercise 4-29):

$$\mathrm{Corr}(C,Y) = \frac{\lambda_C \cdot \phi_{\eta\xi}}{\sqrt{\lambda_C^2 \cdot \phi_\eta^2 + \theta_{\varepsilon_C}^2} \cdot \sqrt{\phi_\xi^2 + \frac{\theta_\varepsilon^2}{n}}}. \qquad (4\text{-}51)$$

Increasing the variance ϕ_ξ^2 increases the denominator thereby decreasing the correlation $\mathrm{Corr}(C,Y)$. Consequently, increasing the reliability of the sum Y results in a reduction of the validity coefficient $\mathrm{Corr}(C,Y)$. However, as noted above this cannot be interpreted as a case of a validity-reliability paradox, since, due to the fact the measures Y_i and Y_C are not measuring the same latent construct, the validity coefficient is not a sensible measure of validity of the sum Y.

Let us now consider the second case where the test items Y_i are valid indicators of the target construct η (cf. Figure 4-29). For this model the validity coefficient turns out to be (Exercise 4-30):

$$\mathrm{Corr}(C,Y) = \frac{\lambda_C}{\sqrt{\lambda_C^2 + \frac{\theta_{\varepsilon_C}^2}{\phi_\eta^2}} \cdot \sqrt{1 + \frac{\theta_\varepsilon^2}{n \cdot \phi_\eta^2}}}. \qquad (4\text{-}52)$$

To increase the reliability of a test Y_i, one hast to either increase the variance ϕ_η^2 or decrease the error variances θ_ε^2. In both cases the validity coefficient $\mathrm{Corr}(C,Y)$ becomes greater since the dominator of Equation (4-52) decreases.

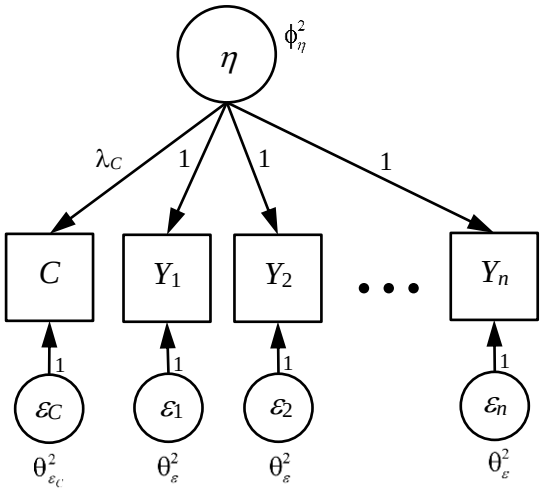


Figure 4-29: *Structural equation model used to illustrate the validity-reliability paradox.*

This result seems to be in contraction to Equation (4-49) according to which the increase of the reliability of a valid test should lead to a decrease of the correlation between the test and the criterion. However, this contradiction is a spurious one since the increase of $\mathrm{Rel}(Y_i)$ is accompanied by an increase of $\mathrm{Corr}(C,Y_i)$ in the nominator of Equation (4-49). Thus the increase of $\mathrm{Rel}(Y_i)$ increases both the nominator and the denominator of Equation (4-49) resulting in a net increase of $\mathrm{Corr}(C,Y)$.

The preceding discussion demonstrates that the validity-reliability paradox that may be observed on the level of observed correlations vanish into thin air as soon as the latent variable structure is taken into account.

Comment 4-12:

It might be argued that the models discussed above do not take unexplained covariances $\theta_{\varepsilon,\varepsilon_C}$ between the residual of the criterion *C* and the error terms of the test items Y_i into account.

Note, however, that unexplained covariances $\theta_{\varepsilon,\varepsilon_C}$ have no effect on the reliability $\mathrm{Rel}(Y)$. Thus, introducing unexplained covariance arcs between *C* and the test items Y_i into the model of Figure 4-29 does not invalidate the given argument.

4.6 Mean Structures

The previous presentation was concerned predominantly with the analysis of the structure of tests. Consequently the analysis of the covariance structure that represents the *structural aspect* of the tests in CTT was in the focus of interest.

In this chapter we turn to the analysis of the means and intercepts that constitute the mean structure of the tests. The means and intercepts represent the *performance aspect* like the difficulty of items or the mean ability of the population in question.

The chapter comprises three parts: Chapter 4.6.1 presents the linear structural equation model used for modeling covariance and mean structures. Chapter 4.6.2 is concerned with the problem of predicting latent construct values on the basis of the observed test scores. Finally, Chapter 4.6.3 discusses the problem of comparing different groups.

4.6.1 Modeling Mean Structures Using Linear Structural Equation Models

Linear structural equation models can be used to model mean structures. The mean structure is represented by two types of parameters:

1. Mean parameters representing the means of exogenous parameters.
2. Intercept parameters associated with endogenous variables. The intercept parameters constitute the regression constants of the linear structural equations.

Since the parameters constituting the mean structure are not random variables they are irrelevant with respect to the covariance structure. Figure 4-30 depicts the general linear structural model with mean and intercept parameters that are represented by triangles. The symbols concerning the parameters of the mean structure have the following meaning:

$\mu_{\eta_1}, \mu_{\eta_2}, \ldots, \mu_{\eta_p}$ denote the means of the latent constructs;

$\mu_{\varepsilon_1}, \mu_{\varepsilon_2}, \ldots, \mu_{\varepsilon_n}$ represent the means of the errors;

$\alpha_{Y_1}, \alpha_{Y_2}, \ldots, \alpha_{Y_n}$ symbolize the intercepts or regression constants.

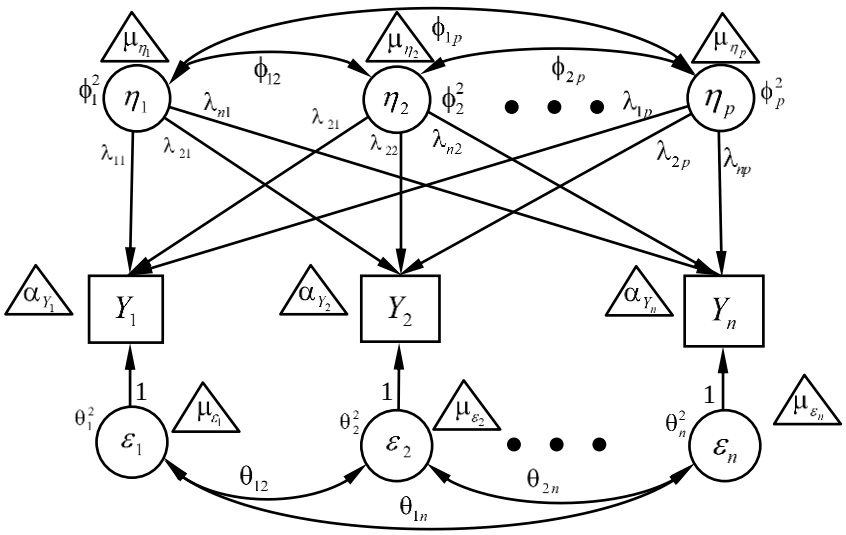


Figure 4-30: *The general factor analytic model of first order with mean and intercept parameters.*

The system of linear equations with mean parameters looks like this:

$$\begin{aligned} Y_1 &= \alpha_{Y_1} + \lambda_{11}\cdot\eta_1 + \lambda_{12}\cdot\eta_2 + \cdots + \lambda_{1p}\cdot\eta_p + \varepsilon_1 \\ Y_2 &= \alpha_{Y_2} + \lambda_{21}\cdot\eta_1 + \lambda_{22}\cdot\eta_2 + \cdots + \lambda_{2p}\cdot\eta_p + \varepsilon_2 \\ &\vdots \qquad\qquad \cdots \\ Y_n &= \alpha_{Y_n} + \lambda_{n1}\cdot\eta_1 + \lambda_{n2}\cdot\eta_2 + \cdots + \lambda_{np}\cdot\eta_p + \varepsilon_n \end{aligned} \quad . \qquad (4\text{-}53)$$

The addition of means and intercepts enables the modeling of the means of the observed test scores:

$$\begin{aligned} \mu_{Y_1} &= \alpha_{Y_1} + \lambda_{11}\cdot\mu_{\eta_1} + \lambda_{12}\cdot\mu_{\eta_2} + \cdots + \lambda_{1p}\cdot\mu_{\eta_p} \\ \mu_{Y_2} &= \alpha_{Y_2} + \lambda_{21}\cdot\mu_{\eta_1} + \lambda_{22}\cdot\mu_{\eta_2} + \cdots + \lambda_{2p}\cdot\mu_{\eta_p} \\ &\vdots \qquad\qquad \cdots \\ \mu_{Y_n} &= \alpha_{Y_n} + \lambda_{n1}\cdot\mu_{\eta_1} + \lambda_{n2}\cdot\mu_{\eta_2} + \cdots + \lambda_{np}\cdot\mu_{\eta_p} \end{aligned} \quad . \qquad (4\text{-}54)$$

Equation (4-54) results from equation (4-53) by applying the rules for expectations and the assumption that the means of the errors are all equal to zero: $\mu_{\varepsilon_i} = 0 \ \ (i = 1, 2, \ldots, n)$.

Comment 4-13: *Fixing the means of the error terms:*

Similar to fixing the loading coefficients associated with the error variables to 1 the means of the error variables are, in general fixed to 0. This represents the assumptions that the errors are not systematically biased.

The setting of the means of the error terms to zero does not constitute any restriction of the model since any systematic influences are represented by the regression constants. This includes systematic biases of the errors.

The system of equations in (4-54) models the means of the test items by means of regression constants as well as the means of the latent constructs and the loading coefficients. The estimation of the mean and intercept parameters is based on the observed means $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_n$. Consequently, *n* observed means are added to the $n \cdot (n+1)/2$ free data points of observed variances and covariances resulting in $n \cdot (n+3)/2$ free data points.
Since that are more mean and intercept parameters than observed sample means the parameters of the mean structure are unidentified as long as there are not further restrictions on these parameters.
The following example demonstrates the usage of mean structures as well as the fixing of the parameters of the mean structure.

Ex. 4-26: τ-equivalent versus essential τ-equivalent model:

In Section 4.2.3.3 (page 56) the distinction between τ-equivalent and essential τ-equivalent tests has been introduced. It was also mentioned that this distinction is relevant only for models representing means structures.

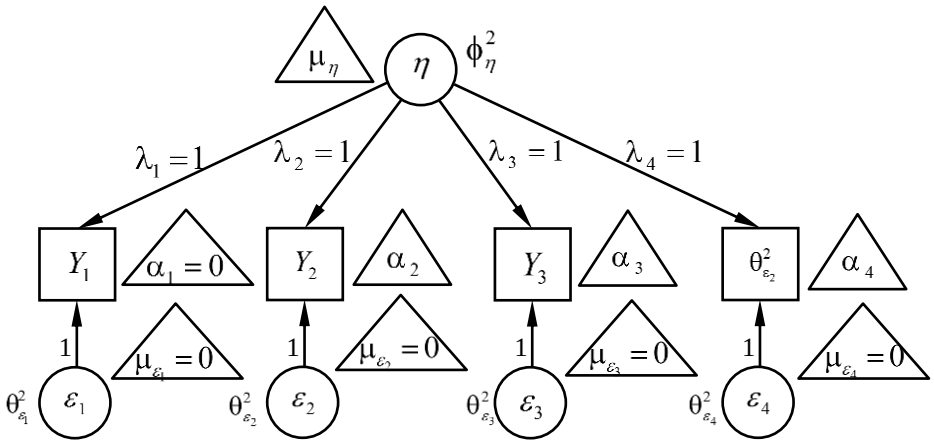


Figure 4-31: *Linear structural equation model of four essential τ-equivalent measures.*

Figure 4-31 illustrates the model of essential τ-equivalent tests. The following restrictions have been imposed:

1. The regression constant of the first test has been fixed to 0: $\alpha_1 = 0$.
2. The mean of the latent construct μ_η as well as the regression constants α_2, α_3 and α_4 are free parameters that are estimated from the data.

For the model in Figure 4-31 (with the given restrictions) the estimated mean $\hat{\mu}_\eta$ corresponds to the observed mean of Y_1. The estimated intercepts $\hat{\alpha}_2, \hat{\alpha}_3$ and $\hat{\alpha}_4$ correspond to the differences between the observed means of the associated variables minus the observed mean of variable Y_1: $\hat{\alpha}_i = \bar{Y}_i - \bar{Y}_1$ $(i = 2,3,4)$.

The model of essential τ-equivalent test does not specify any restrictions on the observed means. Consequently it provides no testable predictions with respect to the mean structure of the tests (Note that the number of free parameters of the mean structure corresponds to the number of observed means).

The model of τ-equivalent tests results from the given model of essential τ-equivalent tests by setting the intercept parameters to zero: $\alpha_2 = 0, \alpha_3 = 0$, and $\alpha_4 = 0$.

Thus, the model of τ-equivalent tests predicts that the observed means are identical.

Figure 4-32 presents an alternative parameterization of the essential τ-equivalent model.

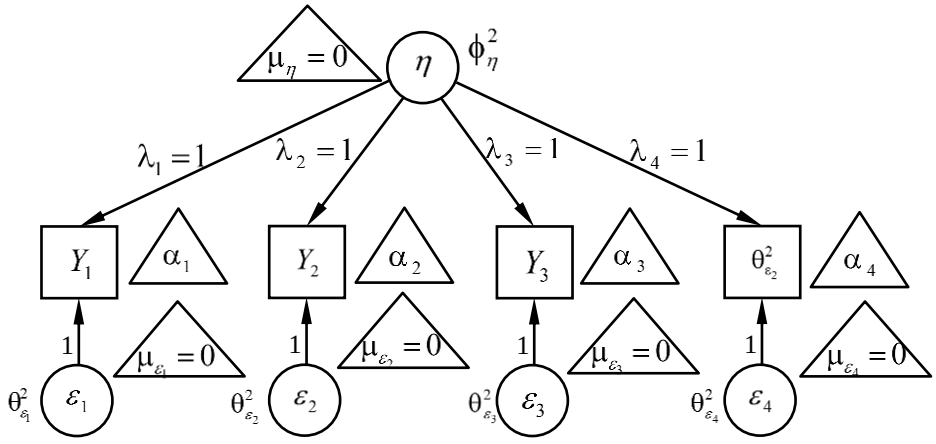


Figure 4-32: *Linear structural equation model of four essential τ-equivalent measures (Alternative parametrization).*

The τ-equivalent model results by equating the intercept parameters: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$. This leads to the same prediction as for the parameterization of Figure 4-31: The model predicted means are all equal.

Notation 4-9:

Since that means of the errors are always fixed to zero these means are no longer shown in the subsequent presentation.

Having specified the structural equation model for modeling the mean structure of tests we next discuss a first application that requires mean structures.

4.6.2 Prediction of Latent Construct Scores from Observed Test Scores

Assume that an examinee has been tested using *n* test items. The observed test scores are $Y_1, Y_2, \ldots, Y_n$. Since the main function of the observed test scores consists in providing information about the latent construct scores one requires a method for inferring the values of the person on the latent constructs on the basis of the observed test scores. There exist different methods to predict the latent construct score from the observed test scores. In the following, two estimators for predicting latent construct scores are discussed:

1. The regression (least squares) predictor, and
2. The maximum likelihood predictor.

4.6.2.1 LEAST SQUARES PREDICTOR OF LATENT CONSTRUCT SCORES

The method of least squares provides the best linear prediction and, in case of normally distributed data the best prediction, in the sense of expected squared deviation (Searle, Casella, & McCulloch, 1992, Chapter 7).

Method 4-8: *Least squares (LS) predictor of the latent construct scores*

Given:

The general factor analytic model of Figure 4-30 (page 139). In the subsequent presentation the following symbols are used:

$\hat{\boldsymbol{\eta}}^{\mathbf{T}} = \left[\tilde{\eta}_1, \tilde{\eta}_2, \ldots, \tilde{\eta}_p\right]$ denotes the vector of the predicted scores for the *p* latent constructs;

$\hat{\boldsymbol{\mu}}_{\boldsymbol{\eta}}^{\mathbf{T}} = \left[\hat{\mu}_{\eta_1}, \hat{\mu}_{\eta_2}, \ldots, \hat{\mu}_{\eta_p}\right]$ represents the vector of estimated mean parameters of the *p* latent constructs;

$\mathbf{Y}^{\mathbf{T}} = \left[Y_1, Y_2, \ldots, Y_n\right]$ denotes the vector of the *n* observed test scores.

$\boldsymbol{\mu}_{\mathbf{Y}}^{\mathbf{T}} = \left[\mu_{Y_1}, \mu_{Y_2}, \ldots, \mu_{Y_n}\right]$ symbolizes the mean vector of observed test scores.

$\hat{\mathbf{\Phi}}$ denotes the ($p \times p$) covariance matrix containing the estimated variances and covariances of the p latent constructs:

$$\hat{\mathbf{\Phi}} = \begin{array}{c} \\ \eta_1 \\ \eta_2 \\ \vdots \\ \eta_p \end{array} \begin{array}{c} \begin{array}{cccc} \eta_1 & \eta_2 & \cdots & \eta_p \end{array} \\ \begin{bmatrix} \hat{\phi}_1^2 & \hat{\phi}_{12} & \cdots & \hat{\phi}_{1p} \\ \hat{\phi}_{21} & \hat{\phi}_2^2 & \cdots & \hat{\phi}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\phi}_{p1} & \hat{\phi}_{p2} & \cdots & \hat{\phi}_p^2 \end{bmatrix} \end{array}.$$

$\hat{\mathbf{\Lambda}}$ represents the ($n \times p$) matrix with the estimated loading coefficients:

$$\hat{\mathbf{\Lambda}} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{array} \begin{array}{c} \begin{array}{cccc} \eta_1 & \eta_2 & \cdots & \eta_p \end{array} \\ \begin{bmatrix} \hat{\lambda}_{11} & \hat{\lambda}_{12} & \cdots & \hat{\lambda}_{1p} \\ \hat{\lambda}_{21} & \hat{\lambda}_{22} & \cdots & \hat{\lambda}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\lambda}_{n1} & \hat{\lambda}_{n2} & \cdots & \hat{\lambda}_{np} \end{bmatrix} \end{array}.$$

$\hat{\mathbf{\Sigma}}_{\mathbf{Y}}$ symbolizes the estimated ($n \times n$) covariance matrix of the observed test scores (i.e. the covariance matrix of the test scores predicted by the model):

$$\hat{\mathbf{\Sigma}}_{\mathbf{Y}} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & \cdots & Y_n \end{array} \\ \begin{bmatrix} \hat{\sigma}_{Y_1}^2 & \hat{\sigma}_{Y_1Y_2} & \cdots & \hat{\sigma}_{Y_1Y_n} \\ \hat{\sigma}_{Y_2Y_1} & \hat{\sigma}_{Y_2}^2 & \cdots & \hat{\sigma}_{Y_2Y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{Y_nY_1} & \hat{\sigma}_{Y_nY_2} & \cdots & \hat{\sigma}_{Y_n}^2 \end{bmatrix} \end{array}.$$

The least squares (LS) predictor $\hat{\boldsymbol{\eta}}$ of latent construct scores is given by the following matrix formula:

$$\tilde{\boldsymbol{\eta}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\eta}} + \hat{\mathbf{\Phi}} \cdot \hat{\mathbf{\Lambda}}^{\mathbf{T}} \cdot \hat{\mathbf{\Sigma}}_{\mathbf{Y}}^{-1} \cdot \left(\mathbf{Y} - \hat{\boldsymbol{\mu}}_{\mathbf{Y}} \right). \qquad (4\text{-}55)$$

Comments concerning the form of the least squares predictor:

1. The least squares predictor has the following structure:

$$\tilde{\boldsymbol{\eta}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\eta}} + \hat{\mathbf{Kov}}\left(\boldsymbol{\eta}, \mathbf{Y}^{\mathbf{T}} \right) \cdot \hat{\mathbf{Var}}\left(\mathbf{Y} \right)^{-1} \cdot \left(\mathbf{Y} - \hat{\boldsymbol{\mu}}_{\mathbf{Y}} \right).$$

$\hat{\mathbf{Kov}}\left(\boldsymbol{\eta}, \mathbf{Y}^{\mathbf{T}} \right)$ symbolizes the estimated covariance matrix between the latent constructs $\boldsymbol{\eta}$ and the test scores $\mathbf{Y}$. This matrix can be computed as follows: $\hat{\mathbf{Kov}}\left(\boldsymbol{\eta}, \mathbf{Y}^{\mathbf{T}} \right) = \hat{\mathbf{\Phi}} \cdot \hat{\mathbf{\Lambda}}^{\mathbf{T}}$.

$\hat{\mathbf{Var}}\left(\mathbf{Y} \right)$ denotes the estimated covariance matrix of the test scores $\mathbf{Y}$: $\hat{\mathbf{Var}}\left(\mathbf{Y} \right) = \hat{\mathbf{\Sigma}}_{\mathbf{Y}}$.

The following equation results from the theory of least squares:

$$\hat{\mathbf{B}} = \mathbf{K\hat{o}v}\left(\boldsymbol{\eta}, \mathbf{Y}^{\mathbf{T}}\right) \cdot \mathbf{V\hat{a}r}\left(\mathbf{Y}\right)^{-1}. \qquad (4\text{-}56)$$

The symbol $\hat{\mathbf{B}}$ denotes the ($p \times n$) matrix of the estimates least squares (regression) coefficients of the regression of the latent constructs on the observed test scores. It has the following structure:

$$\hat{\mathbf{B}} = \begin{array}{c} \\ \eta_1 \\ \eta_2 \\ \vdots \\ \eta_p \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & \cdots & Y_n \end{array} \\ \begin{bmatrix} \hat{\beta}_{11} & \hat{\beta}_{12} & \cdots & \hat{\beta}_{1n} \\ \hat{\beta}_{21} & \hat{\beta}_{22} & \cdots & \hat{\beta}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\beta}_{p1} & \hat{\beta}_{p2} & \cdots & \hat{\beta}_{pn} \end{bmatrix} \end{array}$$

Equation (4-55) can thus be written as:

$$\tilde{\boldsymbol{\eta}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\eta}} + \hat{\mathbf{B}} \cdot \left(\mathbf{Y} - \hat{\boldsymbol{\mu}}_{\mathbf{Y}}\right). \qquad (4\text{-}57)$$

(4-57) represents the following system of equations:

$$\begin{aligned} \tilde{\eta}_1 &= \hat{\mu}_{\eta_1} + \hat{\beta}_{11} \cdot \left(Y_1 - \hat{\mu}_{Y_1}\right) + \hat{\beta}_{12} \cdot \left(Y_2 - \hat{\mu}_{Y_2}\right) + \cdots + \hat{\beta}_{1n} \cdot \left(Y_n - \hat{\mu}_{Y_n}\right) \\ \tilde{\eta}_2 &= \hat{\mu}_{\eta_2} + \hat{\beta}_{21} \cdot \left(Y_1 - \hat{\mu}_{Y_1}\right) + \hat{\beta}_{22} \cdot \left(Y_2 - \hat{\mu}_{Y_2}\right) + \cdots + \hat{\beta}_{2n} \cdot \left(Y_n - \hat{\mu}_{Y_n}\right) \\ \vdots &= \vdots \\ \tilde{\eta}_p &= \hat{\mu}_{\eta_p} + \hat{\beta}_{p1} \cdot \left(Y_1 - \hat{\mu}_{Y_1}\right) + \hat{\beta}_{p2} \cdot \left(Y_2 - \hat{\mu}_{Y_2}\right) + \cdots + \hat{\beta}_{pn} \cdot \left(Y_n - \hat{\mu}_{Y_n}\right) \end{aligned},$$

where:

$$\tilde{\boldsymbol{\eta}} = \begin{bmatrix} \tilde{\eta}_1 \\ \tilde{\eta}_2 \\ \vdots \\ \tilde{\eta}_p \end{bmatrix}, \ \hat{\boldsymbol{\mu}}_{\boldsymbol{\eta}} = \begin{bmatrix} \hat{\mu}_{\eta_1} \\ \hat{\mu}_{\eta_2} \\ \vdots \\ \hat{\mu}_{\eta_p} \end{bmatrix}, \ \mathbf{Y} - \hat{\boldsymbol{\mu}}_{\mathbf{Y}} = \begin{bmatrix} Y_1 - \hat{\mu}_{Y_1} \\ Y_2 - \hat{\mu}_{Y_2} \\ \vdots \\ Y_n - \hat{\mu}_{Y_n} \end{bmatrix}.$$

(The structure of matrix $\hat{\mathbf{B}}$ is shown above).

2. The LS predictor of Equation (4-57) represents the multivariate version of the least squares estimator of the simple linear regression model:
   Let $y = \alpha + \beta \cdot x + \varepsilon$ denote the simple linear regression equation. The predictor of y is given by $\tilde{y} = \hat{\alpha} + \hat{\beta} \cdot x$, with $\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}$ and $\hat{\beta} = s_{XY} / s_X^2$.

The symbols $\bar{x}$ and $\bar{y}$ denote the sample of *x* and *y*. s_{XY} and s_X^2 denote, respectively, the sample covariance between *x* and *y* as well as the sample variance of *x*.

Replacing the estimator $\hat{\alpha}$ by the right-hand side of the equation (cf. above) one gets: $\tilde{y} = \bar{y} + \hat{\beta} \cdot (x - \bar{x})$. This equation has the same structure as Equation (4-57).

Notation 4-10: »Prediction« instead of »estimation«:

In the previous presentation the term »prediction« of the latent construct scores instead of the term »estimation« was used.

In addition, instead of using a »hat« as in case of estimates (e.g. $\hat{\eta}$) the »tilde« notation is employed to denote predictions (e.g. $\tilde{\eta}$).

This notation conforms to the general convention that the latent variable scores are *predicted* whereas the values of parameters are *estimated*.

This distinction is due to the fact that constructs are represented as *random variables* whereas parameters represent *fixed values* (at least in classical statistics).

Ex. 4-27 illustrates the method of predicting a latent construct score by means of the least squares predictor.

Ex. 4-27: Least squares prediction of a latent construct score:

Given:

- The model of Figure 4-18 (page 110).
- Assume that the (estimated) mean of the latent construct was $\hat{\mu}_\eta = 100$.
- The estimated means of the tests are assumed to be: $\hat{\mu}_{Y_1} = \hat{\mu}_{Y_2} = \hat{\mu}_{Z_1} = \hat{\mu}_{Z_2} = \hat{\mu}_{Z_3} = \hat{\mu}_{Z_4} = \hat{\mu}_{Z_5} = 100$.
- The observed values of the examinee on the 7 test are: $Y_i = 150$, $(i = 1,2)$, and $Z_j = 90$, $(j = 1,2,\ldots,5)$.
- The estimated covariance matrix of the latent construct is: $\hat{\phi} = 1$.
- The estimated matrix of loading coefficient is: $\hat{\boldsymbol{\Lambda}}^{\mathrm{T}} = [3 \quad 3 \quad 3 \quad 3 \quad 3 \quad 3 \quad 3]$
- The model implied covariance matrix is given by:

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}} = \begin{bmatrix} 15 & 9 & 9 & 9 & 9 & 9 & 9 \\ 9 & 15 & 9 & 9 & 9 & 9 & 9 \\ 9 & 9 & 90 & 9 & 9 & 9 & 9 \\ 9 & 9 & 9 & 90 & 9 & 9 & 9 \\ 9 & 9 & 9 & 9 & 90 & 9 & 9 \\ 9 & 9 & 9 & 9 & 9 & 90 & 9 \\ 9 & 9 & 9 & 9 & 9 & 9 & 90 \end{bmatrix}$$

The least squares (LS) predictor looks like this::

$$\tilde{\eta} = \hat{\mu}_{\eta} + \phi \cdot \hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Sigma}}_{\mathbf{Y}}^{-1} \left(\mathbf{Y} - \hat{\boldsymbol{\mu}}_{\mathbf{Y}} \right)$$

$$= 100 + 1 \cdot \begin{bmatrix} 3 & 3 & 3 & 3 & 3 & 3 & 3 \end{bmatrix} \cdot \begin{bmatrix} 15 & 9 & 9 & 9 & 9 & 9 & 9 \\ 9 & 15 & 9 & 9 & 9 & 9 & 9 \\ 9 & 9 & 90 & 9 & 9 & 9 & 9 \\ 9 & 9 & 9 & 90 & 9 & 9 & 9 \\ 9 & 9 & 9 & 9 & 90 & 9 & 9 \\ 9 & 9 & 9 & 9 & 9 & 90 & 9 \\ 9 & 9 & 9 & 9 & 9 & 9 & 90 \end{bmatrix}^{-1} \cdot \left(\begin{bmatrix} 150 \\ 150 \\ 90 \\ 90 \\ 90 \\ 90 \\ 90 \end{bmatrix} - \begin{bmatrix} 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \end{bmatrix} \right)$$

$$= \underline{\underline{110.6}}$$

Comment: The mean of the 7 test scores is: 107.1, i.e. lower than the least squares prediction.

4.6.2.2 MAXIMUM LIKELIHOOD PREDICTOR OF LATENT CONSTRUCT SCORES

Method 4-9: *Maximum likelihood (ML) predictor of latent construct scores*

The ML predictor is given by the following equation:

$$\tilde{\boldsymbol{\eta}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\eta}} + \left(\hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Theta}}^{-1} \cdot \hat{\boldsymbol{\Lambda}} \right)^{-1} \cdot \hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Theta}}^{-1} \cdot \left(\mathbf{Y} - \hat{\boldsymbol{\mu}}_{\mathbf{Y}} \right). \qquad (4\text{-}58)$$

The symbol $\hat{\boldsymbol{\Theta}}^{-1}$ denotes the inverse of the estimated covariance matrix $\hat{\boldsymbol{\Theta}}$ of the error variables.

Comment: The latent construct scores given by the ML predictor are also called *Bartlett factor scores*.

Ex. 4-28: Maximum likelihood prediction of a latent construct score (continuation of Ex. 4-27):

Given:

The model of Figure 4-18 (page 110) with the data given in Ex. 4-27.

The covariance matrix of the errors is given by:

$$\hat{\boldsymbol{\Theta}} = \begin{bmatrix} 6 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 81 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 81 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 81 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 81 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 81 \end{bmatrix}.$$

The maximum likelihood (ML) predictor is thus given by:

$$\tilde{\eta} = \hat{\mu}_{\eta} + \left(\hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Theta}}^{-1} \cdot \hat{\boldsymbol{\Lambda}}\right)^{-1} \cdot \hat{\boldsymbol{\Lambda}}^{\mathbf{T}} \cdot \hat{\boldsymbol{\Theta}}^{-1} \cdot \left(\mathbf{Y} - \hat{\boldsymbol{\mu}}_{\mathbf{Y}}\right)$$

$$= 100 + \left(\begin{bmatrix} 3 & 3 & 3 & 3 & 3 & 3 & 3 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{81} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{81} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{81} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{81} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{81} \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 3 \\ 3 \\ 3 \\ 3 \\ 3 \\ 3 \end{bmatrix} \right)^{-1}$$

$$\cdot \begin{bmatrix} 3 & 3 & 3 & 3 & 3 & 3 & 3 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{81} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{81} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{81} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{81} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{81} \end{bmatrix} \cdot \left(\begin{bmatrix} 150 \\ 150 \\ 90 \\ 90 \\ 90 \\ 90 \\ 90 \end{bmatrix} - \begin{bmatrix} 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \end{bmatrix} \right)$$

$$= \underline{\underline{113.5}}$$

Since the matrix $\hat{\boldsymbol{\Theta}}$ is a diagonal matrix its inverse $\hat{\boldsymbol{\Theta}}^{-1}$ is a diagonal matrix with the inverse entries on the diagonal of $\hat{\boldsymbol{\Theta}}$:

$$\hat{\boldsymbol{\Theta}}^{-1} = \begin{bmatrix} \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{81} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{81} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{81} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{81} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{81} \end{bmatrix}.$$

Comment: The ML predictor results in a higher latent score than the LS predictor (110.6).

The ML predictor can be expresses by a simple algebraic expression if there is (a) only a single latent construct, and (b) the error variables are uncorrelated. Specifically, in case of the congeneric test model, the matrix equation (4-58) of the predictor simplifies to:

$$\tilde{\eta} = \hat{\mu}_\eta + \frac{\sum_{i=1}^{n}\left[\frac{\hat{\lambda}_i}{\hat{\theta}_i^2}\cdot\left(y_i - \hat{\mu}_{y_i}\right)\right]}{\sum_{i=1}^{n}\frac{\hat{\lambda}_i^2}{\hat{\theta}_i^2}}. \qquad (4\text{-}59)$$

This can be further simplified in case of τ-equivalent tests:

$$\tilde{\eta} = \hat{\mu}_\eta + \frac{\sum_{i=1}^{n}\left[\frac{\left(y_i - \hat{\mu}_{y_i}\right)}{\hat{\theta}_i^2}\right]}{\sum_{i=1}^{n}\frac{1}{\hat{\theta}_i^2}}. \qquad (4\text{-}60)$$

Finally, in case of parallel tests we get the simple expression:

$$\tilde{\eta} = \hat{\mu}_\eta + \frac{\sum_{i=1}^{n}\left(y_i - \hat{\mu}_{y_i}\right)}{n}. \qquad (4\text{-}61)$$

If the latent mean is fixed to zero $\left(\hat{\mu}_\eta = 0\right)$ then the first term on the right-hand side of Equations (4-59) to (4-61) can be dropped. Note that the latent constructs have no internal location. Thus the location parameter, i.e. the mean, has to be fixed to a specific value. For intelli-

gence tests it is common practice to assume that the mean of the latent construct in the population has the value 100.

4.6.3 Comparison of Different Populations

The comparison of different groups (or populations) is concerned with issue of differences between the groups on the latent constructs. It is important that the comparison concerns the *latent constructs* and not the observed *measurements*. Typical questions concerning differences between populations with respect to latent traits are:

»Do women have a higher disposition to depression than men?«;
»Is the power motive higher in men than in women?«;
»Does the general intelligence differ between Blacks and Whites?«.

This raises the following question:

Issue 4-1:
Which conditions enable the inference of differences or equality of latent construct scores for different groups on the basis of the observed test scores?

Let us specify, as a first step, two necessary conditions that enable an inference from the observed test scores to the latent construct scores:

1. The constructs to be measured are equivalent in both groups.
2. The tests are measuring the latent constructs in both groups in the same way.

The significance of these two conditions is obvious: First, if the latent constructs differ between groups (e.g. they comprise different facets) it does not make much sense to compare the groups on the latent constructs. Second, if the tests measure the constructs in a different way in the two groups (= lack of measurement equivalence) the inference from the test scores to latent construct scores seems unjustified.
Missing measurement equivalence can have different reasons. For instance, test items may be interpreted differently in different groups. In case of rating scales as measurement instruments participants of different groups may use different criteria for selecting a response category.
The preceding discussion illustrates that the assessment of differences between different groups require specific considerations concerning (a) the structure of the latent constructs, and (b) the relationship between latent constructs and the measurements in the different groups.
In order to elucidate the issue further let us, first, discuss the problems that encounters a naïve approach.

4.6.3.1 Inference of Group Differences on the Basis of Observed Scores

The simplest method for the comparison of groups consists in comparing observed test scores. Thus inferences of possible differences are

made on the basis of the observed scores without taking the latent variable structure into account.

Ex. 4-29: Assessment of group differences on the basis of observed test scores:

Given:

- The model of Figure 4-33.
- Two groups (say, man and women);
- For each person in each group two measures are taken: $\mathbf{y}^{(1)} = \left[y_1^{(1)}, y_2^{(1)} \right]^{\mathbf{T}}$ and $\mathbf{y}^{(2)} = \left[y_1^{(2)}, y_2^{(2)} \right]^{\mathbf{T}}$.
- $\hat{\boldsymbol{\Sigma}}^{(1)}$ and $\hat{\boldsymbol{\Sigma}}^{(2)}$ denote the model implied covariance matrices of the measures for the two groups.
- $\bar{\mathbf{y}}^{(1)}$ and $\bar{\mathbf{y}}^{(2)}$ represent the model implied mean vectors of the two groups.

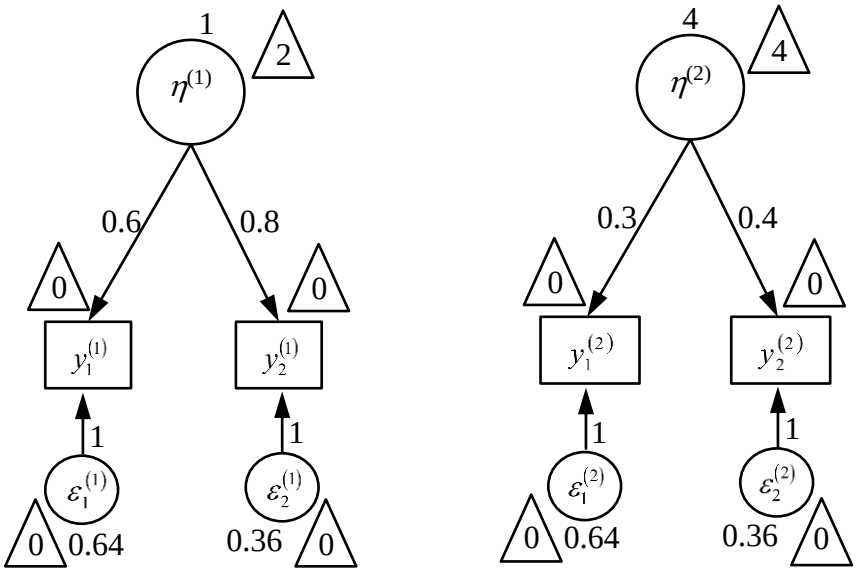


Figure 4-33: *Models for two groups resulting in the same observed covariance and mean structure with differing latent covariance and mean structure for the two groups.*

Question:

Is it possible to draw valid inferences about the mean and covariance structure of the latent constructs on the basis of the observed means and covariances?

Answer:

No, this is impossible as demonstrated by the two models of Figure 4-33:

The mean of the latent construct in Group 2 is double the size of that in Group 1: $\mu_{\eta^{(1)}} = 2$ and $\mu_{\eta^{(2)}} = 4$. However, the model implied means are identical: $\bar{\mathbf{y}}^{(1)} = \bar{\mathbf{y}}^{(2)} = \begin{bmatrix} 1.2 & 1.6 \end{bmatrix}^{\mathbf{T}}$. The same is true for the model implied covariance matrices:

$$\hat{\boldsymbol{\Sigma}}^{(1)} = \hat{\boldsymbol{\Sigma}}^{(2)} = \begin{bmatrix} 1.00 & 0.48 \\ 0.48 & 1.00 \end{bmatrix}.$$

This example makes clear that a naïve approach that takes only observed covariances and means into account can lead to incorrect conclusions with respect to the latent traits. This is a further demonstration of the importance of modern psychometrics that considers the complete measurement model.

4.6.3.2 FACTORIAL INVARIANCE

In the following, conditions are specified that permit a sound conclusion with respect to the equality and differences of latent construct values.

We first consider two different parts of a linear measurement model.

Concept 4-21: *Measurement and structural part of a linear measurement model:*

The *measurement part* of a linear measurement model comprises the following three components of the measurement model:

1. The matrix $\boldsymbol{\Lambda}$ of the factor loadings.
2. The covariance matrix $\boldsymbol{\Theta}$ of the errors.
3. The vector $\boldsymbol{\alpha}$ of intercepts (regression constants).

The *structural part* of a linear measurement model comprises the following two components of the measurement model:

1. The covariance matrix $\boldsymbol{\Phi}$ of the latent constructs.
2. The vector $\boldsymbol{\mu}_{\boldsymbol{\eta}}$ of the means of the latent constructs.

Comment:

The *measurement part* is concerned with those model components that are related to the measurement of the latent constructs. By contrast, the *structural part* concerns the components that are gauged by the test.

The following concepts are of fundamental importance for the specification of conditions that enable valid group comparisons with respect to the latent constructs on the basis of the observed measures.

Concept 4-22: *Strict and strong factorial invariance (Meredith, 1993):*

Strict factorial invariance is given if the following equalities with respect to the measurement part hold:

- The loading matrices are the same for the *G* populations: $\boldsymbol{\Lambda}^{(1)} = \boldsymbol{\Lambda}^{(2)} = \cdots = \boldsymbol{\Lambda}^{(G)}$.
- The vector of intercepts is identical for the *G* groups: $\boldsymbol{\alpha}^{(1)} = \boldsymbol{\alpha}^{(2)} = \cdots = \boldsymbol{\alpha}^{(G)}$.
- The covariance matrix of the errors is the same for the *G* groups: $\boldsymbol{\Theta}^{(1)} = \boldsymbol{\Theta}^{(2)} = \cdots = \boldsymbol{\Theta}^{(G)}$.

Strong factorial invariance is given if the first two equalities (identical factor loadings and identical intercepts) hold.

Comment:

We assume that the means of the errors are zero in all groups:

$$\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}^{(1)} = \boldsymbol{\mu}_{\boldsymbol{\varepsilon}}^{(2)} = \cdots = \boldsymbol{\mu}_{\boldsymbol{\varepsilon}}^{(G)} = \mathbf{0}.$$

The following principle specifies sufficient conditions for a sound conclusion from the observed mean structure on the means of the latent constructs for different groups.

Principle 4-4: *Factorial invariance and valid comparison of group means:*

If strict or strong factorial invariance is present a conclusion from the observed means on the means of the underlying constructs for the different groups is justified. Specifically, observed differences between groups indicate differences on the underlying latent traits.

Justification:

The observed means are given by the system of linear equations(Note that in case of the model being correct, the observed means are equal to the model implied ones, except for sampling errors):

$$\boldsymbol{\mu}_{\mathbf{Y}} = \boldsymbol{\alpha} + \boldsymbol{\Lambda} \cdot \boldsymbol{\mu}_{\boldsymbol{\eta}} + \boldsymbol{\mu}_{\boldsymbol{\varepsilon}}$$

In case of strict or strong factorial invariance the vectors $\boldsymbol{\alpha}$ of intercepts as well as the loading matrices $\boldsymbol{\Lambda}$ are identical for all groups (and the mean vectors of the errors are zero in all groups). Consequently, observed differences of means for different groups can be attributed unambigously to differences between the latent means.

Ex. 4-30: Assessment of group differences on the basis of observed test scores (Continuation of Ex. 4-29):

For the model in Figure 4-33 strict as well as strong factorial invariance is violated since the loading matrices of the two groups are different.

In the context of determining differences of latent means between different populations the distinction between strict and strong invariance is irrelevant. However, the latter distinction is relevant in case of con-

sidering the reliability of the tests as well as in case of conclusions concerning the latent covariance structure on the basis of the observed covariances.
The reliability of the test items is a function of their error variances. In addition the reliability of the sum of test scores also depends on the covariance between errors (cf. Chapter 4.4). By consequence, in case of strong but not strict factorial invariance being present the test may measure the latent construct with different reliability. Thus the precision of the tests as measures of latent traits may be distinct in different groups.
The following principle clarifies the relationship between strict factorial invariance and conclusions concerning the latent covariance structure.

Principle 4-5: *Strict factorial invariance and valid compare-sons of group means and covaraince*

If strict factorial invariance is present a conclusion from the observed means and covariances on the mean and coviarance structure of the underlying constructs for the different groups is justified.

Justification:

The observed covariances are given by the system of linear equations (Note that in case of the model being correct, the observed covariances are equal to the model implied ones, except for sampling errors):

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = \boldsymbol{\Lambda} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\Lambda}^{\mathbf{T}} + \boldsymbol{\Theta} \qquad (4\text{-}62)$$

In case of strict factorial invariance the loading matrices $\boldsymbol{\Lambda}$ as well as the covariance matrix $\boldsymbol{\Theta}$ of the errors are identical for each group. Consequently, observed differences of covariances for different groups can only be due to differences of the covariance matrix $\boldsymbol{\Phi}$ of the latent constructs.

The justification with respect to the means has already been given in Principle 4-4.

Obviously, if only strong but not strict factorial invariance is present, i.e. the covariance matrices of the errors differ in different populations) it is impossible to infer the equality of the covariance structures of the latent constructs from the equality of the observed covariances. This is due to the fact that the model implied covariances are a function of both the latent covariances structure as well as the error covariance structure (cf. Equation 4-62).
Unfortunately, strict and strong factorial invariance constitute ideal cases that are only rarely found in practical applications. This led to the discussion of covariance and mean structures where partial factorial invariance is present only (Byrne, Shavelson, & Muthén, 1989).

4.6.3.3 PARTIAL FACTORIAL INVARIANCE

In case of partial factorial invariance at least one loading coefficient, other than the coefficient used for scaling the latent constructs, as well as the intercept parameter of the associated test are identical in the different groups.

Ex. 4-31 illustrates the problems of interpreting differences between groups in case of partial factorial invariance.

Ex. 4-31: Assessment of group differences in case of partial factorial invariance

Given:

- Two groups: women and men (cf. Figure 4-34);
- Only partial factorial invariance is given since the intercepts of the measures Z_4 und Z_5 differ between the two groups.

A comparison of the two groups with respect to the latent construct η_1 is unproblematic since for this construct we have strict factorial invariance. Consequently, on the basis of the observed identical means it can safely be concluded that the two groups do not differ on the latent trait η_1.

Concerning the latent construct η_2 the situation is different. The identity of the latent means requires different intercepts for the measures Z_4 und Z_5.

Since a difference between groups has been found for two of the five measures the conclusion that both groups do not differ with respect to the latent means of η_2 is problematic.

In case of many indicators with deviations from strong factorial invariance for only very few of them conclusions about equality or differences with respect to the latent construct may be to be justified. In this case one can eliminate the measures exhibiting violations of factorial invariance.

It is, however, useful to further investigate the reasons for the violations of factorial invariance.

The considerations formulated in the context of Ex. 4-31 with respect to conclusions about group differences in case of partial factorial invariance may be summarized as follows:

1. With respect to latent constructs with strong factorial invariance being present conclusions from observed means on latent means are unproblematic even if there does not exist factorial invariance for the other constructs in the model.
2. In case of latent constructs with only partial invariance being present an investigation of possible reasons underlying the observed deviations from factorial invariance seems to be useful. These deviations might indicate the presence of latent factors (not included

into the model) that have a different impact on the measures in different groups.

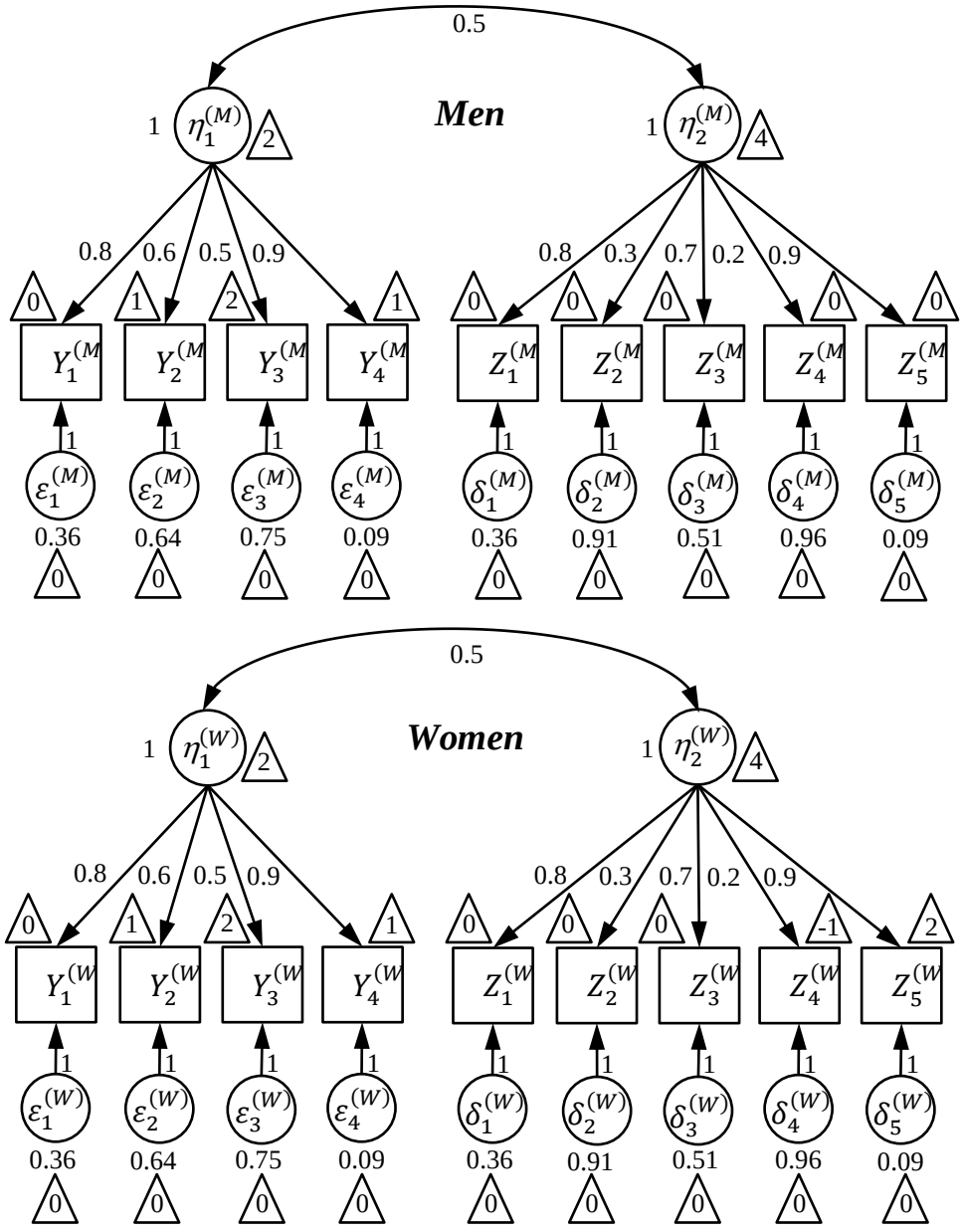


Figure 4-34: *Models of two groups where partial factorial invariance holds.*

4.6.3.4 Conclusion: Group comparisons and factorial invariance

The considerations concerning group comparisons with respect to the latent constructs may be summarized as follows:

1. The equality of or the difference between observed test scores for different groups does, in general, not allow for conclusions with respect to the equality of difference of latent constructs scores.
2. Different loading coefficients indicate that the constructs are measured differentially in the different groups. In this case conclusions from measures on latent constructs are not valid.
3. Strong factorial invariance provides a sufficient condition for sound conclusion concerning the mean structure. Moreover, strict factorial invariance is a sufficient condition for valid inferences with respect to the latent mean and covariance structure.
4. In case of partial invariance limitations with respect to inferences from observed to latent scores concern only those constructs for which factorial does not hold. If violations of factorial invariance concerns only few out of a set of measures one can take into consideration to eliminate those measures for which factorial invariance does not hold.
5. It is instructive to investigate possible reasons for violations of factorial invariance. Differing loading coefficients might indicate that the test items are treated differently in different groups. Differing intercepts and error variances might be an indication that the tests are influenced by latent constructs, not considered so far, that influence the measures differently in the various groups.

4.7 Exercises to Chapter 4

Exercise 4-1: *Computation of the covariance matrix and of variances, covariances, and correlations*

Given: 16 measures of 250 boxes (Excel file: *Data.xlsx*, Sheet: *Boxes*):

X, Y, Z, $\log X$, $\log Y$, $\log Z$, X^2, Y^2, Z^2, $\sqrt{X}$ $\sqrt{Y}$, $\sqrt{Z}$, $X \cdot Y$, $X \cdot Z$, $Y \cdot Z$, $X \cdot Y \cdot Z$

1. Using R, compute the covariance matrix of 16 measures. Then compute on the basis of the computed covariance matrix the following quantities:
   (a) $\mathrm{Var}(X + Y + Z)$
   (b) $\mathrm{Var}(\log X + \log Y + \log Z)$
   (c) $\mathrm{Cov}(X + Y + Z, \log X + \log Y + \log Z)$
   (d) $\mathrm{Corr}(X + Y + Z, \log X + \log Y + \log Z)$
2. Demonstrate the correctness of the computations in (1) by taking the sums of the variables and computing the variances, covariance, and the correlation of the sum variables.

Exercise 4-2: *Computation of covariance matrices within the classical test models*

Given: The test scores $Y_1, Y_2, \ldots, Y_5$ on 5 tests with the true score variables $\tau_1, \tau_2, \ldots, \tau_5$ and error variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_5$.

1. Compute the covariance matrix of the true scores as well as the model implied covariance matrix for the congeneric test model using the following parameters:
   $\lambda_2 = 0.7, \quad \lambda_3 = 0.5, \quad \lambda_4 = 0.8, \quad \lambda_5 = 1.2$
   $\sigma_\tau^2 = 1.5 \quad \left[\sigma_\tau^2 = \sigma_{\tau_1}^2\right]$
   $\sigma_{\varepsilon_1}^2 = 0.8, \quad \sigma_{\varepsilon_2}^2 = 1.3, \quad \sigma_{\varepsilon_3}^2 = 0.7, \quad \sigma_{\varepsilon_4}^2 = 0.4, \quad \sigma_{\varepsilon_5}^2 = 2.1$
2. Compute the covariance matrix of the true scores as well as the model implied covariance matrix for the τ-equivalent test model using the parameters shown above.
3. Compute the covariance matrix of the true scores as well as the model implied covariance matrix for the parallel test model using the parameters shown above, however with the following error variances:
   $\sigma_{\varepsilon_1}^2 = \sigma_{\varepsilon_2}^2 = \sigma_{\varepsilon_3}^2 = \sigma_{\varepsilon_4}^2 = \sigma_{\varepsilon_5}^2 = 0.9$

In order to perform the computations, use matrices (and a program for handling matrices).

Exercise 4-3: *Covariance structure of the congeneric model*

Use covariance algebra to derive the covariance structure of tests for the congeneric model from the linear relationship between true scores.

Exercise 4-4: *Determination of the parameters for the model of congeneric tests*

Given:
The CTT model of three congeneric tests (Figure 4-35):
Determine the estimators (i.e. formulas that represent the parameters as a function of observed variances and covariances) of the parameters.

Hint:
First, use the observed covariances $\mathrm{Cov}\left(Y_i, Y_j\right)$ for determining the loading coefficients λ_i. Second, derive the expressions for the error variances $\theta_{\varepsilon_i}^2$ using the previously determined expressions of λ_i.

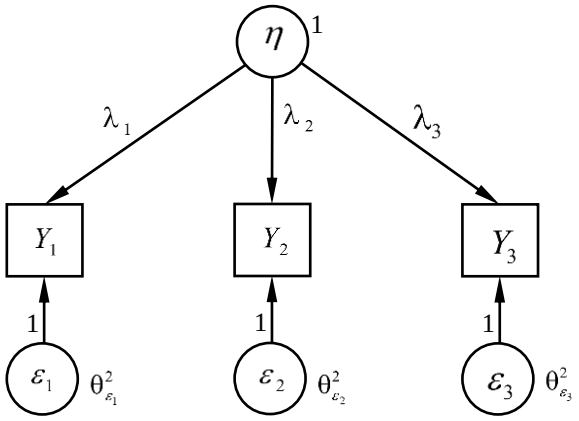


Figure 4-35: *The test model of three congeneric tests.*

Exercise 4-5: *List the parameters of the linear structural equation (LISREL) model and determine the degrees of freedom of the distribution of the test statistics.*

Given:

The LISREL model depicted in Figure 4-36. The variables *F*, *S* and *R* represent the target constructs to be measured. The variables *D*, *K* and *M* denote different methods used for measuring the concepts.

1. List the free model parameters for the different categories:
   (i) Variance parameters of latent constructs;
   (ii) Covariances between latent constructs;
   (iii) Error variances;
   (iv) Loading coefficients.
2. How many free parameters does the model comprise?
3. How many free data points have are modeled?
4. What is the number of degrees of freedom of the distribution of the test statistic?

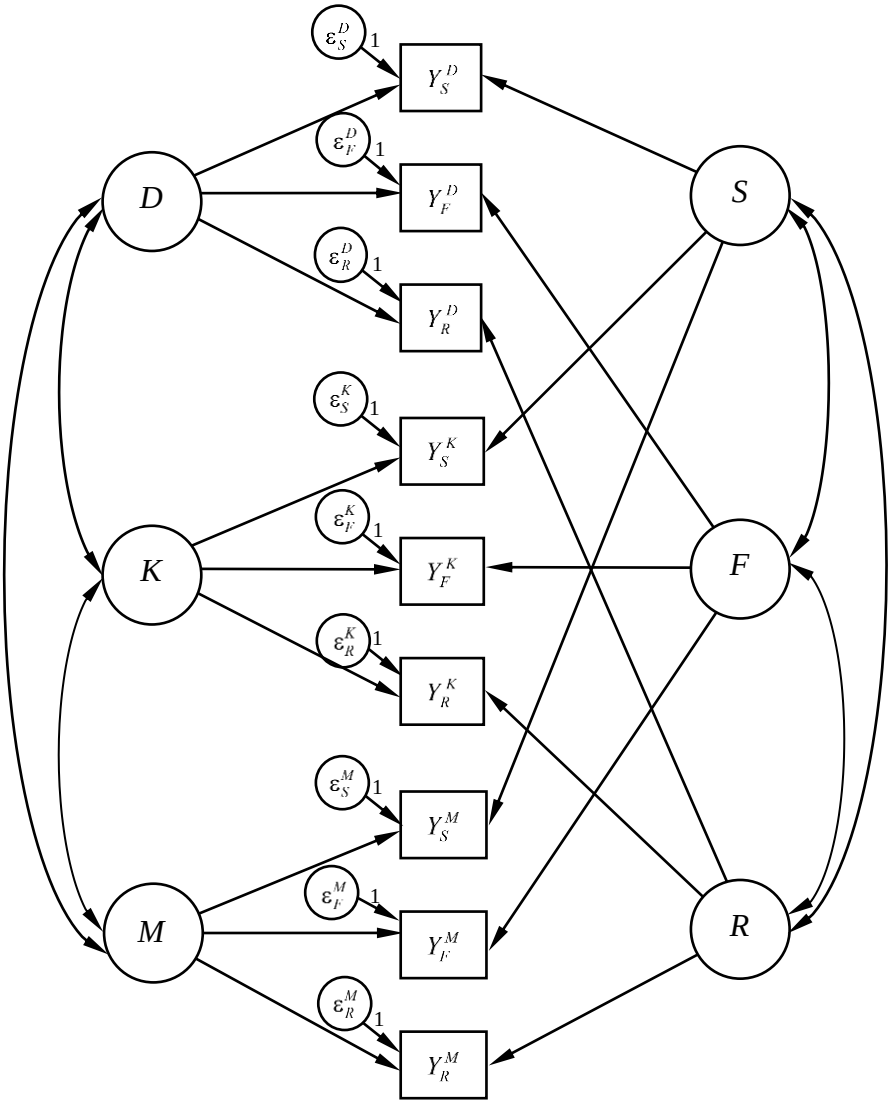


Figure 4-36: *Multitrait–Multimethod-Model.*

Exercise 4-6: *Derivation of the implied covariance matrix of the classical test models*

1. Derive the implied covariance matrix of the LISREL for the general test model (cf. Figure 4-4 on page 73) using covariance algebra.
2. Derive the implied covariance matrix of the LISREL model of 4 congeneric tests (cf. Figure 4-5 on page 74) using covariance algebra.
3. Derive the implied covariance matrix of the LISREL model of 4 congeneric tests (cf. Figure 4-6 on page 76) using covariance algebra.

Exercise 4-7: *Anxiety as a situational factor (Steyer, 1989; Steyer & Eid, 1993)*

Given:

A test for measuring anxiety at two time points with a delay of two month between the two time points.

The test has been divided into two halves with 10 test items in each half. The observed test scores are the sum of the values of the 10 items of each half.

Tab. 4-8 contains the covariance matrix of the test scores from the two test halves at the two time points: S_iA_j = test half j at time point i $(i, j = 1, 2)$.

The sample size was $N = 179$.

Test the following assumptions of Steyer (1989):

1. The two test halves measured at the same time points are parallel (for both time points).
2. Since anxiety is a situational variable and, thus, not constant over time the four measures are not congeneric.

Tab. 4-8: *Covariance matrix of the test scores from two test halves of a test of anxiety that has been applied at two time points: S_iA_j = test half j at time point i (i, j = 1, 2).*

	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

Exercise 4-8: *Test of different test models according to Jöreskog (1971)*

Test the 4 Hypotheses H_1 – H_4 of Ex. 4-5 (page 77) using the covariance matrix given there.

Which of the four hypotheses would you prefer? Justify your judgment.

Exercise 4-9: *Testing hypotheses concerning the covariance structure of a number of tests*

Given:

The covariance matrix of 7 tests: X_1, X_2, X_3, Y_1, Y_2, Z_1, and Z_2. (Tab. 4-9)

The sample size is $N = 350$.

Investigate the subsequent 6 hypotheses concerning the covariance structure of the 7 tests:

Tab. 4-9: *Covariance matrix of the test scores from 7 tests.*

	X_1	X_2	X_3	Y_1	Y_2	Z_1	Z_2
X_1	4.535	1.819	1.576	3.428	2.930	1.219	3.807
X_2	1.819	4.560	1.827	3.320	3.116	1.327	3.792
X_3	1.576	1.827	4.855	3.186	3.052	1.509	3.962
Y_1	3.428	3.320	3.186	13.881	10.490	2.826	7.849
Y_2	2.930	3.116	3.052	10.490	16.514	2.122	6.763
Z_1	1.219	1.327	1.509	2.826	2.122	3.212	3.881
Z_2	3.807	3.792	3.962	7.849	6.763	3.881	15.605

H_1: X_1, X_2 and X_3, as well as Y_1 and Y_2, as well as Z_1 and Z_2 are each congeneric. The 3 groups of variables not congeneric however.

H_2: The 7 tests are congeneric.

H_3: X_1, X_2 and X_3 are parallel. Y_1 and Y_2, as well as Z_1 and Z_2 are each congeneric. The 3 groups of variables not congeneric however.

H_4: X_1, X_2 and X_3 are parallel, Y_1 and Y_2, are τ-equivalent. Z_1 and Z_2 are congeneric. The 3 groups of variables not congeneric however.

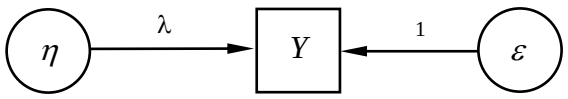
H_5: X_1, X_2 and X_3 are parallel. Y_1 and Y_2, as well as Z_1 and Z_2 are τ-equivalent. The 3 groups of variables not congeneric however.

H_6: X_1, X_2 and X_3, as well as Y_1 and Y_2 are parallel. Z_1 and Z_2 are congeneric. The 3 groups of variables not congeneric however.

Which of the 6 hypotheses would you prefer? Justify your judgment.

Exercise 4-10: *Reliability and the correlation between true score and observed score:*

Given: A simple measurement model:



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

Show that the squared correlation $R^2_{Y.\eta}$ between the observed scores Y and the latent construct η conforms to the reliability of Y:

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

Exercise 4-11: *Error variance and reliability:*

Given:

The linear measurement model of Figure 4-10 on page 86:

Demonstrate the validity of the equation:

$$\mathrm{Var}(\varepsilon) = \mathrm{Var}(Y) \cdot \left[1 - \mathrm{Rel}(Y)\right].$$

Exercise 4-12: *Reliability of a test with two latent variables:*
Given: The model depicted in Figure 4-37:

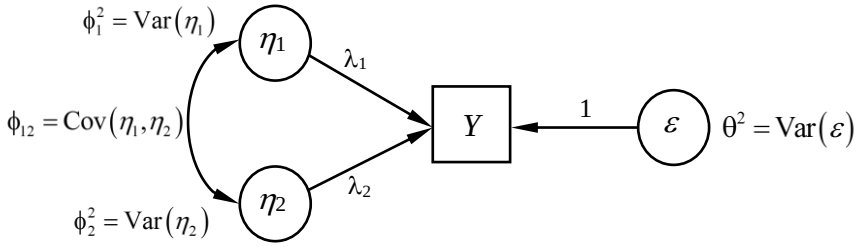


Figure 4-37: *Causal diagram of a measurement model with two latent constructs.*

1. Determine the estimator of $\mathrm{Rel}(Y)$ as a function of the model parameters.

   Hint: Use the structural equations as well as covariance algebra to compute the true score and the total variance of *Y*, predicted by the model.

2. Compute the reliability employing the matrix formula of Method 4-2 on page 85 to compute the reliability of test *Y*, using the data shown in Figure 4-38.
   Compare the result with the one ensuing from application of the estimator developed in the first part of the exercise (Clearly, both methods should lead to the same result).
3. Use a structural equation program with the data of Figure 4-38 to verify the results of your computations

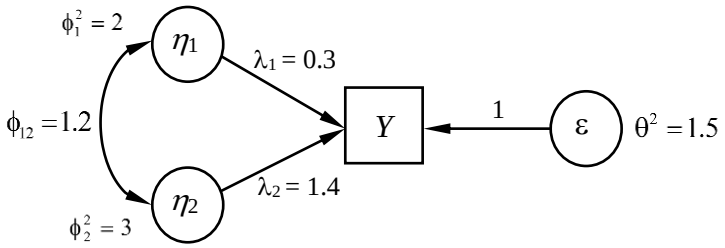


Figure 4-38: *Causal diagram of a measurement model comprising two latent constructs with numerical values for the model parameters.*

Exercise 4-13: *With parallel tests the correlation of the tests corresponds to the reliability of the two tests:*
Given: The measurement model of two parallel tests (Figure 4-39).

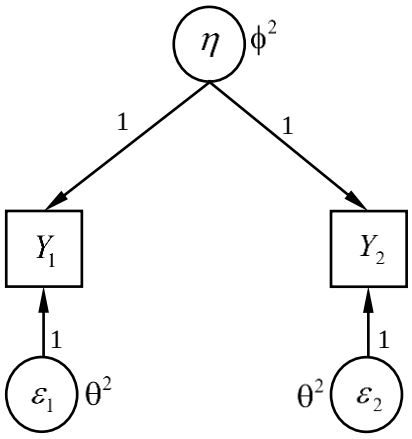


Figure 4-39: *Measurement model of two parallel tests.*

Demonstrate the validity of the following relationships:

$$\mathrm{Rel}(Y_1) = \mathrm{Rel}(Y_2) = \mathrm{Corr}(Y_1, Y_2).$$

Thus, the reliability of two parallel tests corresponds to the correlation of the two tests.

Hint:

- $\mathrm{Corr}(Y_1, Y_2) = \dfrac{\mathrm{Cov}(Y_1, Y_2)}{\sqrt{\mathrm{Var}(Y_1) \cdot \mathrm{Var}(Y_2)}}$.
- Compute the quantities on the right-hand side of the equation using the model parameters and substitute the resulting terms into the equation.
- Show that the resulting expression of the correlation corresponds to the expression of the reliability of Y_1 and Y_2.

Exercise 4-14: *Computation of the reliability of two parallel tests:*

Given: The values of 10 examinees on two parallel tests from Lord and Novick (1968) [Tab. 4-10].

Lord & Novick (1968) got the following estimates:

3. $\hat{\sigma}^2_\eta = 140.67$ [estimated true score variance]
4. $\hat{\sigma}^2_\varepsilon = 9.90$ [estimated error variance]

Determine:

1. The reliability of the tests using the variances estimated by Lord and Novick (1968).
2. The reliability using the correlation of the two (parallel) tests.
3. Estimate the trues score and error variance using structural equations assuming that test are parallel.
4. Compute the covariance between the two tests. Which remarkable result do you get?

Tab. 4-10: *Results of 10 persons on two parallel tests Y_1 and Y_2 (Lord & Novick, 1968, Table 7.3.1 on page 156).*

	Examinee									
Test	1	2	3	4	5	6	7	8	9	10
Y_1	125	119	109	104	101	98	97	94	90	81
Y_2	120	122	107	108	98	106	96	99	93	87

Exercise 4-15: *The reliability of the sum of 2 and m, respectively, τ-equivalent tests conforms to coefficient α:*

Given: The model of two τ-equivalent tests (Figure 4-40).

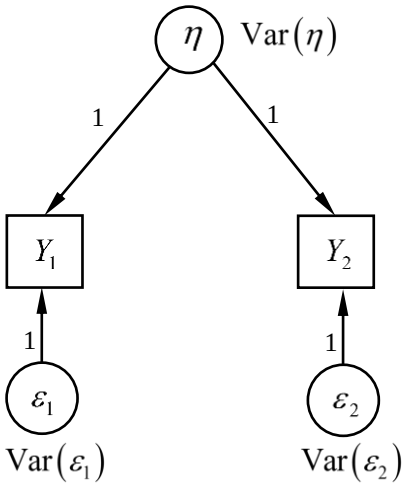


Figure 4-40: *A model of two τ-equivalent tests: Y_1 und Y_2 represent the two tests.*

1. Show that the reliability of $Y = Y_1 + Y_2$ is given correctly by the formula of coefficient α.

Hint: Proceed as follows:

(i) Show that the true score variance of Y is $4 \cdot \mathrm{Var}(\eta)$.

(ii) Show that $\mathrm{Var}(\eta) = \mathrm{Cov}(Y_1, Y_2)$

(iii) Substitute the results in the formula of the reliability:

$$\mathrm{Rel}(Y) = \frac{\text{True score variance}(Y)}{\mathrm{Var}(Y)},$$

and perform the required arithmetic transformations to get the formula for α.

2. Generalize the result to the case of *m* τ-equivalent tests and demonstrate the validity of coefficient α as the coefficient of validity.

Exercise 4-16: *Coefficient α, λ_2, and the reliability of unweighted sums*

Given: The covariance matrix of the test scores for 8 test items $Y_1, Y_2, \ldots, Y_8$ shown in Tab. 4-11 (N = 165).

Tab. 4-11: *Covariance matrix of the test scores for 8 test items.*

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
Y_1	0.249	0.169	0.146	0.190	0.190	0.186	0.168	-0.001
Y_2	0.169	0.251	0.135	0.172	0.172	0.166	0.148	-0.002
Y_3	0.146	0.135	0.243	0.153	0.153	0.145	0.158	0.019
Y_4	0.190	0.172	0.153	0.247	0.198	0.188	0.177	0.014
Y_5	0.190	0.172	0.153	0.198	0.247	0.182	0.177	0.020
Y_6	0.186	0.166	0.145	0.188	0.182	0.250	0.159	0.009
Y_7	0.168	0.148	0.158	0.177	0.177	0.159	0.251	0.016
Y_8	-0.001	-0.002	0.019	0.014	0.020	0.009	0.016	0.250

1. Compute the following quantities using the variances and covariances of the covariance matrix:
   (i) Coefficient α of the sum of the 8 items.
   (ii) Coefficient α of the sum of the 7 items: $Y_1, Y_2, \ldots, Y_7$.
   (iii) Guttman's λ_2 of the sum of the 8 items.
   (iv) Guttman's λ_2 of the sum of the 7 items: $Y_1, Y_2, \ldots, Y_7$.
2. Evaluate the validity of the following statements:
   (i) The 8 test items are congeneric.
   (ii) The 8 test items are τ-equivalent.
   (iii) The 7 test items $Y_1, Y_2, \ldots, Y_7$ are τ-equivalent.
3. Compute the reliability of the sums using the results from the covariance structure analysis of the items:
   (i) The reliability of the sum of the 8 items.
   (ii) The reliability of the sum of the 7 items, $Y_1, Y_2, \ldots, Y_7$, assuming that the 7 items are congeneric.
   (iii) The reliability of the sum of the 7 items, $Y_1, Y_2, \ldots, Y_7$, assuming that the 7 items are τ-equivalent.

Exercise 4-17: *Coefficient α and Guttman's λ_2 are identical, in case of τ-equivalent tests:*

Given: n τ-equivalent tests of a construct.

Show that in this case we have:

$$\alpha = \lambda_2 = \frac{n^2 \cdot \mathrm{Cov}\left(Y_i, Y_j\right)}{\mathrm{Var}\left(Y\right)},$$

i.e., coefficient α is the same as Guttman's λ_2, where:

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

$$\lambda_2 = \frac{\sqrt{\frac{n}{n-1} \cdot \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left[\mathrm{Cov}\left(Y_i, Y_j\right)\right]^2} + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathrm{Cov}\left(Y_i, Y_j\right)}{\mathrm{Var}\left(Y\right)}$$

Hint: In case of τ-equivalent tests the covariances $\mathrm{Cov}\left(Y_i, Y_j\right)$ between any pair of tests, Y_i and Y_j, are the same.

Exercise 4-18: *In case of parallel tests coefficient α corresponds to the Spearman-Brown coefficient:*

Given: n parallel tests:

Show that in this case the formula of coefficient α:

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

corresponds to the Spearman-Brown formula:

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

Hints:

1. Note that, due to the presence of parallel tests, the following relationships are true:
   (i) The correlation $\mathrm{Corr}\left(Y_i, Y_j\right) = \rho$ is the same for each pair Y_i and Y_j $\left(i \neq j\right)$,
   (ii) The variance $\mathrm{Var}\left(Y_i\right) = \sigma^2$ is the same for each Y_i.
2. Count the number of covariances in the nominator and denominator of coefficient α as well as the number of variances in the denominator, and replace the covariance terms $\mathrm{Cov}\left(Y_i, Y_j\right)$ by $\rho \cdot \sigma^2$ and the variance terms $\mathrm{Var}\left(Y_i\right)$ by σ^2.
3. Canceling terms results in the Spearman-Brown formula.

Exercise 4-19: *The Spearman-Brown coefficient overestimates the true reliability in case of correlated errors:*

Given: The model in Figure 4-41:

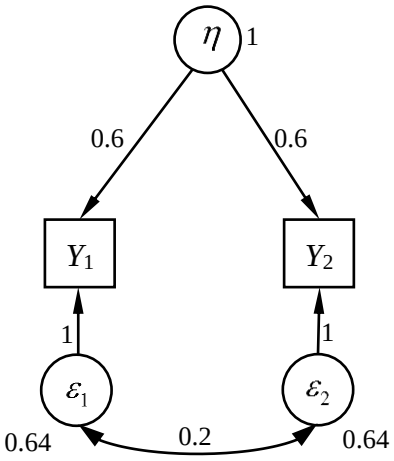


Figure 4-41: *Two tests with correlated errors.*

Note that without the existing covariance between error terms the two tests were parallel. In this case the Spearman-Brown coefficient would be an unbiased estimate of the reliability.
Show that for the given model Spearman-Brown coefficient overestimates the true reliability.
Hint:

1. Compute the true reliability of the sum of the two tests (using covariance algebra or matrices).
2. The correlation $\mathrm{Corr}(Y_1, Y_2)$ can be computed from the model implied covariance matrix (that can be computed by means of matrix algebra using the values shown in Figure 4-41).
3. Use the Spearman-Brown formula to correct the correlation $\mathrm{Corr}(Y_1, Y_2)$ thus getting an estimate of the reliability of sum of the two tests.

Exercise 4-20: *Coefficient α underestimate the reliability of congeneric measures*

Given: n congeneric tests (Model of Figure 4-42):

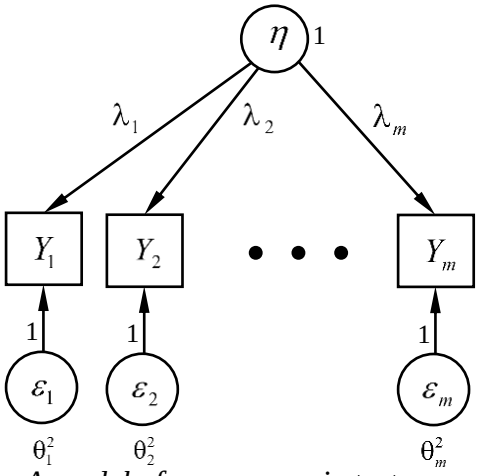


Figure 4-42: *A model of m congeneric tests.*

Let $Y = Y_1 + Y_2 + \cdots + Y_n$ be the sum of the *n* test scores.

Show that coefficient α underestimates the reliability $\mathrm{Rel}(Y)$ of the sum of the *n* test scores.

Hint:

1. Show that

$$\mathrm{Rel}(Y) = \frac{\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_n^2 + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ i \neq j}}^{n} \lambda_i \cdot \lambda_j}{\mathrm{Var}(Y)}$$

Comment:

The double sum comprises $n \cdot (n-1)$ terms.

2. Show that:

$$\alpha = \frac{n}{n-1} \cdot \frac{\sum_{i=1}^{n}\sum_{\substack{j=1 \\ i \neq j}}^{n} \lambda_i \cdot \lambda_j}{\mathrm{Var}(Y)}$$

3. Thus in case of $\mathrm{Rel}(Y) \geq \alpha$ the following inequality hold:

$$\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_m^2 + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ i \neq j}}^{n} \lambda_i \cdot \lambda_j \geq \frac{n}{n-1} \cdot \sum_{i=1}^{n}\sum_{\substack{j=1 \\ i \neq j}}^{n} \lambda_i \cdot \lambda_j$$

or:

$$(n-1) \cdot \left[\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_n^2 + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ i \neq j}}^{n} \lambda_i \cdot \lambda_j \right] \geq n \cdot \sum_{i=1}^{n}\sum_{\substack{j=1 \\ i \neq j}}^{n} \lambda_i \cdot \lambda_j$$

or:

$$(n-1) \cdot \left[\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_n^2 \right] \geq \sum_{i=1}^{n}\sum_{\substack{j=1 \\ i \neq j}}^{n} \lambda_i \cdot \lambda_j$$

4. Demonstrate that this inequality is true.

Hint:
$(\lambda_i - \lambda_j)^2 \geq 0 \Leftrightarrow \lambda_i^2 + \lambda_j^2 \geq 2 \cdot \lambda_i \cdot \lambda_j$, with equality in case of $\lambda_i = \lambda_j$. Otherwise the strict inequality $\lambda_i^2 + \lambda_j^2 > 2 \cdot \lambda_i \cdot \lambda_j$ holds.
Adding the equations for all combinations of λ_i and λ_j,

$$\lambda_1^2 + \lambda_2^2 \geq 2 \cdot \lambda_1 \cdot \lambda_2$$
$$\lambda_1^2 + \lambda_3^2 \geq 2 \cdot \lambda_1 \cdot \lambda_3$$
$$\ldots \quad ,$$
$$\lambda_{n-1}^2 + \lambda_n^2 \geq 2 \cdot \lambda_{n-1} \cdot \lambda_n$$

results in the above inequality:

$$(n-1) \cdot \left[\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_n^2 \right] \geq \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} \lambda_i \cdot \lambda_j$$

Demonstrate the computations using $\lambda_1, \lambda_2, \ldots, \lambda_4$ (4 tests).

Exercise 4-21 *Reliability of the sum of non-congeneric tests:*

Given: The model of Figure 4-17 (Page 108). The loading coefficients are all of the same value: $\lambda = 0.7$.

Compute the reliability of the sum of the 10 tests.

Exercise 4-22: *Reliability of the weighted sum of tests in the general factor analytic model:*

Given: The model of Figure 4-15 on Page 101 (Data: Ex. 4-11, on Page 101).

Compute the reliability of the weighted sum of the 5 tests using the reliabilities of the single test as weights.

Exercise 4-23: *Reliability of the weighted sum of tests:*

Given: The model of Figure 4-18 on Page 110.

Compute the reliability of the weighted sum of the seven tests:
$Z = Y_1 + Y_2 + 0.1 \cdot Z_1 + 0.1 \cdot Z_2 + 0.1 \cdot Z_3 + 0.1 \cdot Z_4 + 0.1 \cdot Z_5$.

Exercise 4-24 *Maximal reliability I: Reliability of unweighted sums vs. the reliability of optimally weighted sums (congeneric tests):*

Given: The model of three congeneric tests (Figure 4-43).

Compute:
1. The reliability $\mathrm{Rel}(Y)$ of the simple sum:

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

2. Die optimal weights w_i that maximize the reliability of the weighted sum.
3. The maximal reliability $\text{Rel}_{\max}$ of the optimally weighted sum, $Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + w_3 \cdot Y_3$, with the optimal weights w_i.

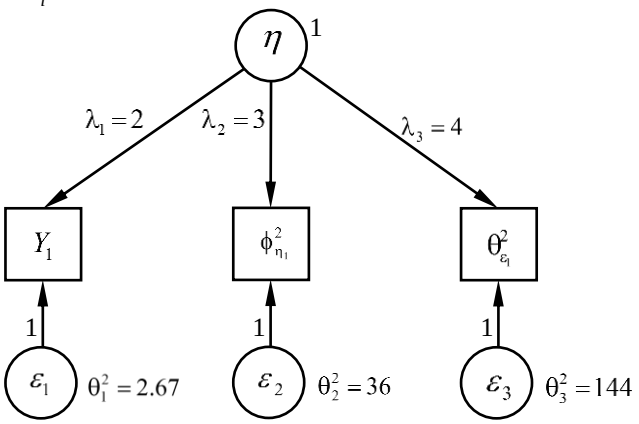


Figure 4-43: *A model of three congeneric tests.*

Exercise 4-25: *Maximal reliability II: Reliability of unweighted sums vs. the reliability of optimally weighted sums (non-congeneric tests):*

Given: The model of three non-congeneric tests of Figure 4-44.

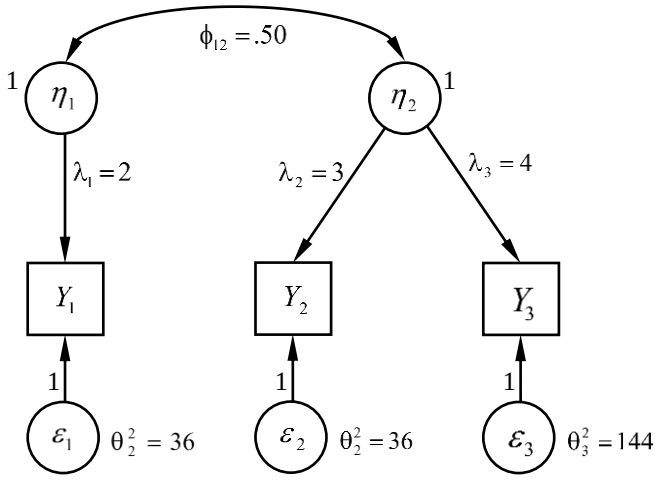


Figure 4-44: *A model of three tests that are not congeneric.*

Compute:

1. The reliability $\theta_1^2 = 2.67$ of the simple sum:

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

2. Die optimal weights $\phi_{\eta_1}^2$ that maximize the reliability of the weighted sum.

3. The maximal reliability $\mathrm{Rel}_{\max}$ of the optimally weighted sum, $Y = w_1 \cdot Y_1 + w_2 \cdot Y_2 + w_3 \cdot Y_3$, with the optimal weights $\phi^2_{\eta_1}$.

Exercise 4-26: *Maximal reliability III:*

Compute the maximal reliability of the model of Figure 4-15 on page 101, as well as the associated optimal weight vector of length 1.

Exercise 4-27: *Unique reliabilities:*

Given: The model of a test with three latent constructs (Figure 4-45).

Compute the unique reliability of the test for each of the three latent constructs.

Comment

The model of Figure 4-45 differs from the one of Figure 4-27 (page 133) only by the higher covariance between the latent constructs. Consequently, the unique reliabilities should by lower than those obtained Ex. 4-25 (page 132).

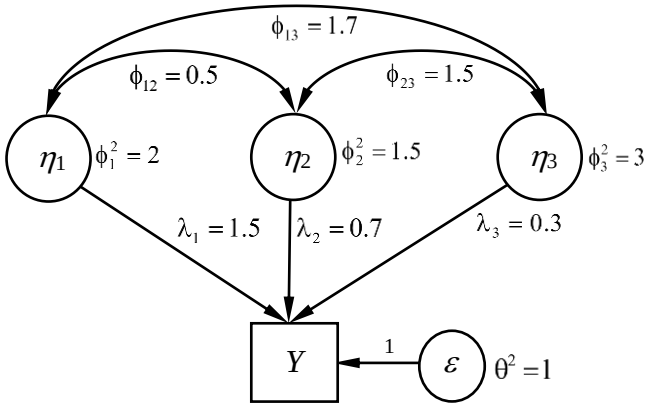


Figure 4-45: *Structural model with three latent constructs and a single indicator.*

Exercise 4-28: *Validity-reliability paradox I*

Given: *n* parallel Tests: $Y_1, Y_2, \ldots, Y_n$;

The criterion *C.*

Show that the correlation $\mathrm{Corr}(C,Y)$ between the criterion *C* and the sum $Y = Y_1 + Y_2 + \cdots + Y_n$ is given by:

$$\mathrm{Corr}(C,Y) = \frac{\sqrt{n} \cdot \mathrm{Corr}(C,Y_i)}{\sqrt{1+(n-1)\cdot \mathrm{Rel}(Y_i)}}.$$

$\mathrm{Corr}(C,Y_i)$ denotes the correlation between the criterion C and the single test Y_i $(i=1,2,\ldots,n)$. Since the n test items are parallel this correlation is the same for all test items.

$\mathrm{Rel}(Y_i)$ denotes the reliability of test Y_i that conforms to the correlation $\mathrm{Corr}(Y_i,Y_j)$ between the single parallel tests Y_i and Y_j (Note that $\mathrm{Rel}(Y_i)$ is the same for each test).

Hints:

1. Since the tests are parallel the all have the same variance.
2. $\mathrm{Cov}(Y_i,Y_j)=\mathrm{Corr}(Y_i,Y_j)\cdot\sqrt{\mathrm{Var}(Y_i)\cdot\mathrm{Var}(Y_j)}$
3. $\mathrm{Rel}(Y_i)=\mathrm{Corr}(Y_i,Y_j)$

Exercise 4-29: *Validity-reliability paradox II:*

Given: The model of Figure 4-46, with n parallel tests, $Y_1,Y_2,\ldots,Y_n$, and a criterion C.

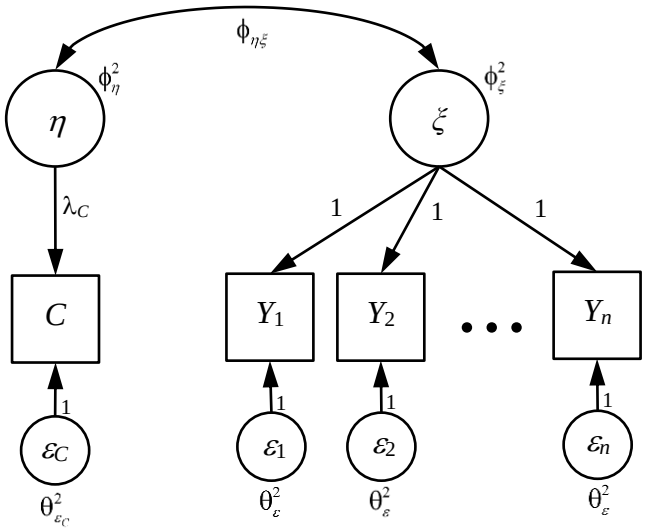


Figure 4-46: *Structural equation model used to illustrate the validity-reliability paradox.*

Show that the reliability $\mathrm{Rel}(Y)$ of the sum of the test scores $Y=Y_1+Y_2+\cdots+Y_n$ is given by:

$$\mathrm{Rel}(Y)=\frac{n\cdot\phi^2_\xi}{n\cdot\phi^2_\xi+\theta^2_\varepsilon}.$$

Show that the correlation $\mathrm{Corr}(C,Y)$ between the criterion C and the sum $Y = Y_1 + Y_2 + \cdots + Y_n$ is given by:

$$\mathrm{Corr}(C,Y) = \frac{\lambda_C \cdot \phi_{\eta\xi}}{\sqrt{\lambda_C^2 \cdot \phi_\eta^2 + \theta_{\varepsilon_C}^2} \cdot \sqrt{\phi_\xi^2 + \frac{\theta_\varepsilon^2}{n}}}$$

Exercise 4-30: *Validity-reliability paradox III:*

Given: The model of Figure 4-47, with n parallel tests, $Y_1, Y_2, \ldots, Y_n$, and criterion C.

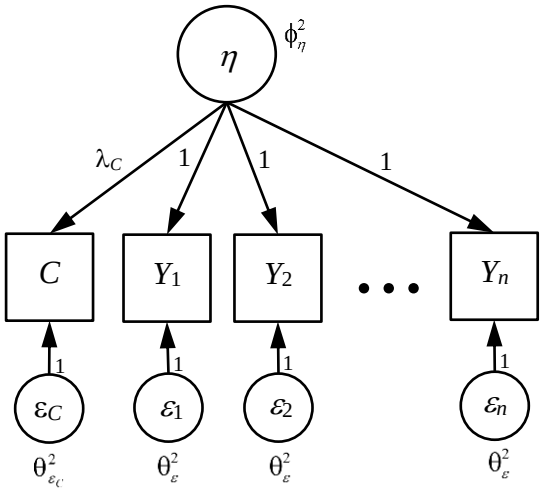


Figure 4-47: *Structural equation model for illustrating the validity-reliability paradox.*

Show that the correlation $\mathrm{Corr}(C,Y)$ between the criterion C and the sum $Y = Y_1 + Y_2 + \cdots + Y_n$ is given by:

$$\mathrm{Corr}(C,Y) = \frac{\lambda_C}{\sqrt{\lambda_C^2 + \frac{\theta_{\varepsilon_C}^2}{\phi_\eta^2}} \cdot \sqrt{1 + \frac{\theta_\varepsilon^2}{n \cdot \phi_\eta^2}}}$$

Exercise 4-31: *Estimating and testing of mean structures:*

Given:

Four measures of the sensitivity parameter d_a of the Gaussian signal detection model (cf. Excel-File: *Data.xlsx*, Sheet: *SDT Data*):

1. SDT6.da: Estimated parameter d_a from new-old recognition using a 6-point rating scale.
2. SDT4.da: Estimated parameter d_a from new-old recognition using a 4-point rating scale.

3. AFC4.da: Estimated parameter d_a from new-old forced choice recognition with repeated choices comprising 4 choice options (on old and 3 new items).
4. AFC3.da: Estimated parameter d_a from new-old forced choice recognition with repeated choices comprising 3 choice options (one old and 2 new items).

Estimate the following models and report the fit statistic G_2 with associated *df* and *p* value, as well as *RMSEA*:

(a) The partial τ-equivalent model assuming equal loadings and intercepts (of the observed variables) for the two estimated parameters resulting from the two rating tasks on the one hand and, on the other hand, the two estimated parameters resulting from the two force choice tasks (all 4 measures are assumed to be congeneric).
(b) The τ-equivalent model, by constraining the intercepts of the 4 measures to be equal.
(c) The strictly parallel model, by constraining the intercepts of the 4 measures to be equal.
(d) The strictly parallel model, by constraining the intercepts of the 4 measures to be zero, and letting, instead, the intercept of the latent construct to be estimated freely.

Which model would you prefer?
Hint:
Model (c) and (d) should result in the same fit indices.

5. Probabilistic Test Theory (PTT)

The following chapter discusses probabilistic test models. These models are used for modeling probabilities of response categories. The term *response category* refers to two different types of events:

1. The *probability of a correct response* (e.g. on a multiple choice test). In this case the response category of interest differs from the response options associated with an item.
2. The *probability of the different response options* available for the item. In this case the response categories of interest correspond to the response options that are available for the participants.

The Birnbaum models presented in Section 5.2 are used for modeling the probabilities of correct responses on different items. The ordered categories response models, Samejima's (1969) graded response model as well as Master's (1982) partial credit model, are used for modeling the distribution of ordered response categories (Section xxxx).
Previously to discussing specific models and methods the relationship between classical and probabilistic test models will be discussed.

5.1 Introduction: Classical and Probabilistic Test Models

Figure 3-1 depicts the general psychometric test model that encompasses the classical and the probabilistic test models as special cases. For convenience the models has been reproduced in Figure 5-1.

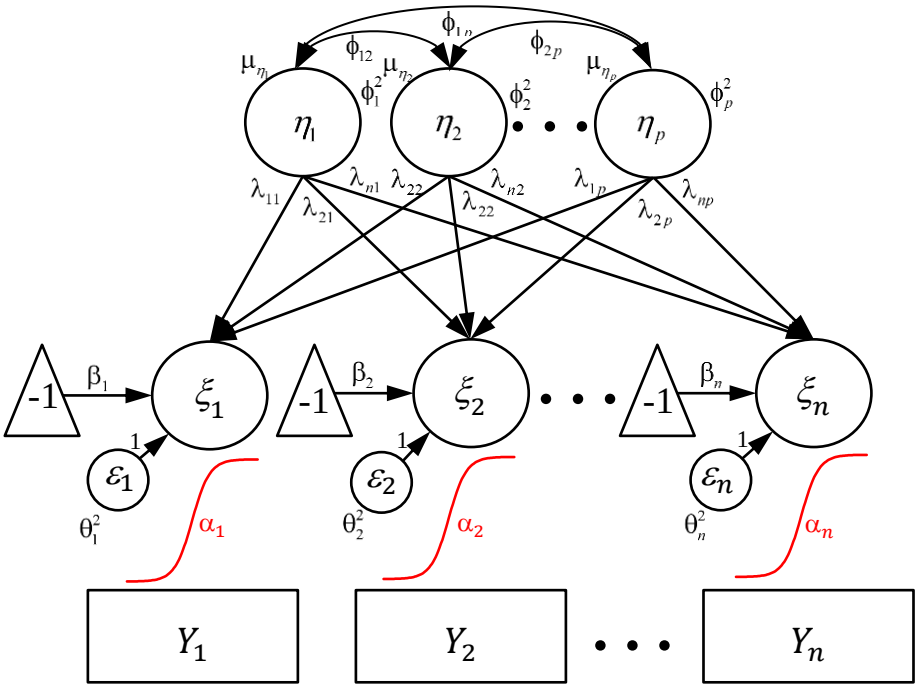


Figure 5-1: *Basic structure of a psychometric model.*

The difference between the linear structural equation model used for modeling the test models of CTT and the PTT models consists in the addition of a non-linear response function for PTT models that maps the value of the latent response processes into the range $[0,1]$ (cf. the sigmoid curves at the bottom of Figure 5-1). Thus both models assume response processes $\xi_1, \xi_2, \ldots, \xi_n$ that are linear function of the exogenous variables. These response processes can be modeled by a set of linear equations:

$$\begin{aligned}
\xi_1 &= \beta_1 + \lambda_{11} \cdot \eta_1 + \lambda_{12} \cdot \eta_2 + \cdots + \lambda_{1p} \cdot \eta_p + \varepsilon_1 \\
\xi_2 &= \beta_2 + \lambda_{21} \cdot \eta_1 + \lambda_{22} \cdot \eta_2 + \cdots + \lambda_{2p} \cdot \eta_p + \varepsilon_2 \\
&\vdots \qquad\qquad \cdots \\
\xi_n &= \beta_n + \lambda_{n1} \cdot \eta_1 + \lambda_{n2} \cdot \eta_2 + \cdots + \lambda_{np} \cdot \eta_p + \varepsilon_n
\end{aligned} \tag{5-1}$$

Notation 5-1: Symbols for denoting intercept parameters:

In this chapter the intercept parameters are denoted by the letter β whereas in Section 4.6 the letter α was used. This change of notation is due to the fact that the letter α is used to denote the slopes of the item response functions.

In case of CTT the response processes $\xi_1, \xi_2, \ldots, \xi_n$ are identical to the observed responses: $Y_1 = \xi_1, Y_2 = \xi_2, \ldots, Y_n = \xi_n$. By contrast, for PTT models the response processes are itself latent variables (hidden responses) that are mapped on the model predicted response probabilities by means of a non-linear function. Consequently the two types of models differ only with respect to the presence of a response function. This is due to the fact that PTT models are used for modeling probabilities of response categories whereas CTT models are used for modeling means and (co-) variances.

5.2 Modeling the Probabilities of Correct Responses: The Birnbaum Models

The Birnbaum models (Birnbaum, 1968) are used for modeling the probabilities in case of two response categories: 1 = correct response, and 0 = wrong response.

The three Birnbaum models are also called the *one-parameter logistic* (*1-PL*), *two-parameter logistic* (*2-PL*) and *three-parameter logistic* (*3-PL*) models. These names are due to the fact that all three models use the logistic response function (cf. Chapter 3):

$$\Psi(\xi) = \frac{\exp(\alpha \cdot \xi)}{1 + \exp(\alpha \cdot \xi)} \qquad (5\text{-}2)$$

In Equation (5-2) the symbols ξ represents the latent response process whereas α denotes the slope of the logistic function (cf. Figure 3-2 on page 28).

Notation 5-2: Using a constant to equate parameters between the logistic and the normal ogive model

As noted in Chapter 3, the constant $D = 1.7$ can be used to adjust the logistic response function to the normal ogive model. By consequence the estimated parameters of both models are similar. In this case the response function has the following form:

$$\Psi(\xi) = \frac{\exp(D \cdot \alpha \cdot \xi)}{1 + \exp(D \cdot \alpha \cdot \xi)}$$

In the subsequent presentation no constant D will be used. This does not provide any restriction with respect to the models capability of fitting the data since the constant can be absorbed into the slope parameter α.

The three models are nested (or embedded) in that the more complex model contain the simpler versions as a special cases that results by fixing one of more parameters (similar to the classical test models). The basic structure of the 1-PL and 2-PL models is depicted in Figure 5-2 (The 3-PL model has a slightly different structure; cf. Figure 5-6 on page 184).

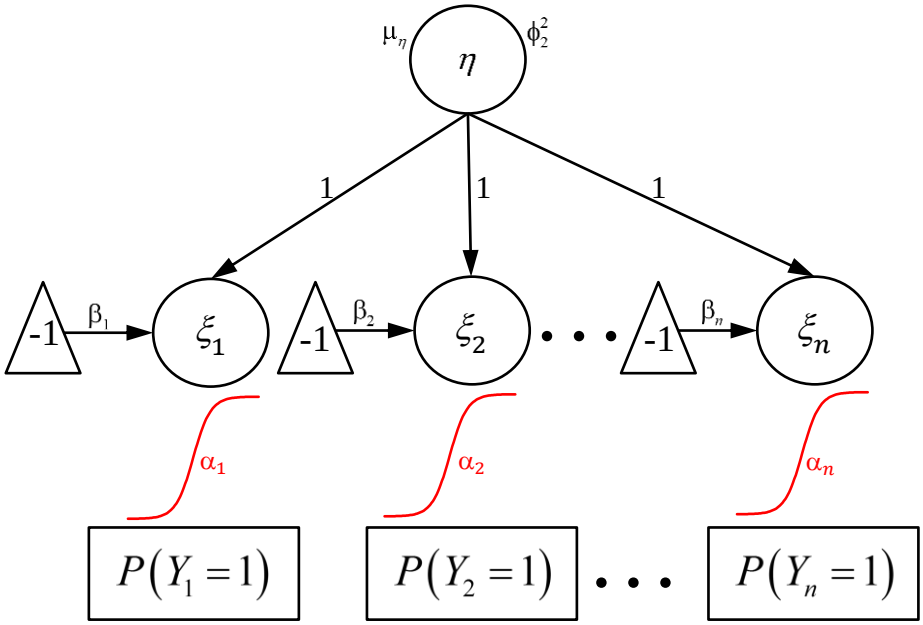


Figure 5-2: *Basic structure of the one- and two-parameter Birnbaum model.*

The model equation describes the probability of a correct response given that the latent variable η takes on value η. The latent variable η is usually interpreted as representing the examinees abilities.

We now turn to a detailed discussion of the three models starting with the most simple model.

5.2.1 The One-Parameter Birnbaum Model: Rasch Model

The one parameter Birnbaum model is also called the *Rasch model*. The model is described by the following model equations:

$$P(Y_i = 1|\eta) = \frac{\exp(\eta - \beta_i)}{1 + \exp(\eta - \beta_i)}$$
$$P(Y_i = 0|\eta) = 1 - P(Y_i = 1|\eta) = \frac{1}{1 + \exp(\eta - \beta_i)} \quad . \qquad (5\text{-}3)$$

The symbols have the following meaning:

η = value of the latent construct (ability) [Person parameter].

β_i = difficulty of item *i* [Item parameter].

$P(Y_i = 1|\eta)$ = the probability of a correct response $(Y_i = 1)$ given a specific value η on the latent construct (ability).

$P(Y_i = 0|\eta)$ = the probability of an incorrect response $(Y_i = 0)$ for item *i* given a specific value η on the latent ability.

Note that the model in Equation (5-3) is but a special case of the model in Equation (5-2) with $\alpha_i = 1$ $(i = 1,\ldots,n)$. The latent response processes are given by (cf. Figure 5-2):

$$\xi_i = \eta - \beta_i \qquad (5\text{-}4)$$

Figure 5-3 depicts the *item characteristic curves* (*ICC*) for five items with the difficulty parameters $\beta_1 = -2$, $\beta_2 = -1$, $\beta_3 = 0$, $\beta_4 = 1$, and $\beta_5 = 2$. The *item characteristic curve* of an item specifies the probability of a correct response to the item as a function of the latent ability variable η. The ICCs of the items conforming to the Rasch model exhibit the following characteristics:

1. The curves all have the same slope. The curves of the single items are thus parallel, i.e. they are shifted versions of the same curve.
2. The effect of the difficulty parameter consist in shifting the curves: Curves representing smaller values of β are located on the left and those with higher value are located on the right. By consequence, for a fixed value of the latent ability, the probability of a correct response decreases with β. This justifies the interpretation of β as a difficulty parameter.
3. By contrast, the probability of a correct response increases with the value of the latent ability variable η.
4. If the value of the latent construct equals the difficulty parameter $(\eta = \beta)$, the probability of a correct response is 0.5 (cf. the grey lines in Figure 5-3).

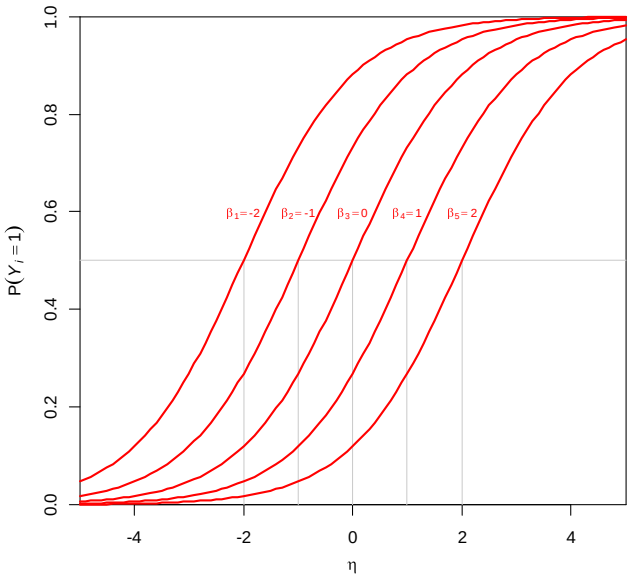


Figure 5-3: *Rasch model: Item characteristic curves for the five items with difficulty parameters β_i.*

The Rasch model incorporates a specific characteristic that renders it particularly attractive. This will be discussed next.

5.2.1.1 LOGIT TRANSFORMATION AND SPECIFIC OBJECTIVITY

The logit transformation is the inverse of the logistic function underlying the Birnbaum models. It is given by the equation

$$\log\left[\frac{P\left(Y_i=1\middle|\eta\right)}{P\left(Y_i=0\middle|\eta\right)}\right]=\eta-\beta_i. \qquad (5\text{-}5)$$

The function log() represents the natural logarithm that is the inverse of the exponential function exp(). Equation (5-5) can be verified easily by replacing the probabilities by the right hand side of Equation (5-3) and simplifying the resulting expression.

Notation 5-3: Logits

The fraction $P\left(Y_i=1\middle|\eta\right)\big/P\left(Y_i=0\middle|\eta\right)$ of the two probabilities is called *odds*. The logarithm of this fraction is called *logarithmic odds* or *log odds*. The term *logit* is simply an abbreviation for logarithmic odds.

The logit transformation thus transforms the probability of a correct response back to the latent response process. By consequence, the logits are linear functions of the latent ability variable as well as of the difficulty parameter. Figure 5-4 depicts the logits for five items with the same difficulty parameter as in Figure 5-3 as a function of the latent ability.

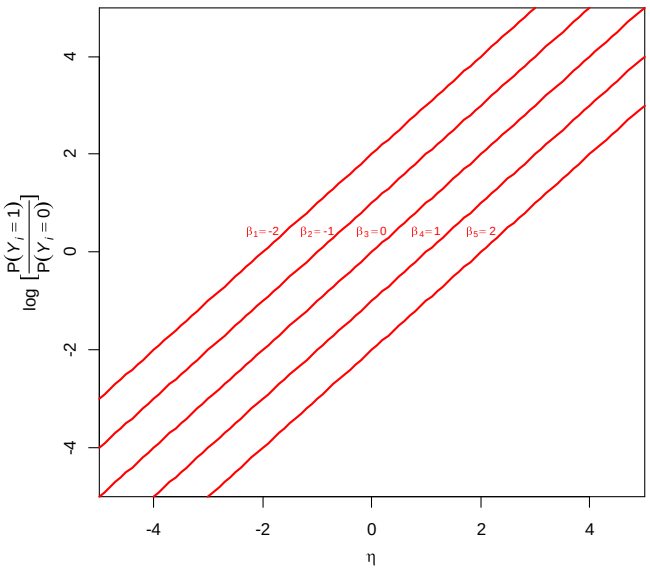


Figure 5-4: *Logit of correct vs. incorrect responses as a function of latent ability for items with different difficulty parameters β_i (Rasch model).*

Using logits reveals a characteristic of the Rasch model that has been termed specific objectivity.

Concept 5-1: *Specific objectivity:*

Specific objectivity in the context of psychometrics comprises two types of invariant comparisons:

(1) Comparisons between persons are invariant with respect to the test items used to measure them.
(2) Comparisons of items are invariant with respect to the examinees used to calibrate them.

This means, that for comparing examinees it does not matter which items are used for the comparison. Similarly, for comparing items it should be irrelevant to which examinees these items have been applied. In each case the outcome has to be the same.

The sort of invariances required by specific objectivity holds on the logit scale. Specifically, differences between logits of different participants tested on the same item does not depend on the item used for comparison:

$$\log\left[\frac{P(Y_i=1|\eta_1)}{P(Y_i=0|\eta_1)}\right]-\log\left[\frac{P(Y_i=1|\eta_2)}{P(Y_i=0|\eta_2)}\right]=(\eta_1-\beta_i)-(\eta_2-\beta_i) \qquad (5\text{-}6)$$

$$=\underline{\underline{\eta_1-\eta_2}}$$

The invariance of the person comparison is evidenced by the fact that the difference of the logits for different examinees with abilities η_1 and η_2 does not depend on the item (difficulty) parameter β_i.

Similarly, the difference between the logits for two different items *i* and *j*, applied to the same participant (or to different participants with the same ability level), does not depend on the value of the ability:

$$\log\left[\frac{P(Y_i=1|\eta)}{P(Y_i=0|\eta)}\right]-\log\left[\frac{P(Y_j=1|\eta)}{P(Y_j=0|\eta)}\right]=(\eta-\beta_i)-(\eta-\beta_j) \qquad (5\text{-}7)$$

$$=\underline{\underline{\beta_j-\beta_i}}$$

The two types of invariances can be read off directly from Figure 5-4:

1. *Invariance of person comparisons:*
   Select two values η_1 and η_2 (on the *x*-axis). Since the curves are parallel lines with identical slopes the difference of the logits (on the *y*-axis) is the same, independently of which item curve is chosen.
2. *Invariance of item comparisons:*
   Select two curves (straight lines). It becomes immediately clear that the vertical separation between the curves is always the same, independently of the level of ability (= the value on the *x*-axis) chosen. The difference depends only on the separation between the two curves. This is due to the fact that the curves are straight lines having the same slope.

According to the Rasch model item difficulties and persons' abilities can be placed on the same latent scale. Consequently, the Rasch model enables a comparison of a person's ability with the difficulty of various test items or with a standard of comparison. Thus, statements are possible about how far the person is located above or below the standard of comparison.

Due to the characteristic of specific objectivity and the possibility of comparing item characteristics (difficulties) with persons' abilities, a set of test items that conform to the Rasch model represents an ideal measurement instrument for measuring the latent ability construct. This characteristic is comparable to that of parallel items in CTT.

Unfortunately most tests do not conform to the Rasch model, and, thus, require more complex measurement models. The two parameter Birnbaum model constitutes a generalization of the Rasch model that enable a more general application.

5.2.2 The Two-Parameter Birnbaum Model (2-PL)

The 2-PL model extends the Rasch model by including an item specific *discrimination parameter* α_i. The resulting model equation is given in Equation (5-8):

$$P(Y_i = 1|\eta) = \frac{\exp[\alpha_i \cdot (\eta - \beta_i)]}{1 + \exp[\alpha_i \cdot (\eta - \beta_i)]}$$
$$P(Y_i = 0|\eta) = 1 - P(Y_i = 1|\eta) = \frac{1}{1 + \exp[\alpha_i \cdot (\eta - \beta_i)]} \quad . \qquad (5\text{-}8)$$

The discrimination parameter α_i affects the slope of the item characteristic curve as illustrated in Figure 5-5. The figure exhibits two items with the same difficulty parameter $(\beta_1 = \beta_2 = 0)$, yet with different discrimination parameters: $\alpha_1 = 1$ and $\alpha_2 = 2$.

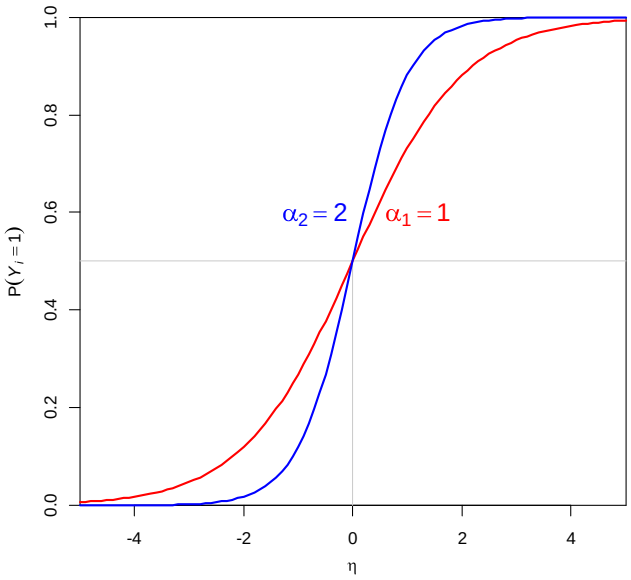


Figure 5-5: *2-PL model: ICC of two items with different discrimination parameters. The item difficulty is $\beta = 0$ for both items.*

The slope of the ICC is higher for the item with the greater discrimination parameter. The item with the greater slope exhibits a higher discrimination (difference in the probability of correct solutions) between two participants whose level of ability is close to the difficulty of the item. However, for participants with ability levels far from the item difficulty the discriminating power of the item is lower than that of the item with a smaller discrimination parameter.

Note also that the relative difficulty of the two items depends on the level of ability of the participants: For participants with an ability below the item difficulty parameter, the item with higher discrimination is more difficult, i.e. the probability of a correct response is lower than for the item with lower discrimination parameter. The opposite pattern is found for participants with ability levels above the difficulty parameter.

Unfortunately, the 2-PL model does not exhibit the nice measurement properties of the Rasch model: There are no invariant comparisons of

items or persons (specific objectivity). In addition there does not exist a common latent scale on which to locate abilities and items. The 2-PL model can be extended by adding a parameter that enables the modeling of guessing.

5.2.3 The Three-Parameter Birnbaum Model (3-PL)

The 3-PL model incorporates an additional parameter γ_i that represents the probability of guessing the right answer for item *i*. The model is represented by the following equation:

$$P\left(Y_i = 1|\eta\right) = \gamma_i + \left(1-\gamma_i\right)\cdot\frac{\exp\left[\alpha_i\cdot\left(\eta-\beta_i\right)\right]}{1+\exp\left[\alpha_i\cdot\left(\eta-\beta_i\right)\right]}$$
$$P\left(Y_i = 0|\eta\right) = 1-P\left(Y_i = 1|\eta\right) = \frac{1-\gamma_i}{1+\exp\left[\alpha_i\cdot\left(\eta-\beta_i\right)\right]} \quad . \qquad (5\text{-}9)$$

According to the model, a correct answer for item *i* may be achieved by either correct guessing or by means of a response process that conforms to the 2-PL model. The probability of guessing the correct answer is γ_i whereas the probability π_i of a correct answer in case of no correct guessing is given by the model equation of the 2-PL model. The structure of the model is illustrated in Figure 5-6.

Figure 5-7 exhibits the ICCs of two items one with no correct guessing $\left(\gamma_1 = 0\right)$, and the other one with a correct guessing probability of $\gamma_2 = 0.2$. The other two item parameters (difficulty and slope parameter) are the same for both items. The ICCs exhibit two characteristics that are specific to the 3-PL model:

1. The minimal probability of a correct response corresponds to the probability γ_i of guessing. By contrast, the minimal probability of a correct response is zero for the 1-PL and 2-PL models.
2. The probability of a correct response in case of the ability η being identical to the item difficulty β_i is no longer 0.5 in case of guessing. This is also in contrast to the 1-PL and 2-PL model.

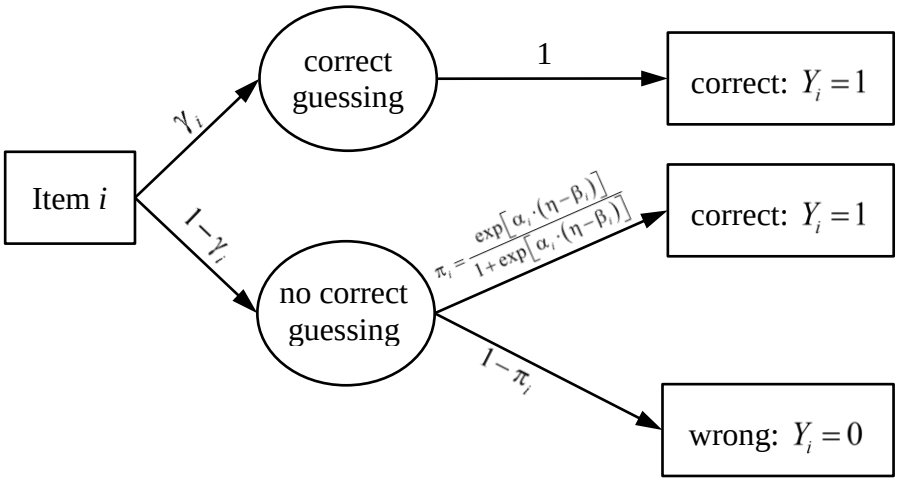


Figure 5-6: *3-PL model: The model structure.*

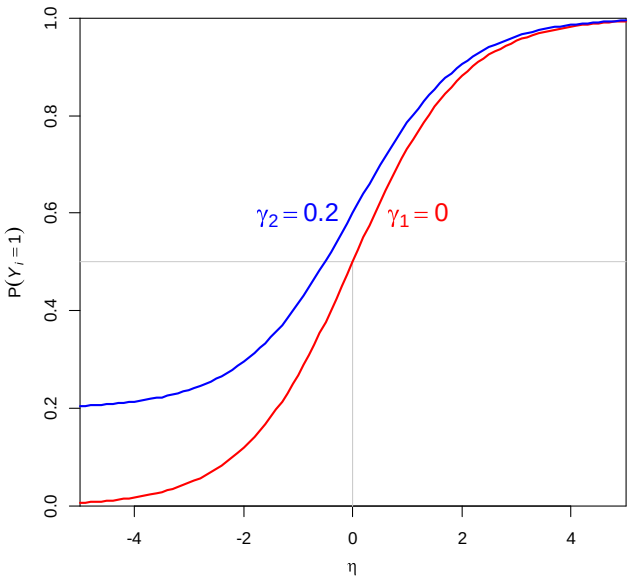


Figure 5-7: *3-PL model: ICC of two items with different guessing parameters. The item difficulty parameter is $\beta = 0$ and the discrimination parameter is $\alpha = 0$ for both items.*

5.3 The Information Functions

The information function of a test item and a test, respectively, serves a similar function as the reliability in CTT. It provides for each latent score value of the psychometric information that the item or test can provide for the specific latent score. Formally, the information is given by the following equation (cf. Embretson & Reise, 2009, p.184):

$$I\left(\eta|\beta_i,\alpha_i,\gamma_i\right) = \alpha_i^2 \cdot \frac{1-P_i}{P_i} \cdot \left(\frac{P_i-\gamma_i}{1-\gamma_i}\right)^2 \qquad (5\text{-}10)$$

The symbols have the following meaning:

η	=	value of the latent construct (ability).
β_i	=	difficulty of item *i*.
α_i	=	slope parameter of item *i*.
γ_i	=	guessing parameter of item *i*.
$I\left(\eta\middle\vert\beta_i,\alpha_i,\gamma_i\right)$	=	the information of item *i* with the given item parameters for the latent construct score η.
P_i	=	the probability of a correct response to item *i* for a given latent score (cf Equation 5-9). $$P_i=\gamma_i+\left(1-\gamma_i\right)\cdot\frac{\exp\left[\alpha_i\cdot\left(\eta-\beta_i\right)\right]}{1+\exp\left[\alpha_i\cdot\left(\eta-\beta_i\right)\right]}.$$

Notation 5-4: Functions with parameters

$I\left(\eta\middle\vert\beta_i,\alpha_i,\gamma_i\right)$ is a function of the latent construct scores η for item *i* that is characterized by the item parameters β_i, α_i and γ_i. The latter are fixed for a given item. Thus they are located on the right side of the slash.

Exercise 5-1: *Item characteristic curves (ICC) of the Rasch model*

Given: 10 test items Y_1 – Y_{10}

- ❑ The difficulty parameters of the 10 items are:
  $\beta_1 = -1.603$, $\beta_2 = -0.958$, $\beta_3 = 0.909$, $\beta_4 = 0.705$, $\beta_5 = -0.115$, $\beta_6 = 0.645$, $\beta_7 = 1.180$, $\beta_8 = -1.490$, $\beta_9 = 1.413$, $\beta_{10} = -0.686$.
- ❑ The slope parameter are $\alpha_i = 1.0$ ($i = 1, 2, \ldots, 10$).
- ❑ No guessing takes place: $\gamma_i = 1.0$ ($i = 1, 2, \ldots, 10$).

Create a figure of the item characteristic curves of the ten items in the range of $-5.0 \le \eta \le 5.0$.

Hint:
Use the function `ICC.Rasch.matrix()` of the PTT toolbox for computing the values of the ICC curves and the function `matplot()` for plotting the curves.

Exercise 5-2: *Item characteristic curves (ICC) of Birnbaum models*

Given: 3 test items Y_1 – Y_3

- ❑ The difficulty parameters of the 3 items are:
  $\beta_1 = -1$, $\beta_2 = 1$, $\beta_3 = -1$.
- ❑ The slope parameter are:
  $\alpha_1 = 0.5$, $\alpha_2 = 1.0$, $\alpha_3 = 1.0$.
- ❑ The guessing parameters are:
  $\gamma_1 = 0$, $\gamma_2 = 0.25$, $\gamma_3 = 0$.

Create a figure of the item characteristic curves of the three items in the range of $-5.0 \leq \eta \leq 5.0$.

Hint:
Use the function `ICC.Rasch.matrix()` of the PTT toolbox for computing the values of the ICC curves and the function `matplot()` for plotting the curves.

Exercise 5-3: *Conditional probability of a response pattern for the two-parameter Birnbaum-model*

Given:

- ❑ 5 test $Y_1 - Y_5$
- ❑ The item parameters of the five items are:
  $\alpha_1 = 1.0$, $\alpha_2 = 1.7$, $\alpha_3 = 1.2$, $\alpha_4 = 0.9$, $\alpha_5 = 0.5$,
  $\beta_1 = 1.0$, $\beta_2 = 2.0$, $\beta_3 = 1.5$, $\beta_4 = 2.5$, $\beta_5 = 0.8$.

The following response pattern has been observed:
$\mathbf{Y} = (1, 0, 1, 0, 1)$ (1 = correct, 0 = wrong).

Compute the conditional probability $P\left(\mathbf{Y} = (1,0,1,0,1) \middle| \, \theta = 1.7\right)$ for the two-parameter Birnbaum-model.
Use the function `P.Rasch()` of the PTT toolbox to check the correctness of your computation.

Exercise 5-4: *Information functions of the Rasch model*

Given: 10 test items $Y_1 - Y_{10}$

- ❑ The difficulty parameters of the 10 items are:
  $\beta_1 = -1.603$, $\beta_2 = -0.958$, $\beta_3 = 0.909$, $\beta_4 = 0.705$, $\beta_5 = -0.115$,
  $\beta_6 = 0.645$, $\beta_7 = 1.180$, $\beta_8 = -1.490$, $\beta_9 = 1.413$, $\beta_{10} = -0.686$.
- ❑ The slope parameter are $\alpha_i = 1.0$ ($i = 1, 2, \ldots, 10$).
- ❑ No guessing takes place: $\gamma_i = 1.0$ ($i = 1, 2, \ldots, 10$).

Create a figure of the information functions of the single items as well as of the whole test in the range of $-5.0 \leq \eta \leq 5.0$.

Hint:
Use the function `I.Rasch.matrix()` of the PTT toolbox for computing the values of the ICC curves and the function `matplot()` for plotting the curves.

References

Ackerman, P. L., Beier, M.E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131,* 30-60.

Agresti, A. (2002*). Categorical data analysis* (2nd edition). New York: Wiley..

Aronson, E., Wilson, T. D., & Akert, R. M. (2010). *Social psychology* (7th edition). Boston: Pearson. [HPAED- Y 2039A]

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4,* 417-423.

Baker, F. B., & Kim, S-H. (2004). *Item response theory: Parameter estimation techniques* (2nd edition). New York: Dekker. [HPAED PC-292]

Bargh, J. A. (1994). The four horseman of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Handbook of social cognition* (2nd edition, Vol.1, pp. 1-40). Hillsdale, NJ: Erlbaum.

Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer, Jr. (ed.), *Advances in social cognition* (Vol. 10, pp. 1-61). Mahwah, NJ: Erlbaum.

Bechtoldt, H. P. (1959). Construct validity: A critique. *American Psychologist, 14,* 619-629.

Beier, M. E., & Ackerman, P.L. (2005). Working memory and intelligence: Different constructs. Reply to Oberauer et al. (2005) and Kane et al. (2005). *Psychological Bulletin, 131,* 72-75.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, F. M., & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, Mass.: Addison Wesley.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46,* 443-459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomous scored items. *Psychometrika, 35,* 179-197.

Böckenholt, U. (2001). Hierarchical modeling of paired comparison data. *Psychological Methods, 6*, 49-66.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley. [HPAED Pc-152]

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University Press. [HPAED Pb-412]

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.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456-466.

Batchelder, W. H. (2010). Cognitive psychometrics: Using multinomial processing tree models. In: S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 71-93). Washington, DC: American Psychological Assocation. http://dx.doi.org/10.1037/12074-004

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika, 39,* 324-345.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validity by the multitrait-mulitmethod matrix. *Psychological Bulletin, 56,* 81-105.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review, 97,* 404-431.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671-684.

Cronbach, L. J., & Meehl, P. F. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurment and psychometrics using R.* Boca Raton, FL: CRC-Press. [HPEAD Pc-488]

Dörner D. (1983). *Lohhausen: Vom Umgang mit Unbestimmtheit und Komplexität.* Bern: Huber.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories In R. J. Sternberg, & J. E. Davidson (Eds.), *The nature of insight* (356-395). Cambridge, Mass.: MIT-Press. [HPAED Q-1427]

Dyrenforth, P. S., Kashy, D. A., Donnellan, M . B., & Lucas, R. E. (2010). Predicting relationship and life satisfaction from personality in nationally representative samples from three countries: The relative importance of actor, partner, and similarity effects. *Journal of Personality and Social Psychology*, *99,* 690-702.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Applications to abstract reasoning. *Psychological Methods, 3*, 380-396.

Embretson, S. E. (2010), *Measuring psychological constructs: Advances in model-based approaches*. Washington, DC: American Psychological Assocation. http://dx.doi.org/10.1037/12074-004 [HPAED Pb-423]

Embretson, S. E. (2010a). Measuring psychological constructs with model-based approaches: An introduction. In S. E. Embretson, S. E. (2010), *Measuring psychological constructs: Advances in model-based approaches* (pp. 1-7). Washington, DC: American Psychological Assocation. http://dx.doi.org/10.1037/12074-004

Embretson, S. E., Reise, S. P. (2009). *Item response theory for*

psychologists. London: Psycholgy Press. [HPAED Pb-474]

Festinger, L. & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance.
Journal of Abnormal and Social Psychology, 58, 203-210.

Flynn, J. R. (2009). *What is intelligence? Beyond the Flynn effect.* Cambridge, UK: Cambridge University Press.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255-282.

Herrnstein, R. J., & Murray, C. A. (1996). *The Bell curve: Intelligence and class structure in American life.* New York: Simon & Schuster.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55*, 577-602.

Jensen, A. R. (1985). The nature of the black–white difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8*, 193-263.

John, O. P., & Soto, C. J. (2007). The importance of being valid: Reliability and the process of construct validation. In: R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 461-494). Guilford: New York.

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

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier and Boyle (2005). *Psychological Bulletin, 131,* 66-71.

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

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues* (9th edition). Boston, MA: Cengage Learning. [HPAED PB-494]

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: Macmillan.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Lakatos, I. (1978). *The methodology of scientific research programs*. Cambridge, UK: Cambridge University Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison Wesley. [HPEAD Pb-72]

Mair, P. (2018). *Modern psychometriccs with R*. Cham: Springer. [HPAED Pc-491]

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525-543.

Mischel, Walter (2015). *Der Marshmallow-Test: Willensstärke, Belohnungsaufschub und die Entwicklung der Persönlichkeit.* München Siedler.

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control.* Cambridge, UK: Cambridge University Press.[HPEAD Q-1502]

Mulaik, S. A. (2009). *Linear causal modeling with structural equations.* Boca Raton: Chapman & Hall.

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*, 301-319.

Oberauer, K., Schulze, R., Wilhelm, O., & Süss, H.M. (2005). Working memory and intelligence – their correlation and their relation: Comment on Ackerman, Beier and Boyle (2005). *Psychological Bulletin, 131,* 61-65.

Pearl, J. (2009). *Causality: Models, reasoning, inference* (2nd edition). Cambridge, UK: Cambridge University Press.

Pylyshyn, Z.W.(1984). *Cognition and computation: Toward a foundation for cognitive science*. Cambridge, MA: MIT Press, A Bradford Book.

Rosseel, Y. (2012). Laavan: An R package for structural equation modeling. *Journal of Statistical Software, 48,* 1-36.

Rost, D. H. (2013). *Handbuch Intelligenz*. Weinheim: Beltz. [HPAED Q-2078]

Rost, J. (2004). *Lehrbuch Testtheorie Testkonstruktion* (2.Auflage). Bern: Huber.

Schmid, H. (1992). *Psychologische Tests : Theorie und Konstruktion.* Bern: Huber.

Schott, J. R. (2005). *Matrix analysis for statistics* (2nd editon). New York; Wiley. [HPAED PC-353]

Schurz, G., & Gebharter, A. (2016). Causality as a theoretical concept: Explanatory warrant and empirical content of the theory of causal nets. *Synthesis, 193:* 1073-1103.

Searle, S. R, Casella, G., & McCulloch, C. E. (1992). *Variance components*. Chichester: Wiley.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search.* New York: Springer. .

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodica, 3,* 25-60.

Steyer, R., & Eid, M. (2001). *Messen und Testen* (2.Auflage). Berlin: Springer. [HPAED Pb-243A+A]

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering, & R. Ostini, (Eds.), *Handbook of polytomous item response models* (pp. 43-75). New York: Taylor & Francis.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273-286.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford: Clarendon 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. [HPAED PC-89]

Appendix A: lavaan Instructions

In this appendix instructions concerning the usage of the R-package lavaan for provided. As aworking example Ex. 4-5 on p. 77 is used.

1. Setting up the model

A model is created by specifying the latent variables on the left hand side and the variables representing the tests on the right-hand side. For example, to set up a congeneric model with 4 tests, X_1, X_2, Y_1, Y_2, the following equation is specified:

```
H4 <- 'eta =~ X1 + X2 + Y1 + Y2'
```

The term `'eta =~ X1 + X2 + Y1 + Y2'` sets up the model:

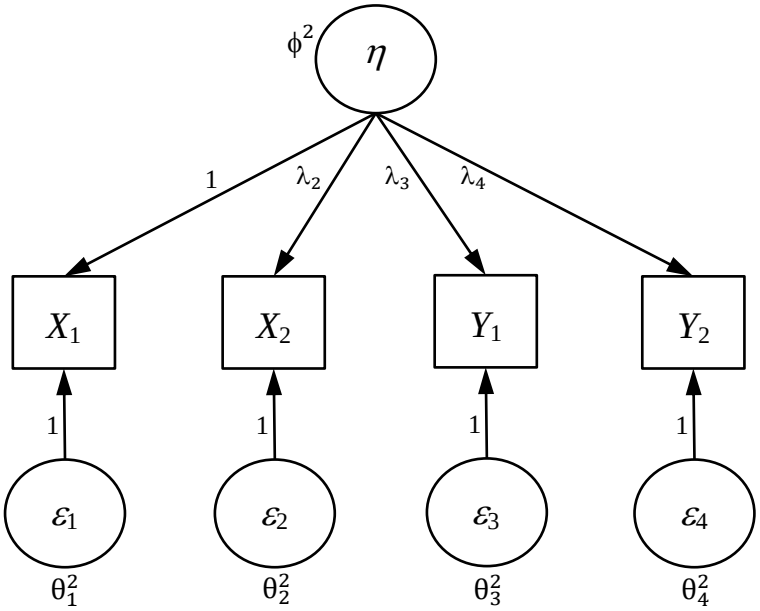


Figure A-1: *SEM model H_1.*

This model is assigned to the variable H4 (cf. the R command above). Note that the whole model specificaton has to be encosed between apostrophes.

By default lavaan assumes that the first loading coefficient is fixed at one and the variance of the latent construct is assumed to be zero.

A model with more thatn a single latent variable is set up by specifying an equation for each latent variable, for example:

```
H3 <- 'eta1 =~ X1 + X2
       eta2 =~ Y1 + Y2'
```

sets up the model:

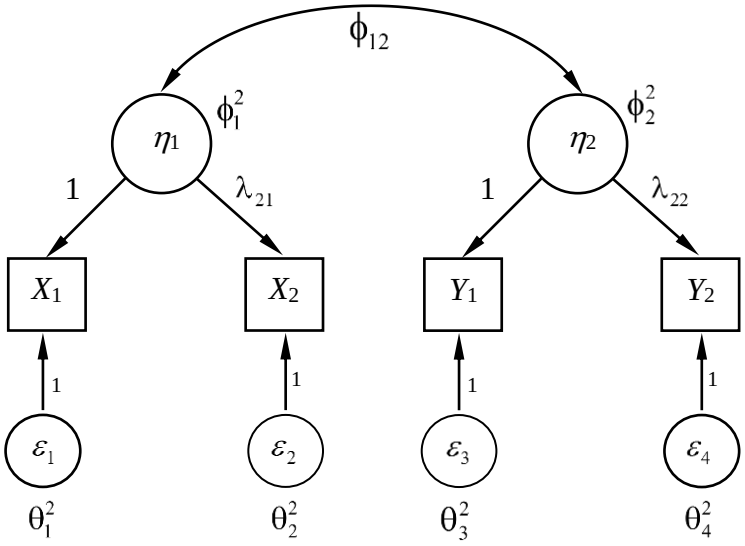


Figure A-2: *SEM model H_3.*

In order to fix the loadings to a specific value, say 1, the variable name multiplied by the value has to be specified, e.g.:

```
H1 <- 'eta1 =~ X1 + 1*X2
       eta2 =~ Y1 + 1*Y2'
```

The resulting model looks like this:

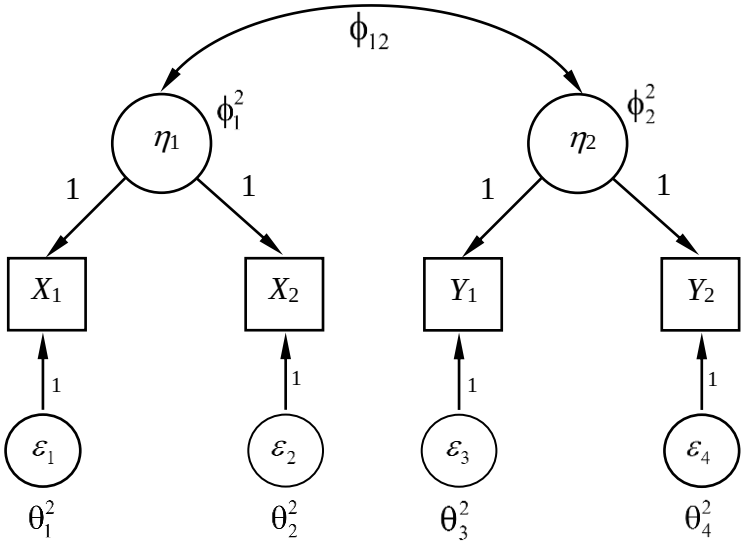


Figure A-3: *SEM model H_1.*

To set equality constraints on the loading coefficients, the same names are used. The following model dspeicification:

```
H2 <- 'eta =~ NA*X1 + 1*X1 + 1*X2 + lambda*Y1 +
              lambda*Y2'
```

results in the following model:

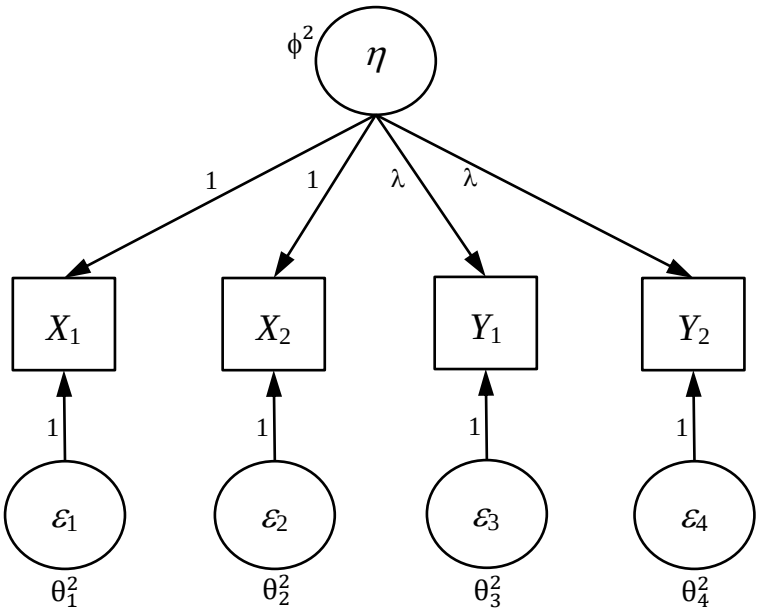


Figure A-4: *SEM model H_2.*

The specification:

```
NA*X1
```

tells the program to not fix the loading coefficient of variably X_1 to 1.

Comment A-1: A problem of lavaan:

The specification above:

```
'eta =~ NA*X1 + 1*X1 + 1*X2 + lambda*Y1 +
        lambda*Y2'
```

does not appear to be very logical since it tells the program to not set the first loading coefficient to 1 and then sets the coefficient to 1. A more logical specification:

```
'eta =~ X1 + 1*X2 + lambda*Y1 + lambda*Y2'
```

does not work, however. This seems to be an internal problem of lavvan.

2. *Fixing Variances and Covariances of Latent Constructs and Error Terms*

The parallel test model requires that the error variances are all equal. This is accomplished by adding additional terms in the model specification, for example:

```
H2 <- 'eta =~ NA*X1 + 1*X1 + 1*X2 + lambda*Y1 +
              lambda*Y2
       X1 ~~ e1*X1
       X2 ~~ e1*X2
       Y1 ~~ e2*Y1
       Y2 ~~ e2*Y2'
```

tells lavvan to set the varainces of the error terms of X_1 and X_2 to be equal by using the same coefficient `e1` in both cases. Similarly, the variances of Y_1 and Y_2 are specified to be equal by using the same coefficient `e2`. Since the two coefficients are not equal, the error variances of the four variables are not the same. Using the same term, for example:

```
H2 <- 'eta =~ NA*X1 + 1*X1 + 1*X2 + lambda*Y1 +
              lambda*Y2
       X1 ~~ e*X1
       X2 ~~ e*X2
       Y1 ~~ e*Y1
       Y2 ~~ e*Y2'
```

tells the program to set all four error variances to be the same.

The fixing of variances and covariances of latent constructs proceeds in a similar fashion, for example:

```
H2 <- 'etaX =~ NA*X1 + lambda1*X1 + lambda1*X2
       etaY =~ NA*Y1 + lambda2*Y1 + lambda2*Y2

# Fix covariance structure of latent constructs

       etaX ~~ 1*etaX   # Var(etaX) = 1
       etaY ~~ 1*etaY   # Var(etaY) = 1
       etaX ~~ 1*etaY'  # Cov(etaX, etaY) = 1
```

The resulting model corresponds to the congeneric model since the varainces of the latent constructs are fixed to 1, and, thus, the covariance between the two constructs is equal to the respective correlation. By consequence, setting the covariance to 1 amounts to setting the correlation to 1. As a result, the two constructs have the same variances (and means) and a perfect correlation. They can thus be not differentiated, and the model corresponds to a model with a single latent construct only.

3. Fixing and Constraining Intercept Parameters

The τ-equivalent as well as the strictly parallel model assume that the means of the measures are all the same. This can be achieved by setting the intercepts of the measures to be equal. The following piece of R-code illustrates how to tell lavaan to constrain the intercepts of the four observed variables `X1`, `X2`, `X3`, and `X4` to be equal.

```
X1 ~ 1
X2 ~ equal("X1 ~ 1") * 1
X3 ~ equal("X1 ~ 1") * 1
X4 ~ equal("X1 ~ 1") * 1
```

In the first line an intercept is assigned to variable `X1`. Thus the program now takes into account the mean structure. The following lines tell the program to set the intercepts of the other variables equal to that of `X1`.

4. Estimating the Model

For estimating the model different fucntions are available. We use the function `cfa()` (confirmative factor analysis) which is a simplified version of the more general function `lavvan()`. The variable of the model together with additional parameters are passed to the function, for example:

```
H1.Fit <- cfa(H1, meanstructure = F, sample.cov =
              data.mat, sample.nobs = 649, likelihood
              = "Wishart")
```

The arguments passed to the function have the followoing meaning:

`H1`	The name of the specified model.
`meanstructure = F`	Only covariances are estimated and no means.
`sample.cov = data.mat`	A sample covariance matrix, called `data.mat` and a data frame with raw data is provided.
`sample.nobs = 649`	Sample size (only required in case of a covariance and mean vector is provided.
`likelihood = "Wishart"`	Maximum likelihood estimation should be performed.

Further arguments concerning specific fit options:

`missing = "fiml"`	In case of missing data this option tells the program to use *full information likelihood* (which is usually the best method).
`data = data.file`	If a data.frame containing the raw data of the different measures (instead of a covariance matrix) the name of data frame (in the actual case namede `data.file`) is given to the `data` argument. In this case the argument `sample.nobs` is not required since the program computes the sample size on the basis of the entries in the data file.

5. Extracting Results

A list of functions for displaying results is provided in Rosseel (2012), Table 4, on page 13. Here we focus on the function `inspect()` that enables the extraction of different entities from the fit object. The function has the following basic structure:

```
inspect(object, what)
```

where:

`object`	The name of the object returned by the fit function: `cfa()`, `lavaan()`.
`what`	Tells the function which information to ex-

tract.

In following we discuss the extraction of fit indices and of the matrices with model parameters.

Extraction of Fit Indices

The extraction of fit indices is performed by passing the value `"fit"` to the argument `what` of the function `inspect()`, for example:

```
H1.Res <- inspect(H1.Fit, what = "fit")
```

The function returns a vector with different fit indices together with other information, like degrees of freedom. Relevant indices as well as their postion within the vector are shown in Tab A-1.

Tab. A-1: *Important Fit Indices Returned by the Function Inspect, and their Positions within the Vector of Results.*

Fit index	Position	Comment
npar	1	Number of free parameters.
G^2	3	Chi-square statistic.
df	4	Degrees of freedom associated with G^2.
p	5	*P*-value associated with G^2.
AIC	19	Akaike's information criterion.
BIC	20	Bayesian information criterion.
N	21	Number of observations.
RMSEA	23	Root mean squared error of approximation.
RMSEA (L)	24	Lower bound of 90% CI of RMSEA.
RMSEA (U)	25	Upper bound of 90% CI of RMSEA.
RMSEA (p)	26	*P*-value associated with RMSEA.

Thus, the command:

```
inspect(H1.Fit, what = "fit")[3:5]
```

retruns a vector with the fit index G^2, the associated degrees of freedom and the associated *p*-value.

Extraction of Matrices with Parameters

The extraction of parameter matrices is performed by passing the value `"est"` to the argument `what` of the function `inspect()`, for example:

```
inspect(H3.Fit, what = "est")
```

returns the following list of matrices:

```
$lambda
    eta1  eta2
X1 1.000 0.000
X2 1.027 0.000
```

```
Y1 0.000 1.000
Y2 0.000 1.019

$theta
   X1     X2     Y1     Y2
X1 30.139
X2  0.000 26.931
Y1  0.000  0.000 24.876
Y2  0.000  0.000  0.000 22.561

$psi
     eta1   eta2
eta1 56.258
eta2 57.350 72.409
```

The matrix `lambda` denotes the matrix $\mathbf{\Lambda}$ of loading coefficients. `theta` denotes the covariance matrix $\mathbf{\Theta}$ of the error terms. The matrix `psi` denotes the covariance matrix $\mathbf{\Phi}$ of the latent constructs.

It is easy to convert the output covariance matrices to simple R matrices, for example, the following sequence of command:

```
H3.Mat <- inspect(H3.Fit, what = "est")
Psi <- matrix(H3.Mat$psi, nrow(H3.Mat$psi))
```

extracts the list of matrices (first line). The command in the second line converts the matrix `psi` to an R matrix (i.e. an object of class `matrix`) that looks like this:

```
          [,1]     [,2]
[1,] 56.25848 57.34987
[2,] 57.34987 72.40918
```

These matrices are quite useful in reliability computations (cf. Chapter 4.4).